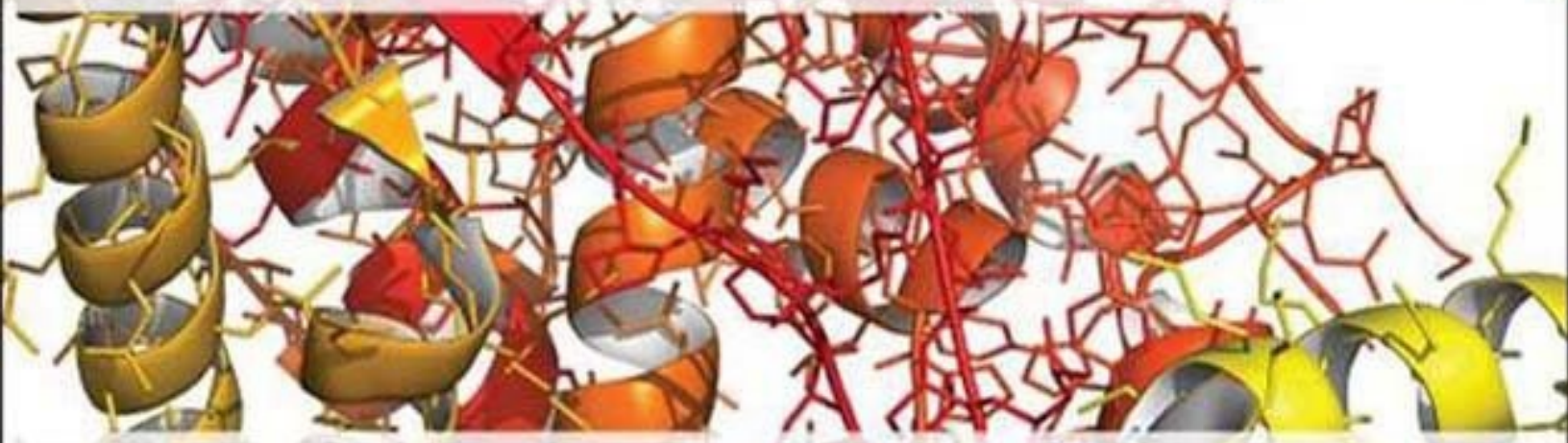
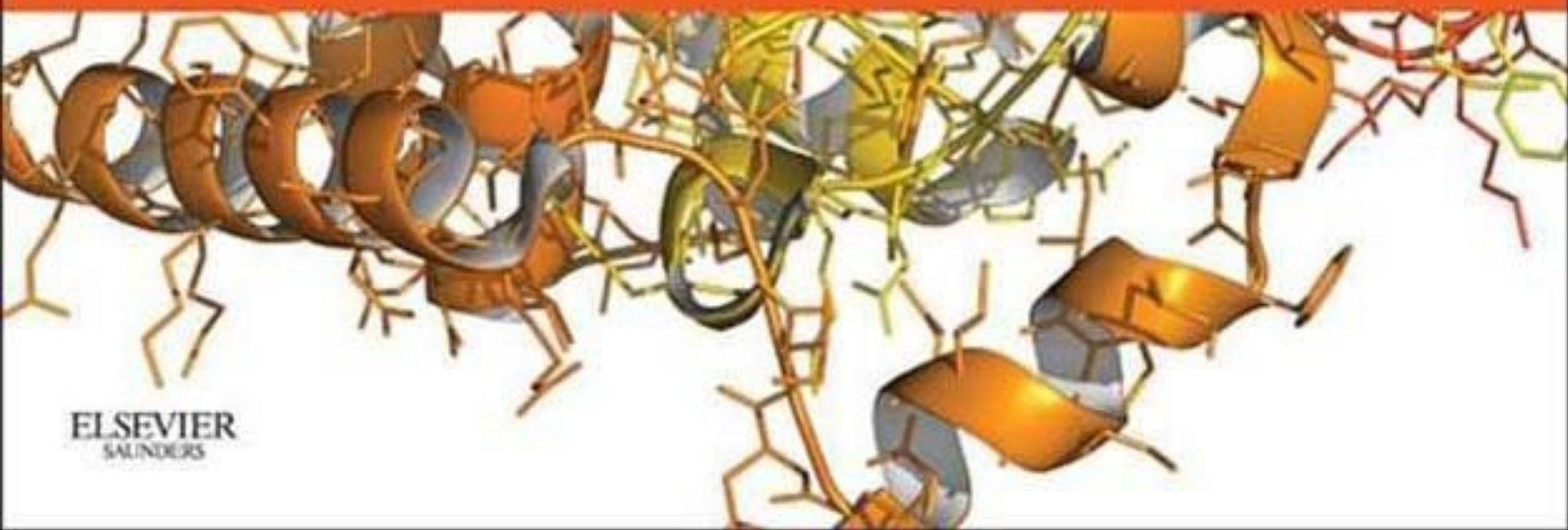
Study smart with Student Consult

John W Baynes
Marek H Dominiczak



FOURTH EDITION

Medical BIOCHEMISTRY



ELSEVIER
SAUNDERS

Medical Biochemistry

FOURTH EDITION

John W. Baynes, PhD

Carolina Distinguished Professor Emeritus, Department of Pharmacology, Physiology and Neuroscience, University of South Carolina School of Medicine, Columbia, South Carolina, USA

Marek H. Dominiczak, MD (Dr Hab Med) FRCPath FRCP (Glas)

Professor of Clinical Biochemistry and Medical Humanities, College of Medical, Veterinary and Life Sciences, University of Glasgow, United Kingdom
Docent in Laboratory Medicine, University of Turku, Finland
Consultant Biochemist, Clinical Biochemistry Service, National Health Service (NHS) Greater Glasgow and Clyde,, Gartnavel General Hospital, Glasgow, United Kingdom

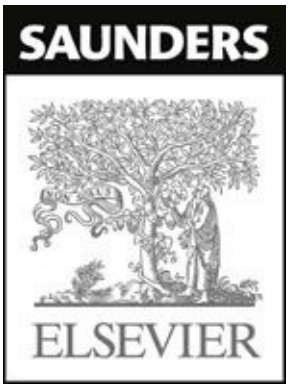

SAUNDERS
ELSEVIER

Table of Contents

Copyright

First edition 1999
Second edition 2005
Third edition 2009
Fourth edition 2014

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information,

methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

With respect to any drug or pharmaceutical products identified, readers are advised to check the most current information provided (i) on procedures featured or (ii) by the manufacturer of each product to be administered, to verify the recommended dose or formula, the method and duration of administration, and contraindications. It is the responsibility of practitioners, relying on their own experience and knowledge of their patients, to make diagnoses, to determine dosages and the best treatment for each individual patient, and to take all appropriate safety precautions.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

ISBN: 978-1-4557-4580-7

e-book ISBN: 978-1-4557-4581-4

Printed in China

Last digit is the print number: 9 8 7 6 5 4 3 2 1

Preface

We now present the 4th edition of *Medical Biochemistry*. Our aim remains, as before, to provide biochemical foundation for the study of clinical medicine – with down-to-earth practical relevance.

A textbook is a snapshot of a field as it exists at the time of writing. Such ‘photographic’ metaphor is appropriate here, because biochemistry undergoes constant change; in the period since the publication of the 3rd edition it has probably changed faster than ever before.

While core metabolic pathways remain largely unchanged, our understanding of underlying regulatory mechanisms is better, thanks to the progress in identifying signaling pathways. In many instances, these pathways have become targets for drugs, and underpin the impressive therapeutic progress in fields such as oncology.

Since completion of the Human Genome Project, genome-wide association studies and bioinformatic analyses have allowed us to put together a new picture of genetic regulation, the hallmarks of which are interactions between multiple, heterogeneous transcription factors and gene promoters, and the emerging field of epigenetics.

Behind this are, as had happened many times before in the history of science, major advances in methodology, including rapidly expanding genetic screening. The common denominator between methodologies now employed in genetic research laboratories and hospital clinical labs has been the advent of robotics and bioinformatics, and therefore the ability to process – and interpret – an ever-increasing amount of data.

This edition has again been substantially updated. We have rewritten the chapters on lipids, glucose homeostasis, nutrition and biochemical endocrinology, and added a section on the effects of exercise on muscle development and cardiovascular health. The chapter on the -omics incorporates new directions in proteomics, metabolomics and recombinant DNA technology.

This edition also benefits from the expertise of new authors who have shared

their perspectives on signaling, fat and glycoconjugate metabolism, exercise biochemistry, nutrition, and blood coagulation processes.

We have expanded the chapter on the GI tract as an important interface between the organism and the environment, and now have a separate short chapter on kidney function. In both we provide more information on membrane transport systems. We remain convinced that the biochemistry of water and electrolyte balance is as important for future clinicians as the key metabolic pathways – and deserve more emphasis in the biochemistry curricula.

We have updated literature and web references throughout the textbook. At the same time we were able to eliminate some web links in this edition, because search engines and websites such as Wikipedia and YouTube now provide quick access to so many rapidly evolving resources.

Throughout the text we strive to explain complex issues as simply as possible, but try hard not to become superficial. Unfortunately, new fields come with new terminologies and numerous additions to scientific slang. The discovery of new genes and new signaling pathways means new names and acronyms. We identify them here not as material to be committed to memory, but to help build a knowledge framework without oversimplification. The fact that some chapters may seem complex to the uninitiated may also reflect the true state of knowledge – the complexity, or even a touch of confusion, often present before a coherent picture emerges.

The Question Bank (Self-Assessment) and many more resources are available at the Elsevier website, www.studentconsult.com, to which the reader is referred. Student Consult also provides links to other Elsevier biomedical textbooks which integrate and build on knowledge of medical biochemistry. There is also a companion publication, Medical Biochemistry Flash Cards, which provides means for quick revision.

As before, we welcome comments, criticisms and suggestions from our readers. Many of these suggestions are incorporated in this 4th edition. There is no better way to continue the improvement of this text.

Contributors

Catherine N. Bagot, BSc MBBS MD MRCP FRCPath, Consultant Haematologist
Honorary Clinical Senior Lecturer
Glasgow Royal Infirmary
Glasgow, UK

Gary A. Bannon, PhD, Director
Section on Protein Analytics
Regulatory Division
Monsanto
St Louis, MO, USA

John W. Baynes, PhD, Carolina Distinguished Professor Emeritus
Department of Pharmacology, Physiology and Neuroscience
University of South Carolina School of Medicine
Columbia, SC, USA

Graham Beastall, Formerly Consultant Clinical Scientist
Department of Clinical Biochemistry
Royal Infirmary
Glasgow, UK

Hanna Bielarczyk, PhD, Assistant Professor
Head of Department of Laboratory Medicine
Department of Laboratory Medicine
Medical University of Gdańsk
Poland

John I. Broom, DSc MBChB FRCPath FRCP(Glas) FRCPE, Professor and Director
Centre for Obesity Research and Epidemiology
Robert Gordon University
Aberdeen, Scotland, UK

Wayne E. Carver, PhD, Professor of Cell Biology and Anatomy
Department of Cell Biology and Anatomy
University of South Carolina School of Medicine
Columbia, SC, USA

Marek H. Dominiczak, MD Dr Hab Med FRCPath FRCP (Glas), Hon Professor of Clinical Biochemistry and Medical Humanities

College of Medical, Veterinary and Life Sciences, University of Glasgow, UK
Docent in Laboratory Medicine
University of Turku, Finland
Consultant Biochemist
Clinical Biochemistry Service
National Health Service (NHS) Greater Glasgow and Clyde,
Gartnavel General Hospital
Glasgow, UK

†Alan D. Elbein, PhD, Professor and Chair
Department of Biochemistry and Molecular Biology
University of Arkansas for Medical Sciences
Little Rock, AR, USA

†Alex Farrell, FRCPath, Consultant Immunologist
Formerly Head of Department of Immunology and Immunopathology
Histocompatibility and Immunogenetics
Western Infirmary
Glasgow, UK

William D. Fraser, BSc MD MRCP FRCPath, Professor of Medicine
Norwich Medical School
University of East Anglia
Norwich, UK

Norma Frizzell, PhD, Assistant Professor
Department of Pharmacology, Physiology and Neuroscience
University of South Carolina School of Medicine
Columbia, SC, USA

Junichi Fujii, PhD, Professor of Biochemistry and Molecular Biology
Graduate School of Medical Science
Yamagata University
Yamagata, Japan

Helen S. Goodridge, BSc PhD, Research Scientist
Immunobiology Research Institute
Cedars-Sinai Medical Center
Los Angeles, CA, USA

J Alastair Gracie, PhD BSc (Hons), Senior University Teacher
School of Medicine
College of Medical, Veterinary and Life Sciences
University of Glasgow
Glasgow, UK

Alejandro Gugliucci, MD PhD, Director of Research and Sponsored Programs
Associate Dean of Research
Professor of Biochemistry
Touro University California

Wales, UK

Edward J. Thompson, PhD MD DSc FRCPath FRCP, Emeritus Professor of Neurochemistry
Department of Neuroimmunology
Institute of Neurology
National Hospital for Nervous Diseases
London, UK

Robert Thornburg, PhD, Professor of Biochemistry
Department of Biochemistry, Biophysics and Molecular Biology
Iowa State University
Ames, IA, USA

†A. Michael Wallace, BSc MSc PhD FRCPath, Professor,
University of Strathclyde
Consultant Clinical Scientist
Department of Clinical Biochemistry
Royal Infirmary
Glasgow, UK

Dedication

To inspirational academics

Inquisitive students

And all those who want to be good doctors

Acknowledgments

First of all, we wish to thank our contributors for sharing their expertise with us and for fitting the writing – again – into their busy research, teaching and clinical schedules. In the 4th edition, we welcome several new contributors: Catherine Bagot, Norma Frizzel, Koichi Honke, Fredrik Karpe, Matthew Kostek, Jennifer Logue, Alison Michie, Matthew Priest, Ryoji Nagai and Ian Salt. We are delighted that they have joined us.

We were saddened by the death of our good friends and contributors to previous editions, A. Michael Wallace and Alan D. Elbein.

As in the previous editions, we greatly valued the excellent secretarial assistance of Jacky Gardiner in Glasgow.

We are very grateful to students and academics from universities around the world who continue to provide us with comments, criticisms and suggestions.

The key to the whole project has been, of course, the Elsevier team. Our thanks go to Nani Clansey, Senior Development Editor, who enthusiastically steered the project through, and also to Meghan K. Ziegler and Madelene Hyde who formulated the strategy. We are very grateful to the production staff, Anne Collett, Samuel Crowe and Andrew Riley who gave the book its final form.

Our inspiration to change and improve this text comes also from ‘the field’ – from the issues, questions and decisions that arise in our everyday clinical practice, in the outpatient clinics and during wardrounds. Therefore a final thank you goes to all our clinical colleagues and doctors in training.

Abbreviations

A adenine
ABC ATP-bindimg cassette
ACE angiotensin-converting enzyme
acetyl-CoA acetyl coenzyme A
ACh acetylcholine
ACP acyl carrier protein
ACTase aspartate carbamoyl transferase
ACTH adrenocorticotropic hormone
ADC AIDS-dementia complex
ADH alcohol dehydrogenase
ADH antidiuretic hormone (also known as AVP)
ADP adenosine diphosphate
AE anion exchanger
AFP α-fetoprotein
AGE advanced glycoxidation (glycation) endproduct
AHF antihemophilic factor
AICAR 5-aminoimidazole-4-carboxamide ribonucleotide
AIDS acquired immunodeficiency syndrome
AIR 5-aminoimidazole ribonucleotide
ALDH aldehyde dehydrogenase
ALP alkaline phosphatase
ALT alanine aminotransferase
AML acute myeloblastic leukemia
AMP adenosine monophosphate
ANP atrial natriuretic peptide
APC adenomatous polyposis coli (gene)
apoA, B, etc. apolipoprotein A, B, *etc.*
APRT adenosine phosphoribosyl transferase
APTT activated partial thromboplastin time

AQP aquaporin
ARDS acute respiratory distress syndrome
ARE antioxidant response element
AST aspartate aminotransferase
ATF activation transcription factor
ATM ataxia telangiectasia-mutated gene
ATP adenosine triphosphate
AVP arginine-vasopressin (same as antidiuretic hormone)
AZT azido-2′,3′-dideoxythymidine
Bcl-2 B cell lymphoma protein 2
BMI body mass index
BMR basal metabolic rate
BNP brain natriuretic peptide
bp base pair
2,3-BPG 2,3-bisphosphoglycerate
BUN blood urea nitrogen equivalent of (but not the same as) serum urea
bw body weight
C cytosine
CA carbonic anhydrase
CAD caspase-dependent endonuclease
CAIR carboxyaminoimidazole ribonucleotide
cAMP cyclic AMP
CAT catalase
CD cluster designation: classification system for cell surface molecules
CDG congenital disorders of glycosylation
CDGS carbohydrate-deficient glycoprotein syndromes
CDK cyclin-dependent kinase
CDKI cyclin-dependent kinase inhibitor
CDP cytidine diphosphate
CFTR cystic fibrosis transmembrane conductance regulator
cGMP cyclic GMP
CGRP calcitonin gene-related peptide
CML chronic myeloid leukemia
CMP cytidine monophosphate
CNS central nervous system
COAD chronic obstructive airways disease (synonym: COPD)
COMT catecholamine-*O*-methyl transferase

COPD chronic obstructive pulmonary disease (synonym: COAD)
CoQ_{10} coenzyme Q_{10} (ubiquinone)
COX-1 cyclooxygenase-1
CK creatine phosphokinase (also CPK)
CPK creatine phosphokinase (also CK)
CPS I, II carbamoyl phosphate synthetase I, II
CPT I, II carnitine palmitoyl transferase I, II
CREB cAMP-response element-binding protein
CRGP calcitonin-related gene peptide
CRH corticotropin-releasing hormone
CRP C-reactive protein
CSF cerebrospinal fluid
CT calcitonin
CTP cytidine triphosphate
CVS chorionic villous sampling
DAG diacylglycerol
DCC 'delete in colon carcinoma' gene
dNPs deoxynucleotides
ddNPs dideoxynucleotides
DEAE diethylaminoethyl
DGGE denaturing-gradient gel electrophoresis
DHAP dihydroxyacetone phosphate
DIC disseminated intravascular coagulation
DIPF, DFP diisopropylphosphofluoride
DNA deoxyribonucleic acid
DNP 2,4-dinitrophenol
dNTPs deoxynucleotides triphosphates
Dol-P dolichol phosphate
Dol-PP-GlcNAc dolichol pyrophosphate-*N*- acetylglucosamine
DOPA dihydroxyphenylalanine
DPPC dipalmitoyl phosphatidyl choline
DVT deep vein thrombosis
EBV Epstein–Barr virus
ECF extracellular fluid
ECM extracellular matrix
EDRF endothelium-derived relaxing factor (nitric oxide)
EDTA ethylenediaminetetraacetic acid

EF-1, 2 elongation factor-1,2
EFAs essential fatty acids
EGF epidermal growth factor
eIF-3 eukaryotic initiation factor 3
EMSA electrophoretic mobility shift assay
ENaC epithelial sodium channel
ER endoplasmic reticulum
ERK extracellular signal-regulated kinase
ESR erythrocyte sedimentation rate
FACIT fibril-associated collagen with interrupted triple helices
FAD flavin adenine dinucleotide
FADD a ‘death domain’ accessory protein
$FADH_2$ reduced flavin adenine dinucleotide
FAICAR 5-formylaminoimidazole-4-carboxamide ribonucleotide
FAP familial adenomatous polyposis
Fas apoptosis signaling molecule: a ‘death domain’ accessory protein (CD95)
FBPase fructose bisphosphatase
FDP fibrin degradation product
FGAR formylglycinamide ribonucleotide
FGF fibroblast growth factor
FHH familial hypocalciuric hypercalcemia
FMN flavin mononucleotide
$FMNH_2$ reduced flavin mononucleotide
FRAXA fragile X syndrome
Fru-1,6-BP fructose-1,6 bisphosphate
Fru-2,6-BP fructose-2,6 bisphosphate
Fru-2,6-BPase fructose-2,6 bisphosphatase
Fru-6-P fructose-6-phosphate
FSF fibrin-stabilizing factor
FSH follicle-stimulating hormone
G guanine
G3PDH glyceraldehyde-3-phosphate dehydrogenase
GABA γ-amino butyric acid
GAG glycosaminoglycan
Gal galactose
Gal-1-P galactose-1-phosphate
GalNAc *N*-acetylgalactosamine

$GalNH_2$ galactosamine
GAP guanosine–triphosphatase activating protein
GAPDH glyceraldehyde-3-phosphate dehydrogenase
GAR glycinamide ribonucleotide
GDH glutamate dehydrogenase
GDP guanosine diphosphate
GDP-Fuc guanosine diphosphate-L-fucose
GDP-Man guanosine diphosphate-mannose
GFAP glial fibrillary acidic protein
γGT γ-glutamyl transferase
GH growth hormone
GHRH growth hormone-releasing hormone
GIP glucose-dependent insulinotropic peptide
GIT gastrointestinal tract
GK glucokinase
Glc glucose
Glc-1-P glucose-1-phosphate
Glc-6-P glucose-6-phosphate
Glc-6-Pase glucose-6-phosphatase
GlcN-6-P glucosamine-6-phosphate
GlcNAc *N*-acetylglucosamine
GlcNAc-1P *N*-acetylglucosamine-1-phosphate
GlcNAc-6-P *N*-acetylglucosamine-6-phosphate
$GlcNH_2$ glucosamine
GlcUA D-glucuronic acid
GLP-1 glucagon-like peptide-1
GLUT glucose transporter (GLUT-1 to GLUT-5)
GM_1 monosialoganglioside 1
GMP guanosine monophosphate
GnRH gonadotropin-releasing hormone
GP1b−IXa (etc.) glycoprotein receptor 1b−IXa (etc.)
GPx glutathione peroxidase
GRE glucocorticoid response element
GSH reduced glutathione
GSSG oxidized glutathione
GTP guanosine triphosphate
GTPase guanosine triphosphatase

5-HIAA 5-hydroxyindoleacetic acid
5-HT 5-hydroxytryptamine
Hb hemoglobin
HBOC Hb-based oxygen carrier
HCM hypercalcemia associated with malignancy
Hct hematocrit
HDL high-density lipoprotein
HGF-R hepatocyte growth factor receptor
HGP Human Genome Project
HGPRT hypoxanthine-guanine phosphoribosyl transferase
HIV human immunodeficiency virus
HLA human leukocyte antigen (system)
HLH helix-loop-helix (motif)
HMG hydroxymethylglutaryl
HMWK high-molecular-weight kininogen
HNPCC hereditary nonpolyposis colorectal cancer
hnRNA heteronuclear ribonucleic acid
HPLC high-performance liquid chromatography
HPT hyperparathyroidism
HRT hormone replacement therapy
HTGL hepatic triglyceride lipase
HTH helix-turn-helix (motif)
ICAM-1 intracellular cell adhesion molecule-1
ICF intracellular fluid
IDDM insulin-dependent diabetes mellitus (term now substituted by Type 1 diabetes mellitus)
IDL intermediate-density lipoprotein
IdUA L-iduronic acid
IEF isoelectric focusing
IFN-γ interferon-γ
Ig immunoglobulin
IGF insulin growth factor
IGF-1 insulin-like growth factor 1
IL interleukin (IL-1 to IL-29)
IMP inosine monophosphate
Inr initiator (nucleotide sequence of a gene)
IP_1 I-1-P_1, I-4-P_1 (etc.) inositol monophosphates

IP_2, I-1,3-P_2, I-1,4-P_2 inositol bisphosphates
$IP_{3,}$ I-1,4,5-P_3 inositol trisphosphate
IRE iron response element
IRE-BP IRE-binding protein
IRMA immunoradiometric assay
ITAM immunoreceptor tyrosine activation motif
ITIM immunoreceptor tyrosine inhibition motif
JAK Janus kinase
JNK Jun *N*-terminal kinase
K equilibrium constant
kb kilobase
kbp kilobase pair(s)
KCCT kaolin-cephalin clotting time
KIP2 cell cycle regulatory molecule
K_m Michaelis constant
LACI lipoprotein-associated coagulation inhibitor
LCAT lecithin: cholesterol acyltransferase
LDH lactate dehydrogenase
LDL low-density lipoprotein
LH luteinizing hormone
LPL lipoprotein lipase
LPS lipopolysaccharide
LRP LDL-receptor-related protein
M-CSF-R macrophage colony-stimulating factor receptor
malonyl-CoA malonyl coenzyme A
Man mannose
Man-1-P mannose-1-phosphate
Man-6-P mannose-6-phosphate
MAO monoamine oxidase
MAPK mitogen-activated protein kinase (a superfamily of signal-transducing kinases)
Mb myoglobin
MCHC mean corpuscular hemoglobin concentration
MCP-1 monocyte chemoattractant protein-1
MCV mean corpuscular volume
MDR multidrug resistance
MEKK mitogen-activated protein kinase kinase

MEN IIA multiple endocrine neoplasia type IIA
met-tRNA methionyl-tRNA
MGUS monoclonal gammopathy of uncertain significance
MHC major histocompatibility complex
miRNAs micro RNAs
MMPs matrix metalloproteinases
MPO myeloperoxidase
mRNA messenger ribonucleic acid
MRP multidrug resistance-associated protein
MS mass spectrometry
MSH melanocyte-stimulating hormone
MSUD maple syrup urine disease
MyoD muscle cell-specific transcription factor
Na^+/K^+-ATPase sodium-potassium ATPase
NABQI *N*-acetyl benzoquinoneimine
NAC *N*-acetylcysteine
NAD^+ nicotinamide adenine dinucleotide (oxidized)
NADH nicotinamide adenine dinucleotide (reduced)
$NADP^+$ nicotinamide adenine dinucleotide phosphate (oxidized)
NADPH nicotinamide adenine dinucleotide phosphate (reduced)
NANA *N*-acetylneuraminic acid (sialic acid)
ncRNAs noncoding RNAs
NF nuclear factor
NF-II type II neurofibromatosis
NGF nerve growth factor
NHE sodium-hydrogen exchanger
NIDDM noninsulin-dependent diabetes mellitus (term now substituted by Type 2 diabetes mellitus)
NKCC1 Na-K-Cl co-transporter 1 (etc.)
NMDA *N*-methyl-D-aspartate
NPY neuropeptide Y
NSAID nonsteroidal antiinflammatory drug
nt nucleotide (as measure of size/length of a nucleic acid)
$1,25(OH)_2D_3$ 1,25-dihydroxy vitamin D_3
OGTT oral glucose tolerance test
8-oxoG 8-oxo-2′-deoxyguanosine
OxS oxidative stress

3-PG 3-phosphoglycerate
p38RK p38-reactivating kinase
Pa pascal
PA phosphatidic acid
PAF platelet-activating factor
PAGE polyacrylamide gel electrophoresis
PAI-1 plasminogen activator inhibitor-1
PAPS phosphoadenosine phosphosulfate
PC phosphatidyl choline
PC pyruvate carboxylase
PCR polymerase chain reaction
PDE phosphodiesterase
PDGF platelet-derived growth factor
PDH pyruvate dehydrogenase
PDK phosphatidylinositol dependent kinase
PE phosphatidyl ethanolamine
PEP phosphoenolpyruvic acid
PEPCK phosphoenolpyruvate carboxykinase
PF3 platelet factor 3
PFK-1 (-2) phosphofructokinase 1 (2)
PGG_2 prostaglandin G_2 (etc.)
PGK phosphoglycerate kinase
PGM phosphoglucomutase
PHHI persistent hyperinsulinemic hypoglycemia of infancy
PHP pseudohypoparathyroidism
Pi inorganic phosphate
PI-3-K phosphoinositide-3-kinase
PIP_2/PIP_3 phosphatidylinositol bisphosphate/trisphosphate
PK pyruvate kinase
PKA/PKC protein kinase A/C
PKU phenylketonuria
PL phospholipase A, *etc.*
PLA/PLC phospholipase A/C
PMA phorbol myristic acetate
PNS peripheral nervous system
PPARs peroxisome proliferator-activated receptors
PPi inorganic pyrophosphate

PRL prolactin
PrP prion protein
PRPP 5-phosphoribosyl-α-pyrophosphate
PS phosphatidylserine
PT prothrombin time
PTA plasma thromboplastin antecedent
PTH parathyroid hormone
PTHrP parathyroid hormone-related protein
PTK protein tyrosine kinase
PTPase phosphotyrosine phosphatase
Py pyrimidine base (in a nucleotide sequence)
R receptor (with qualifier, not alone)
RAIDD a ‘death domain’ accessory protein
Rb retinoblastoma protein
RBC red blood cell
RDS respiratory distress syndrome
RER rough endoplasmic reticulum
RFLP restriction fragment length polymorphism
RIP a ‘death domain’ accessory protein
RKK p38RK homologue of MEK
RNA ribonucleic acid
RNAi RNA interference
RNAPol I/II RNA polymerase I/II
RNR ribonucleotide reductase
RNS reactive nitrogen species
ROS reactive oxygen species
S Svedberg unit
SACAIR 5-aminoimidazole-4-(*N*-succinylocarboxamide) ribonucleotide
SAM *S*-adenosyl methionine
SAPK stress-activated protein kinase
SCIDS severe combined immunodeficiency syndrome
scuPA single-chain urinary-type plasminogen activator
SD standard deviation
SDS sodium dodecyl sulfate
SDS-PAGE sodium dodecyl sulfate-polyacrylamide gel electrophoresis
SEK SAPK homologue of MEK
SER smooth endoplasmic reticulum

ser-P serine phosphate
SGLT Na^+-coupled glucose symporter
SH (Figs only) steroid hormone
SH2 Src-homology region-2
SIADH syndrome of inappropriate antidiuretic hormone secretion
siRNAs small interfering RNAs
snRNA small nuclear RNA
SOD superoxide dismutase
SPCA serum prothrombin conversion accelerator
Src a protein tyrosine kinase
SRE steroid response element
SREBPs sterol regulatory element-binding proteins (SREBP 1a, SREBP 1c)
SRP signal recognition particle
SSCP single-strand conformational polymorphism
SSRI selective serotonin reuptake inhibitor
STAT signal transducer and activator of transcription
SUR sulfonylurea receptor
T thymine
T_3 tri-iodothyronine
T_4 thyroxine
TAG triacylglycerol (triglyceride)
TAP transporter associated with antigen presentation
TB tuberculosis
TBG thyroid-binding globulin
TCA tricarboxylic acid cycle
TCT thrombin clotting time
TEG thromboelastography
TF transcription factor (with qualifier)
TFPI tissue factor pathway inhibitor
TG triglyceride (triacylglycerol)
TGF(-β) transforming growth factor-β
T_{max} Km for facilitated transport protein
TNF tumor necrosis factor
TNF-R tumor necrosis factor receptor
tPA tissue-type plasminogen activator
TRADD a ‘death domain’ accessory protein
TRAFS a ‘death domain’ accessory protein

TRH thyrotropin-releasing hormone
TSH thyroid-stimulating hormone (thyrotropin)
TTP thymidine triphosphate
TXA_2 thromboxane A_2
U uridine
UCP uncoupling protein
UDP uridine diphosphate
UDP-Gal UDP-galactose
UDP-GalNAc UDP-*N*-acetylgalactosamine
UDP-Glc UDP-glucose
UDP-GlcNAc UDP-*N*-acetylglucosamine
UDP-GlcUA UDP glucuronic acid
UMP uridine monophosphate
uPA urinary-type plasminogen activator
UTR untranslated region
UV ultraviolet
VCAM-1 vascular cell adhesion molecule 1
VDCC voltage-dependent calcium channel
VIP vasoactive intestinal peptide
VLDL very low-density lipoprotein
vWF von Willebrand factor
WAF1 cell cycle regulator
X-SCID X-linked severe combined immunodeficiency
XMP xanthine monophosphate
XO xanthine oxidase
ZP3 zona pellucida 3 glycoprotein

CHAPTER 1

Introduction

John W. Baynes and Marek H. Dominiczak

Biochemistry and clinical medicine

We call this book '*Medical Biochemistry*' because it focuses on aspects of biochemistry relevant to medicine: on explaining how the body works as a chemical system and how it malfunctions during illness. Biochemistry provides a foundation for understanding the action of new drugs, such as antidepressants, drugs used to treat diabetes, hypertension and heart failure, and those that lower blood lipids. It describes clinical applications of recombinant proteins, viral vectors and the '-omics': proteomics, genomics and metabolomics. By providing insight into nutrition and exercise, and metabolic stress, it contributes to understanding how diet and lifestyle influence our health and performance, as well as how the organism ages. It describes how cellular signaling and communications systems respond to endogenous and environmental stress. It also incorporates enormous progress made in recent years in understanding human genetics, and links it to the emerging fields of nutrigenomics and pharmacogenomics, that will hopefully create a basis for therapies customized to an individual's genetic make-up.

One studies biochemistry to understand the interplay of nutrition, metabolism and genetics in health and disease

The human organism is, on the one hand, a tightly controlled, integrated and self-contained metabolic system. On the other, it is a system that is open and communicates with its environment. Despite these two seemingly contradictory characteristics, the body manages to maintain its internal environment for decades. We regularly top up our fuel (consume food) and water, and take up oxygen from inspired air to use for oxidative metabolism (which is in fact a chain of low-temperature combustion reactions). We then use the energy generated from metabolism to perform work and to maintain body temperature. We get rid of (exhale or excrete) carbon dioxide, water and nitrogenous waste. The amount and quality of food we consume have significant impact on our health – malnutrition on the one hand and obesity and diabetes on the other, are currently major public health issues worldwide.

The entire biochemistry on two pages

It is said that any text can be shortened. Thus, we took a plunge and attempted to condense our book to less than two pages. This is meant to give the reader a general overview and to create a framework for the study of subsequent chapters. The items highlighted below take you through the contents of chapters in this book.

The major structural components of the body are carbohydrates, lipids and proteins

Proteins are building blocks and catalysts; as structural units, they form the 'architectural' framework of tissues; as enzymes, together with helper molecules (**coenzymes** and **cofactors**), they catalyze biochemical reactions. **Lipids**, such as cholesterol and phospholipids, form the backbone of biological membranes.

Carbohydrates and **lipids** as monomers or relatively simple polymers are our major energy sources. They can be stored in tissues as glycogen and triglycerides. However, carbohydrates can also be linked to both proteins and lipids, and form complex structures (glycoconjugates) essential for cell signaling systems and processes such as cell adhesion and immunity.

Chemical variables, such as **pH**, **oxygen tension**, and **inorganic ion and buffer concentrations**, define the homeostatic environment in which metabolism takes place. Minute changes in this environment, for example, less than a fifth of a pH unit or just a few degrees' change in body temperature, can be life-threatening.

The blood is a unique transport medium that participates in the exchange of gases, fuels, metabolites – and information – between tissues. Moreover, the plasma, which can be easily sampled and analyzed, is a 'window' on metabolism and a rich source of clinical information.

Biological membranes partition metabolic pathways into different cellular compartments. Their water-impermeable structure is dotted with an array of 'doors and gates' (membrane transporters) and 'locks' that accept a variety of keys (hormone, cytokine and other receptors) and generate intracellular signals. They play a fundamental role in **ion** and **metabolite transport**, and in **signal transduction** both from one cell to another, and within individual cells. The fact that most of the body's energy is consumed to maintain ion and metabolite

gradients across membranes emphasizes the importance of these processes. Also, cells throughout the body are critically dependent on membrane potentials for nerve transmission, muscle contraction, nutrient transport and the maintenance of cell volume.

Energy released from nutrients is distributed in the form of adenosine triphosphate

Energy capture in biological systems occurs through **oxidative phosphorylation** which takes place in the mitochondrion. This process involves oxygen consumption, or **respiration**, by which the organism uses the energy of fuels to produce a hydrogen ion gradient across the mitochondrial membrane and capture this energy as **adenosine triphosphate (ATP)**. Biochemists call ATP the 'common currency of metabolism' because it allows energy from fuel metabolism to be used for work, transport and biosynthesis.

Metabolism is a sophisticated network of chemical processes

Carbohydrates and **lipids** are our primary sources of energy, but our nutritional requirements also include amino acids (components of **proteins**), inorganic molecules containing sodium potassium phosphate and other atoms, and micronutrients – **vitamins** and **trace elements**. Glucose is metabolized through **glycolysis**, a universal non-oxygen requiring (anaerobic) pathway for energy production. It yields pyruvate, setting the stage for oxidative metabolism in the mitochondria. It also generates metabolites that are the starting points for synthesis of **amino acids**, **proteins**, **lipids** and **nucleic acids**.

Glucose is the most important fuel for the brain: therefore maintaining its concentration in plasma is essential for survival. Glucose supply is linked to the metabolism of **glycogen**, its short-term storage form. Glucose homeostasis is regulated by the hormones that coordinate metabolic activities among cells and organs – primarily insulin and glucagon, and also epinephrine and cortisol.

Oxygen is essential for energy production but can also be toxic

During aerobic metabolism, pyruvate is transformed into **acetyl coenzyme A (acetyl-CoA)**, the common intermediate in the metabolism of carbohydrates, lipids and amino acids. Acetyl-CoA enters the central metabolic engine of the

cell, the **tricarboxylic acid cycle (TCA cycle)** located in the mitochondria. Acetyl-CoA is oxidized to **carbon dioxide** and reduces the important coenzymes **nicotinamide adenine dinucleotide (NAD^+)** and **flavin adenine dinucleotide (FAD)**. Reduction of these nucleotides captures the energy from fuel oxidation. They in turn become substrates for the final pathway, **oxidative phosphorylation**, where the electrons they carried reduce molecular oxygen through a chain of **electron transport** reactions, providing the energy for the **synthesis of ATP**. While oxygen is essential for metabolism, it can also cause oxidative stress and widespread tissue damage during inflammation. Powerful **antioxidant defenses** exist to protect cells and tissues from damaging effects of oxygen.

Metabolism continuously cycles between fasting and post-eating modes

The direction of the main pathways of carbohydrate and lipid metabolism changes in response to food intake. In the fed state, the active pathways are **glycolysis**, **glycogen synthesis**, **lipogenesis** and **protein synthesis**, rejuvenating tissues and storing the excess of metabolic fuel. In the fasting state, the direction of metabolism reverses: glycogen and lipid stores are degraded through **glycogenolysis** and **lipolysis**, providing a constant stream of **substrates for energy production**. As glycogen stores became depleted, proteins are sacrificed to make glucose through **gluconeogenesis**, guaranteeing a constant supply, while other biosynthetic pathways are slowed down. Common conditions such as diabetes mellitus, obesity and atherosclerosis that are currently major public health issues, result from impairment of fuel metabolism and transport.

Tissues perform specialized functions

Such functions include muscle contraction, nerve conduction, bone formation, immune surveillance, hormonal signaling, maintenance of pH, fluid and electrolyte balance, and detoxification of foreign substances. Specialized compounds, such as **glycoconjugates** (glycoproteins, glycolipids and proteoglycans), are needed for tissue organization and cell-to-cell communications. Recent progress in understanding cellular signaling systems has improved our insight into **cell growth**, and **repair mechanisms**. Their time-dependent decline leads to **aging**, and their failure causes diseases such as **cancer**.

The genome underpins it all

The genome provides the mechanism for conservation and transfer of genetic information, through regulation of the expression of constituent genes and their control of protein synthesis. The synthesis of individual proteins is controlled by information encoded in **deoxyribonucleic acid (DNA)** and transcribed into **ribonucleic acid (RNA)**, which is then translated into peptides that fold into **functional protein molecules**. The spectrum of expressed proteins and the control of their temporal expression during development, adaptation and aging are responsible for our protein make-up. In the last few years bioinformatics, genome-wide association studies and progress in understanding of epigenetics, provided truly fascinating insights into the complexity of genetic regulatory networks. Further, applications of **recombinant DNA** technology have revolutionized the work of clinical laboratories during the last decade. The recent ability to scan the entire genome and the potential of **proteomics** and **metabolomics** provides yet new insights into gene-driven protein synthesis.

This chapter is summarized in Figure 1.1. To think about it, the figure resembles the plan of the London Tube (see Further reading). Look at it now and don't be intimidated by the many as yet unfamiliar terms. Refer back to this figure as you progress in your studies, and you will notice how your understanding of biochemistry improves.

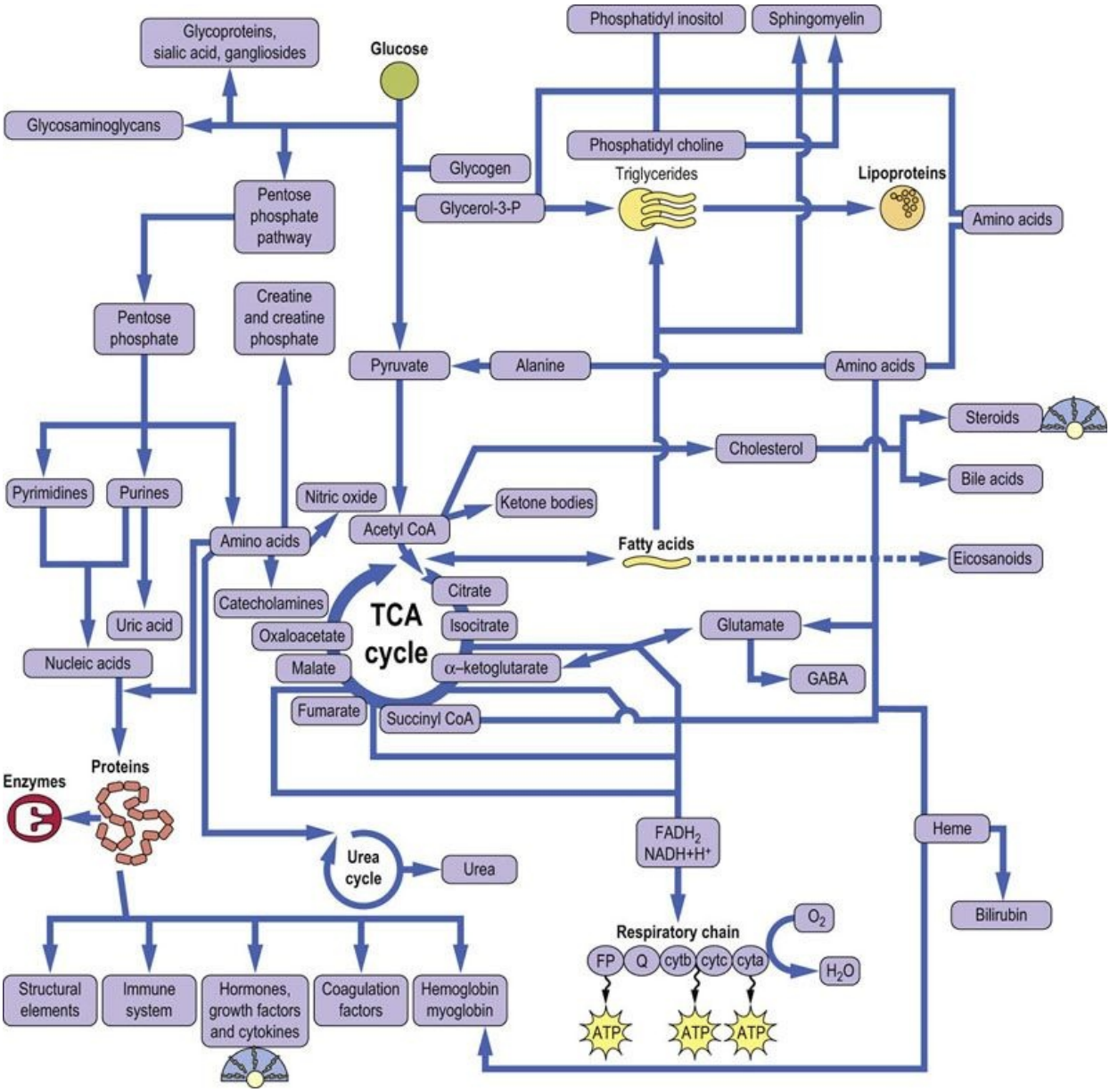


FIG. 1.1 Biochemistry: all in one.
This figure has been designed to give you a bird's-eye view of the field. It may help to structure your study or revision. Refer back to it as you study the following chapters and see how you gain perspective on biochemistry. GABA, γ-aminobutyrate; glycerol-3-P, glycerol-3-phosphate; CoA, coenzyme A; TCA cycle, tricarboxylic acid cycle; cyt, cytochrome; FP, flavoprotein; Q, coenzyme Q_{10}; ATP, adenosine 5'-triphosphate.

What this book is – and isn't

In today's medical education, acquired knowledge should be a framework for career-long study. Studying medicine piecemeal by narrow specialties is seen as less valuable than integrated learning, which places acquired knowledge in a wider context. This book attempts to do just that for biochemistry.

Keep in mind that *Medical Biochemistry* is not designed to be a review text or resource for preparation for multiple choice exams. These resources are provided separately on our website. This text is a strongly clinically oriented presentation of the science of biochemistry. It is a resource for your clinical career. It is shorter than many of the heavy tomes in our discipline, and it focuses on **explanation of key concepts and relationships** that we hope you will retain in your recall memory, and use in your future clinical practice.

As you study, remember that this is just one among the available textbooks for students and physicians. On our website, you can connect to other medical textbooks, moving readily from the biochemical aspects of a system or disease to its anatomy, physiology, pharmacology, clinical chemistry and pathology. *Medical Biochemistry* is also conveniently hyperlinked to other resources, such as clinical associations and key guidelines.

A textbook is a snapshot of rapidly changing knowledge

What only a few years ago was pure biochemical theory is now a part of the clinicians' vocabulary at the ward rounds and case conferences. A doctor (or a future doctor) does not learn biochemistry to gain theoretical brilliance: he or she learns it to be prepared for future developments in clinical practice.

We wrote *Medical Biochemistry* because we are convinced that understanding biochemistry helps in the practice of medicine. The question we asked ourselves many times during the writing process was ‘how could this piece of information improve your clinical reasoning?’ The text constantly links basic science to situations which a busy physician encounters at the bedside, in the doctor's office and when requesting tests from the clinical laboratories, which is what you will have to do when you start practicing medicine. We hope that the concepts you learn here will help you then – and benefit your patients.

Further reading

Cooke, M, Irby, DM, Sullivan, W, *et al.* American medical education 100 years after the Flexner report. *N Engl J Med*. 2006; 355:1339–1344.

Dominiczak, MH. Teaching and training laboratory professionals for the 21st century. *Clin Chem Lab Med*. 1998; 36:133–136.

Jolly B, Rees L, eds. Medical education in the millennium. Oxford University Press: Oxford, 1998; 1–268.

Ludmerer, KM. Learner-centered medical education. *N Engl J Med*. 2004; 351:1163–1164.

Tube map. www.tfl.gov.uk/assets/downloads/standard-tube-map.pdf.

CHAPTER 2

Amino Acids and Proteins

Ryoji Nagai and Naoyuki Taniguchi

Learning objectives

After reading this chapter you should be able to:

■ Classify the amino acids based on their chemical structure and charge.

■ Explain the meaning of the terms pK_a and pI as they apply to amino acids and proteins.

■ Describe the elements of the primary, secondary, tertiary, and quaternary structure of proteins.

■ Describe the principles of ion exchange and gel filtration chromatography, and electrophoresis and isoelectric focusing, and describe their application in protein isolation and characterization.

Introduction

Proteins are major structural and functional polymers in living systems

Proteins have a broad range of activities, including catalysis of metabolic reactions and transport of vitamins, minerals, oxygen, and fuels. Some proteins make up the structure of tissues, while others function in nerve transmission, muscle contraction and cell motility, and still others in blood clotting and immunologic defenses, and as hormones and regulatory molecules. Proteins are synthesized as a sequence of amino acids linked together in a linear polyamide (polypeptide) structure, but they assume complex three-dimensional shapes in performing their function. There are about 300 amino acids present in various animal, plant and microbial systems, but **only 20 amino acids are coded by DNA to appear in proteins**. Many proteins also contain modified amino acids and accessory components, termed prosthetic groups. A range of chemical techniques is used to isolate and characterize proteins by a variety of criteria, including mass, charge and three-dimensional structure. Proteomics is an emerging field which studies the full range of expression of proteins in a cell or organism, and changes in protein expression in response to growth, hormones, stress, and aging.

Amino acids

Amino acids are the building blocks of proteins

Stereochemistry: configuration at the α-carbon, D-and L-isomers

Each amino acid has a central carbon, called the α-carbon, to which four different groups are attached (Fig. 2.1):

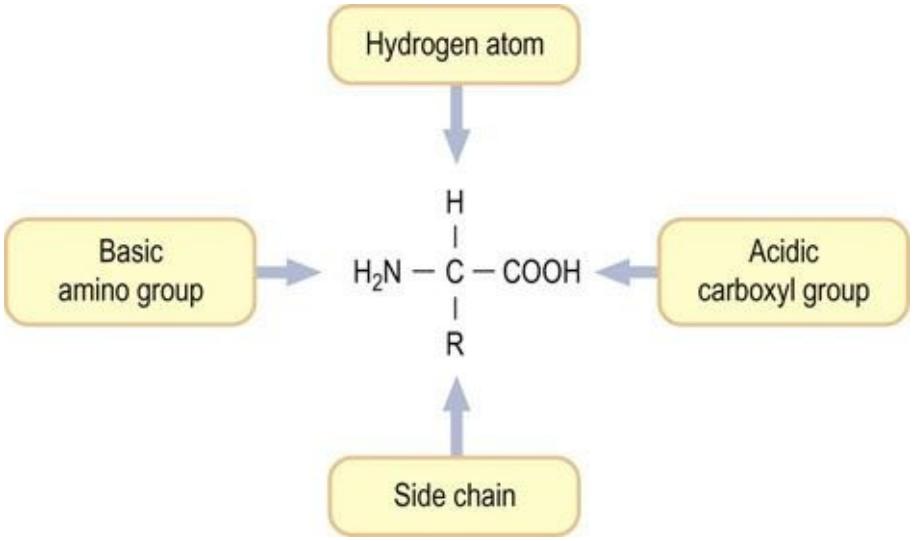


FIG. 2.1 Structure of an amino acid.
Except for glycine, four different groups are attached to the α-carbon of an amino acid. Table 2.1 lists the structures of the R groups.

- a basic amino group (—NH_2)
- an acidic carboxyl group (—COOH)
- a hydrogen atom (—H)
- a distinctive side chain (—R).

One of the 20 amino acids, proline, is not an α-amino acid but an α-imino acid (see below). Except for glycine, all amino acids contain at least one asymmetric carbon atom (the α-carbon atom), giving two isomers that are **optically active**,

i.e. they can rotate plane-polarized light. These isomers, referred to as stereoisomers or enantiomers, are said to be chiral, a word derived from the Greek word for hand. Such isomers are nonsuperimposable mirror images and are analogous to left and right hands, as shown in Figure 2.2. The two amino acid configurations are called D (for dextro or right) and L (for levo or left). **All amino acids in proteins are of the L-configuration**, because proteins are biosynthesized by enzymes that insert only L-amino acids into the peptide chains.

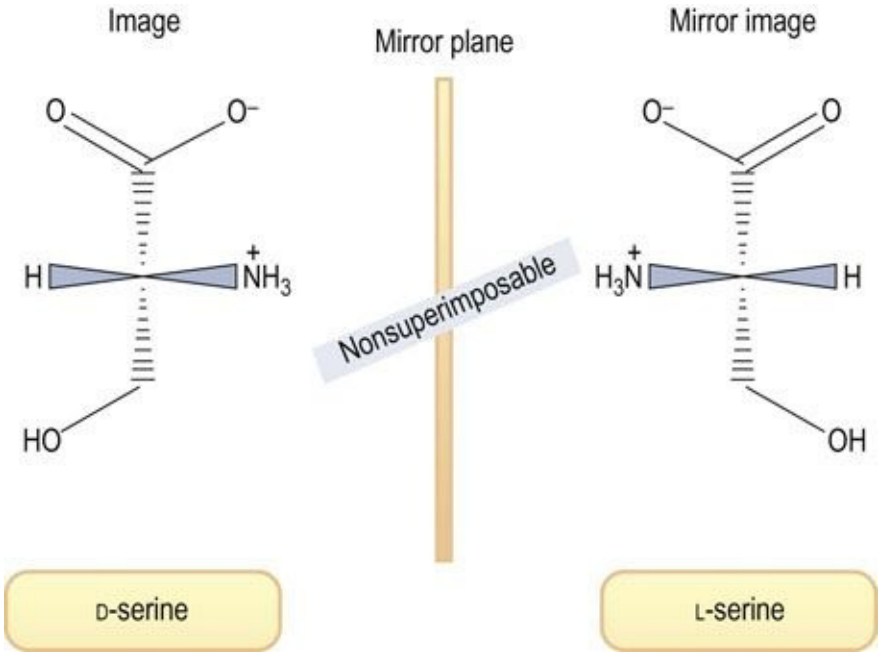


FIG. 2.2 Enantiomers.
The mirror-image pair of amino acids. Each amino acid represents nonsuperimposable mirror images. The mirror-image stereoisomers are called enantiomers. Only the L-enantiomers are found in proteins.

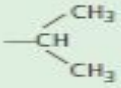
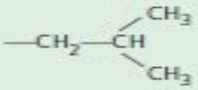
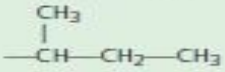
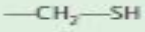
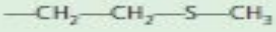
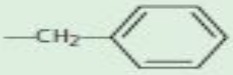
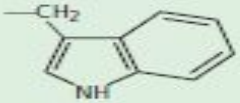
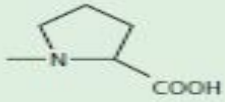
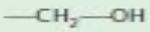
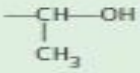
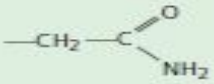
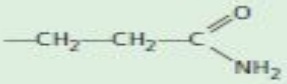
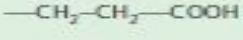
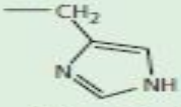
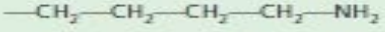
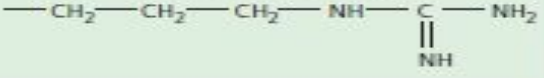
Classification of amino acids based on chemical structure of their side chains

The properties of each amino acid are dependent on its side chain (—R), which determines; the side chains are the functional groups that the structure and function of proteins, as well as the electrical charge of the molecule. Knowledge of the properties of these side chains is important for understanding methods of

analysis, purification, and identification of proteins. Amino acids with charged, polar or hydrophilic side chains are usually exposed on the surface of proteins. The nonpolar hydrophobic residues are usually buried in the hydrophobic interior or core of a protein and are out of contact with water. The 20 amino acids in proteins encoded by DNA are listed in Table 2.1 and are classified according to their side chain functional groups.

Table 2.1

The 20 Amino Acids found in proteins.*

Amino Acids	Structure of R moiety
Aliphatic Amino Acids	
glycine (Gly, **G**)	
alanine (Ala, **A**)	
valine (Val, **V**)	
leucine (Leu, **L**)	
isoleucine (Ile, **I**)	
Sulfur-containing Amino Acids	
cysteine (Cys, **C**)	
methionine (Met, **M**)	
Aromatic Amino Acids	
phenylalanine (Phe, **F**)	
tyrosine (Tyr, **Y**)	
tryptophan (Trp, **W**)	
Imino acid	
proline (Pro, **P**)	
Neutral Amino Acids	
serine (Ser, **S**)	
threonine (Thr, **T**)	
asparagine (Asn, **N**)	
glutamine (Gln, **Q**)	
Acidic Amino Acids	
aspartic acid (Asp, **D**)	
glutamic acid (Glu, **E**)	
Basic Amino Acids	
histidine (His, **H**)	
lysine (Lys, **K**)	
arginine (Arg, **R**)	

*The three-letter and single-letter abbreviations in common use are given in parentheses.

Aliphatic amino acids

Alanine, valine, leucine, and isoleucine, referred to as aliphatic amino acids, have saturated hydrocarbons as side chains. Glycine, which has only a hydrogen side chain, is also included in this group. Alanine has a relatively simple structure, a side chain methyl group, while leucine and isoleucine have *sec*- and *iso*-butyl groups. All of these amino acids are hydrophobic in nature.

Advanced concept box Nonprotein amino acids

Some amino acids occur in free or combined states, but not in proteins. Measurement of abnormal amino acids in urine (aminoaciduria) is useful for clinical diagnosis (see Chapter 19). In plasma, free amino acids are usually found in the order of 10–100 mmol/L, including many that are not found in protein. Citrulline, for example, is an important metabolite of L-arginine and a product of nitric oxide synthase, an enzyme that produces nitric oxide, an important vasoactive signaling molecule. Urinary amino acid concentration is usually expressed as µmol/g creatinine. Creatinine is an amino acid derived from muscle and is excreted in relatively constant amounts per unit body mass per day. Thus, the creatinine concentration in urine, normally about 1 mg/mL, can be used to correct for urine dilution. The most abundant amino acid in urine is glycine, which is present as 400–2000 mg/g creatinine.

Aromatic amino acids

Phenylalanine, tyrosine, and tryptophan have aromatic side chains

The nonpolar aliphatic and aromatic amino acids are normally buried in the protein core and are involved in hydrophobic interactions with one another. Tyrosine has a weakly acidic hydroxyl group and may be located on the surface

of proteins. Reversible phosphorylation of the hydroxyl group of tyrosine in some enzymes is important in the regulation of metabolic pathways. **The aromatic amino acids are responsible for the ultraviolet absorption of most proteins, which have absorption maxima ~280 nm.** Tryptophan has a greater absorption in this region than the other two aromatic amino acids. The molar absorption coefficient of a protein is useful in determining the concentration of a protein in solution, based on spectrophotometry. Typical absorption spectra of aromatic amino acids and a protein are shown in Figure 2.3.

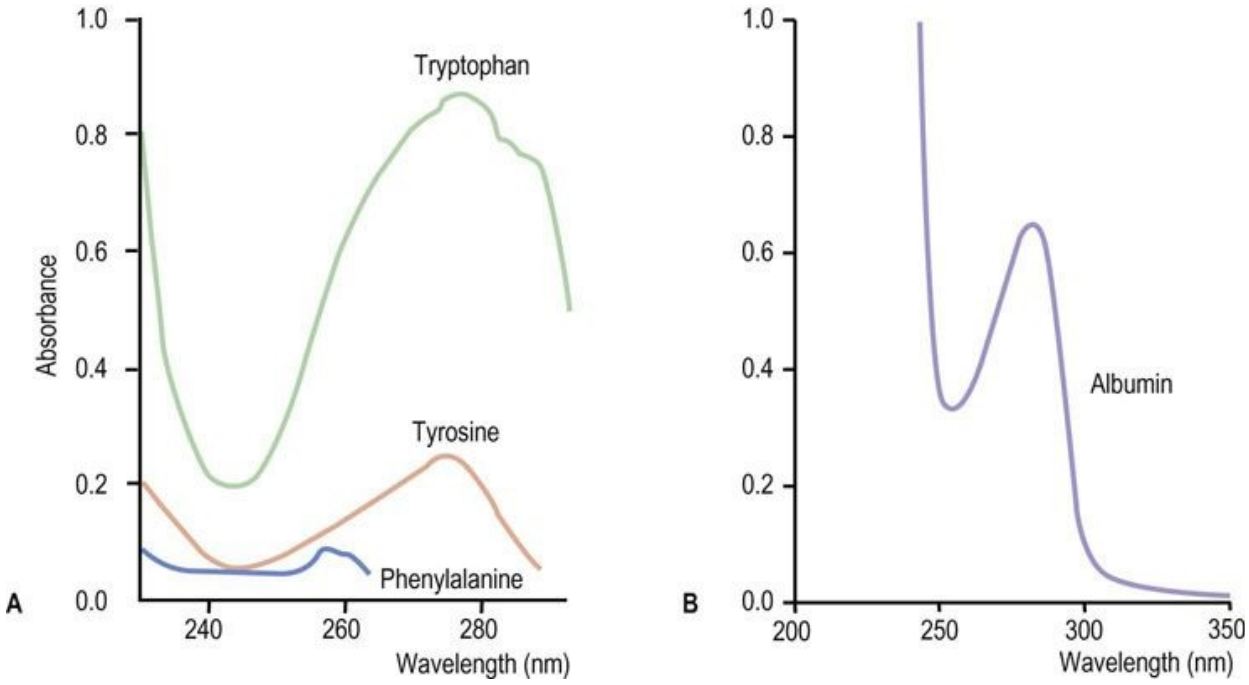


FIG. 2.3 Ultraviolet absorption spectra of the aromatic amino acids and bovine serum albumin.
(A) Aromatic amino acids such as tryptophan, tyrosine, and phenylalanine have absorbance maxima at ~280 nm. Each purified protein has a distinct molecular absorption coefficient at around 280 nm, depending on its content of aromatic amino acids. **(B)** A bovine serum albumin solution (1 mg dissolved in 1 mL of water) has an absorbance of 0.67 at 280 nm using a 1 cm cuvette. The absorption coefficient of proteins is often expressed as $E_{1\%}$ (10 mg/mL solution). For albumin, $E_{1\%}280_{nm} = 6.7$. Although proteins vary in their Trp, Tyr, and Phe content, measurements of absorbance at 280 nm are useful for estimating protein concentration in solutions.

Neutral polar amino acids

Neutral polar amino acids contain hydroxyl or amide side chain groups. Serine and threonine contain hydroxyl groups. These amino acids are sometimes found at the active sites of catalytic proteins, enzymes (Chapter 6). Reversible phosphorylation of peripheral serine and threonine residues of enzymes is also involved in regulation of energy metabolism and fuel storage in the body (Chapter 13). **Asparagine and glutamine have amide-bearing side chains. These are polar but uncharged under physiologic conditions.** Serine, threonine and asparagine are the primary sites of linkage of sugars to proteins, forming glycoproteins (Chapter 26).

Acidic amino acids

Aspartic and glutamic acids contain carboxylic acids on their side chains and are ionized at pH 7.0 and, as a result, carry negative charges on their β-and γ-carboxyl groups, respectively. In the ionized state, these amino acids are referred to as aspartate and glutamate, respectively.

Basic amino acids

The side chains of lysine and arginine are fully protonated at neutral pH and, therefore, positively charged. Lysine contains a primary amino group (NH_2) attached to the terminal ε-carbon of the side chain. The ε-amino group of lysine has a $pK_a \approx 11$. Arginine is the most basic amino acid ($pK_a \approx 13$) and its guanidine group exists as a protonated guanidinium ion at pH 7.0.

Histidine ($pK_a \approx 6$) has an **imidazole** ring as the side chain and functions as a general acid–base catalyst in many enzymes. The protonated form of imidazole is called an imidazolium ion.

Sulfur-containing amino acids

Cysteine and its oxidized form, cystine, are sulfur-containing amino acids characterized by low polarity. Cysteine plays an important role in stabilization of protein structure, since it can participate in formation of a disulfide bond with other cysteine residues to form cystine residues, crosslinking protein chains and stabilizing protein structure. Two regions of a single polypeptide chain, remote from each other in the sequence, may be covalently linked through a disulfide bond (intrachain disulfide bond). **Disulfide bonds** are also formed between two polypeptide chains (interchain disulfide bond), forming covalent protein dimers. These bonds can be reduced by enzymes or by reducing agents such as 2-

mercaptoethanol or dithiothreitol, to form cysteine residues. Methionine is the third sulfur-containing amino acid and contains a nonpolar methyl thioether group in its side chain.

Proline, a cyclic imino acid

Proline is different from other amino acids in that its side chain **pyrrolidine ring** includes both the α-amino group and the α-carbon. This imino acid forces a 'bend' in a polypeptide chain, sometimes causing abrupt changes in the direction of the chain.

Classification of amino acids based on the polarity of the amino acid side chains

Table 2.2 depicts the functional groups of amino acids and their polarity (hydrophilicity). Polar side chains can be involved in hydrogen bonding to water and to other polar groups and are usually located on the surface of the protein. Hydrophobic side chains contribute to protein folding by hydrophobic interactions and are located primarily in the core of the protein or on surfaces involved in interactions with other proteins.

Table 2.2

Summary of the functional groups of amino acids and their polarity

Amino acids	Functional group	Hydrophilic (polar) or hydrophobic (nonpolar)	Examples
acidic	carboxyl, –COOH	polar	Asp, Glu
basic	amine, $–NH_2$ imidazole guanidino	polar polar polar	Lys His Arg
neutral	glycine, –H amides, $–CONH_2$ hydroxyl, –OH sulfhydryl, –SH	nonpolar polar polar nonpolar	Gly Asn, Gln Ser, Thr, Cys
aliphatic	hydrocarbon	nonpolar	Ala, Val, Leu, Ile, Met, Pro
aromatic	C-rings	nonpolar	Phe, Trp, Tyr

Ionization state of an amino acid

Amino acids are amphoteric molecules – they have both basic and acidic groups

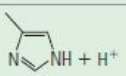
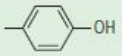
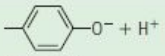
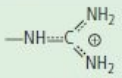
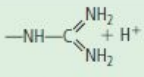
Monoamino and monocarboxylic acids are ionized in different ways in solution, depending on the solution's pH. At pH 7, the 'zwitterion' $^{+}H_3N{-}CH_2{-}COO^{-}$ is the dominant species of glycine in solution, and the overall molecule is therefore electrically neutral. On titration to acidic pH, the α-amino group is protonated and positively charged, yielding the cation $^{+}H_3N{-}CH_2{-}COOH$, while titration with alkali yields the anionic $H_2N{-}CH_2{-}COO^{-}$ species.

$$^{+}H_3N\ CH_2\ COOH \overset{H^+}{\rightleftharpoons} {}^{+}H_3N\ CH_2\ COO^{-} \overset{OH^-}{\rightleftharpoons} H_2N\ CH_2\ COO^{-}$$

pK_a values for the α-amino and α-carboxyl groups and side chains of acidic and basic amino acids are shown in Table 2.3. The overall charge on a protein depends on the contribution from basic (positive charge) and acidic (negative charge) amino acids, but the actual charge on the protein varies with the pH of the solution. To understand how the side chains affect the charge on proteins, it is worth recalling the Henderson–Hasselbalch equation.

Table 2.3

pK_a values for ionizable groups in proteins.

Group	Acid (protonated form) (conjugate acid)	H^+ + Base (unprotonated form) (conjugate base)	pK_a
terminal carboxyl residue (α-carboxyl)	–COOH (carboxylic acid)	$-COO^- + H^+$ (carboxylate)	3.0–5.5
aspartic acid (β-carboxyl)	–COOH	$-COO^- + H^+$	3.9
glutamic acid (γ-carboxyl)	–COOH	$-COO^- + H^+$	4.3
histidine (imidazole)	 (imidazolium)	 (imidazole)	6.0
terminal amino (α-amino)	$-NH_3^+$ (ammonium)	$-NH^+ + H^+$ (amine)	8.0
cysteine (sulfhydryl)	–SH (thiol)	$-S^- + H^+$ (thiolate)	8.3
tyrosine (phenolic hydroxyl)	 (phenol)	 (phenolate)	10.1
lysine (ε-amino)	$-NH_3^+$	$-NH_2 + H^+$	10.5
arginine (guanidino)	 (guanidinium)	 (guanidino)	12.5

Actual pK_a values may vary by as much at three pH units, depending on temperature, buffer, ligand binding, and especially neighboring functional groups in the protein.

Henderson–Hasselbalch equation and pK_a

The H-H equation describes the titration of an amino acid and can be used to predict the net charge and isoelectric point of a protein

The general dissociation of a weak acid, such as a carboxylic acid, is given by the equation:

$$HA \rightleftarrows H^+ + A^- \quad \textbf{(1)}$$

where HA is the protonated form (conjugate acid or associated form) and A^- is the unprotonated form (conjugate base, or dissociated form).

The dissociation constant (K_a) of a weak acid is defined as the equilibrium constant for the dissociation reaction (1) of the acid:

$$K_a = \frac{[H^+][A^-]}{[HA]} \quad \textbf{(2)}$$

The hydrogen ion concentration $[H^+]$ of a solution of a weak acid can then be calculated as follows. Equation (2) can be rearranged to give:

$$[H^+] = K_a \times \frac{[HA]}{[A^-]} \quad \textbf{(3)}$$

Equation (3) can be expressed in terms of a negative logarithm:

$$-\log[H^+] = -\log K_a - \log\frac{[HA]}{[A^-]} \quad \textbf{(4)}$$

Since pH is the negative logarithm of $[H^+]$, *i.e.* $-\log[H^+]$ and pK_a equals the negative logarithm of the dissociation constant for a weak acid, i.e. $-\log K_a$, the

Henderson–Hasselbalch equation (5) can be developed and used for analysis of acid–base equilibrium systems:

$$\mathrm{pH} = \mathrm{p}K_a + \log\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \quad \textbf{(5)}$$

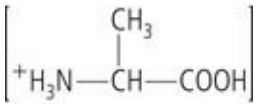
For a weak base, such as an amine, the dissociation reaction can be written as:

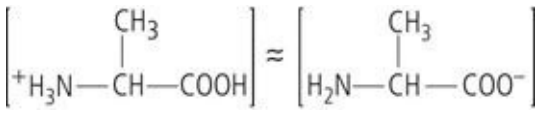
$$\mathrm{RNH_3}^+ \rightleftarrows \mathrm{H}^+ + \mathrm{RNH_2} \quad \textbf{(6)}$$

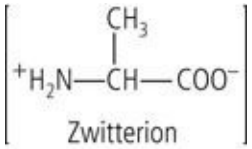
and the Henderson–Hasselbalch equation becomes:

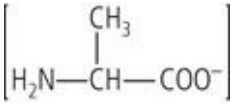
$$\mathrm{pH} = \mathrm{p}K_a + \log\frac{[\mathrm{RNH_2}]}{[\mathrm{RNH_3}^+]} \quad \textbf{(7)}$$

From equations (5) and (7), it is apparent that the extent of protonation of acidic and basic functional groups, and therefore the net charge will vary with the pK_a of the functional group and the pH of the solution. For alanine, which has two functional groups with pK_a = 2.4 and 9.8, respectively (Fig. 2.4), the net charge varies with pH, from +1 to −1. At a point intermediate between pK_{a1} and pK_{a2}, alanine has a net zero charge. This pH is called its isoelectric point, pI (Fig. 2.4).









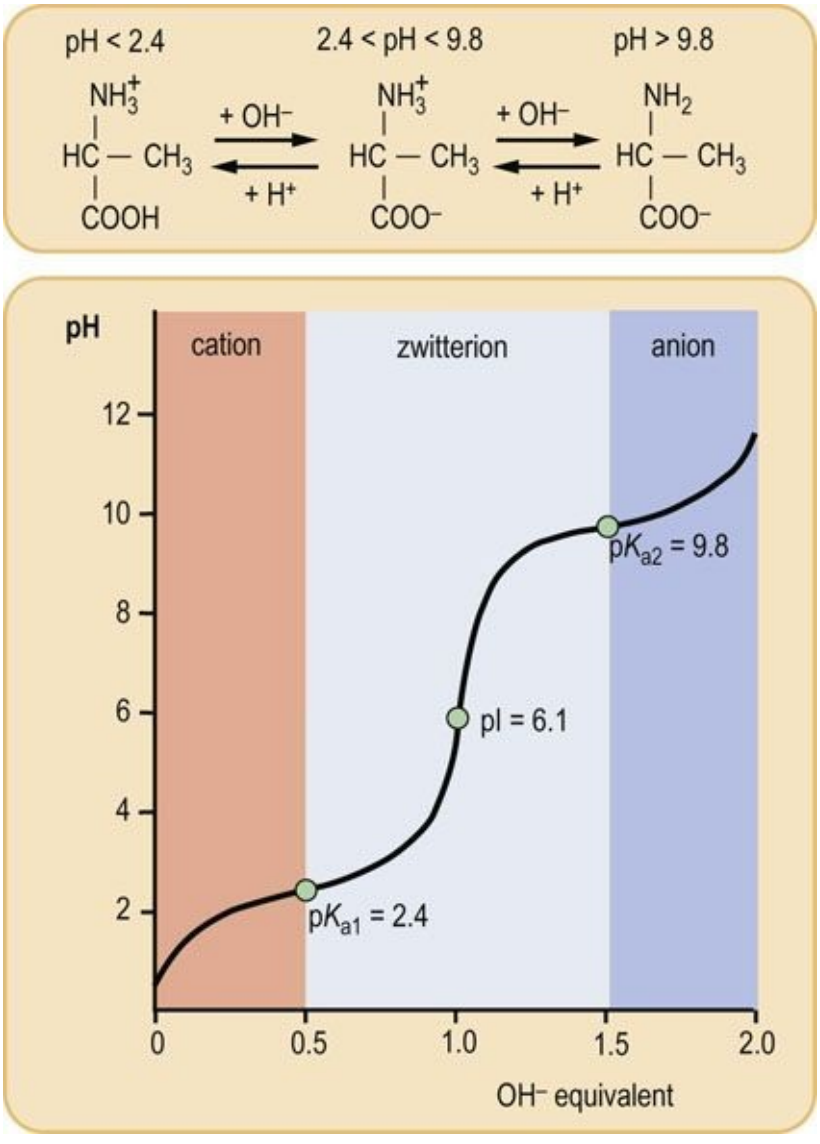


FIG. 2.4 Titration of amino acid.
The curve shows the number of equivalents of NaOH consumed by alanine while titrating

the solution from pH 0 to pH 12. Alanine contains two ionizable groups: an α-carboxyl group and an α-amino group. As NaOH is added, these two groups are titrated. The pK_a of the α-COOH group is 2.4, whereas that of the α-NH_3^+ group is 9.8. At very low pH, the predominant ion species of alanine is the fully protonated, cationic form:

At the mid-point in the first stage of the titration (pH 2.4), equimolar concentrations of proton donor and proton acceptor species are present, providing good buffering power.

At the mid-point in the overall titration, pH 6.1, the zwitterion is the predominant form of the amino acid in solution. The amino acid has a net zero charge at this pH – the negative charge of the carboxylate ion being neutralized by the positive charge of the ammonium group.

The second stage of the titration corresponds to the removal of a proton from the $\_NH_3^+$ group of alanine. The pH at the mid-point of this stage is 9.8, equal to the pK_a for the $\_NH_3^+$ group. The titration is complete at a pH of about 12, at which point the predominant form of alanine is the unprotonated, anionic form:

The pH at which a molecule has no net charge is known as its isoelectric point, pI. For alanine, it is calculated as:

$$\mathrm{pI} = \frac{\mathrm{p}K_{a1} + \mathrm{p}K_{a2}}{2} = \frac{(2.4 + 9.8)}{2} = 6.1$$

Buffers

Amino acids and proteins are excellent buffers under physiological conditions

Buffers are solutions that minimize a change in $[H^+]$, i.e. pH, on addition of acid or base. A buffer solution, containing a weak acid or weak base and a counter-ion, has maximal buffering capacity at its pK_a, *i.e.* when the acidic and basic forms are present at equal concentrations. The acidic, protonated form reacts with added base, and the basic unprotonated form neutralizes added acid, as shown below for an amino compound:

$$RNH_3^+ + OH^- \rightleftarrows RNH_2 + H_2O$$

$$RNH_2 + H^+ \rightleftarrows RNH_3^+$$

An alanine solution (Fig. 2.4) has maximal buffering capacity at pH 2.4 and 9.8, *i.e.* at the pK_a of the carboxyl and amino groups, respectively. When dissolved in water, alanine exists as a dipolar ion, or zwitterion, in which the carboxyl group is unprotonated (—COO^-) and the amino group is protonated (—NH_3^+). The pH of the solution is 6.1, the pI, half-way between the pK_a of the amino and carboxyl groups. The titration curve of alanine by NaOH (Fig. 2.4) illustrates that alanine has minimal buffering capacity at its pI, and maximal buffering capacity at a pH equal to the pK_{a1} or pK_{a2}.

Peptides and proteins

Primary structure of proteins

The primary structure of a protein is the linear sequence of its amino acids

In proteins, the carboxyl group of one amino acid is linked to the amino group of the next amino acid, forming an amide (peptide) bond; water is eliminated during the reaction (Fig. 2.5). The amino acid units in a peptide chain are referred to as amino acid residues. A peptide chain consisting of three amino acid residues is called a tripeptide, *e.g.* glutathione in Figure 2.6. By convention, the amino terminus (N-terminus) is taken as the first residue, and the sequence of amino acids is written from left to right. When writing the peptide sequence, one uses either the three-letter or the one-letter abbreviations of amino acids, such as Asp-Arg-Val-Tyr-Ile-His-Pro-Phe-His-Leu or D-R-V-Y-I-H-P-F-H-L (see Table 2.1). This peptide is angiotensin, a peptide hormone that affects blood pressure. The amino acid residue having a free amino group at one end of the peptide, Asp, is called the N-terminal amino acid (amino terminus), whereas the residue having a free carboxyl group at the other end, Leu, is called the C-terminal amino acid (carboxyl terminus). Proteins contain between 50 and 2000 amino acid residues. The mean molecular mass of an amino acid residue is about 110 dalton units (Da). Therefore the molecular mass of most proteins is between 5500 and 220,000 Da. Human carbonic anhydrase I, an enzyme that plays a major role in acid–base balance in blood (Chapter 24), is a protein with a molecular mass of 29,000 Da (29 kDa).

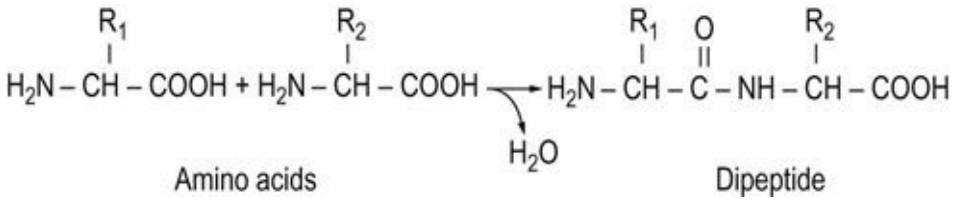


FIG. 2.5 Structure of a peptide bond.

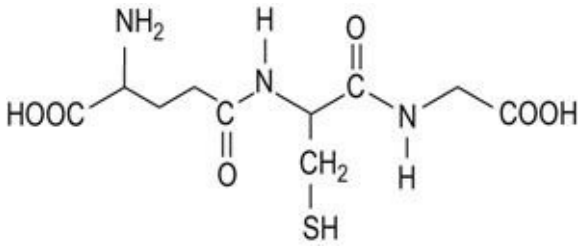


FIG. 2.6 Structure of glutathione.

Advanced concept box
Glutathione

Glutathione (GSH) is a tripeptide with the sequence γ-glutamyl-cysteinyl-glycine (Fig. 2.6). If the thiol group of the cysteine is oxidized, the disulfide GSSG is formed. GSH is the major peptide present in the cell. In the liver, the concentration of GSH is ~5 mmol/L. GSH plays a major role in the maintenance of cysteine residues in proteins in their reduced (sulfhydryl) forms and in antioxidant defenses (Chapter 37). The enzyme γ-glutamyl transpeptidase is involved in the metabolism of glutathione and is a plasma biomarker for some liver diseases, including hepatocellular carcinoma and alcoholic liver disease.

Amino acids side chains contribute both charge and hydrophobicity to proteins

The amino acid composition of a peptide chain has a profound effect on its physical and chemical properties. Proteins rich in aliphatic or aromatic amino groups are relatively insoluble in water and are likely to be found in cell membranes. Proteins rich in polar amino acids are more water soluble. Amides are neutral compounds so that the amide backbone of a protein, including the α-amino and α-carboxyl groups from which it is formed, does not contribute to the charge of the protein. Instead, the charge on the protein is dependent on the side chain functional groups of amino acids. Amino acids with side chain acidic (Glu, Asp) or basic (Lys, His, Arg) groups will confer charge and buffering capacity to

a protein. The balance between acidic and basic side chains in a protein determines its isoelectric point (pI) and net charge in solution. Proteins rich in lysine and arginine are basic in solution and have a positive charge at neutral pH, while acidic proteins, rich in aspartate and glutamate, are acidic and have a negative charge. Because of their side chain functional groups, all proteins become more positively charged at acidic pH and more negatively charged at basic pH. Proteins are an important part of the buffering capacity of cells and biological fluids, including blood.

Secondary structure of proteins

The secondary structure of a protein is determined by hydrogen bonding interactions between amino acid side chain functional groups

The secondary structure of a protein refers to the local structure of the polypeptide chain. This structure is determined by hydrogen bond interactions between the carbonyl oxygen group of one peptide bond and the amide hydrogen of another nearby peptide bond. There are two types of secondary structure: the α-helix and the β-pleated sheet.

The α-helix

The α-helix is a rod-like structure with the peptide chain tightly coiled and the side chains of amino acid residues extending outward from the axis of the spiral. Each amide carbonyl group is hydrogen-bonded to the amide hydrogen of a peptide bond that is four residues away along the same chain. There are on average 3.6 amino acid residues per turn of the helix, and the helix winds in a right-handed (clockwise) manner in almost all natural proteins (Fig. 2.7A).

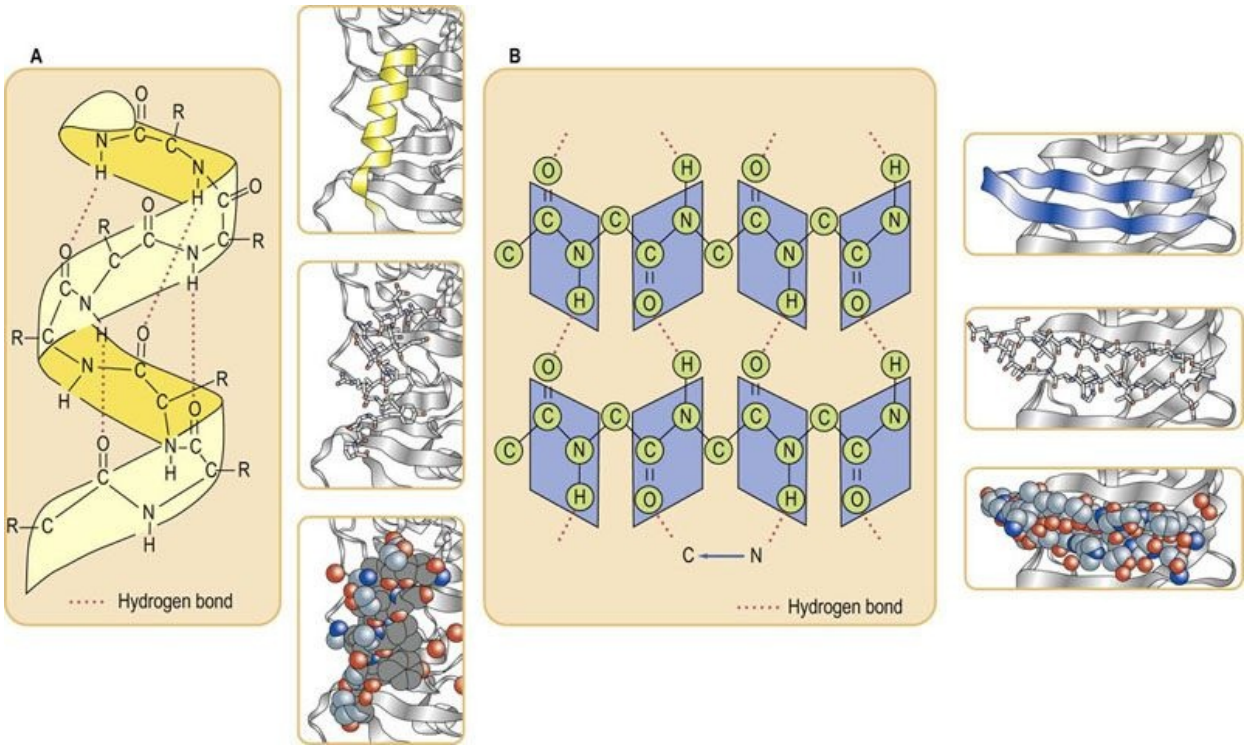


FIG. 2.7 Protein secondary structural motifs.
(A) An α-helical secondary structure. Hydrogen bonds between 'backbone' amide NH and C=O groups stabilize the α-helix. Hydrogen atoms of OH, NH or SH group (hydrogen donors) interact with electron pairs of the acceptor atoms such as O, N or S. Even though the bonding energy is lower than that of covalent bonds, hydrogen bonds play a pivotal role in the stabilization of protein molecules. R, side chain of amino acids which extend outward from the helix. Ribbon, stick and space-filling models are shown. **(B)** The parallel β-sheet secondary structure. In the β-conformation, the backbone of the polypeptide chain is extended into a zigzag structure. When the zigzag polypeptide chains are arranged side by side, they form a structure resembling a series of pleats. Ribbon, stick and space-filling models are also shown.

The β-pleated sheet

If the H-bonds are formed laterally between peptide bonds, the polypeptide sequences become arrayed parallel or antiparallel to one another in what is commonly called a β-pleated sheet. The β-pleated sheet is an extended structure as opposed to the coiled α-helix. It is pleated because the carbon–carbon (C–C) bonds are tetrahedral and cannot exist in a planar configuration. If the polypeptide chain runs in the same direction, it forms a parallel β-sheet (Fig. 2.7B), but in the opposite direction, it forms an antiparallel structure. The β-turn or β-bend refers to the segment in which the polypeptide abruptly reverses direction. Glycine (Gly) and proline (Pro) residues often occur in β-turns on the

surface of globular proteins.

Advanced concept box Collagen

Human genetic defects involving collagen illustrate the close relationship between amino acid sequence and three-dimensional structure. Collagens are the most abundant protein family in the mammalian body, representing about a third of body protein. Collagens are a major component of connective tissue such as cartilage, tendons, the organic matrix of bones, and the cornea of the eye.

Comment.

Collagen contains 35% Gly, 11% Ala, and 21% Pro plus Hyp (hydroxyproline). The amino acid sequence in collagen is generally a repeating tripeptide unit, Gly-Xaa-Pro or Gly-Xaa-Hyp, where Xaa can be any amino acid; Hyp = hydroxyproline. This repeating sequence adopts a left-handed helical structure with three residues per turn. Three of these helices wrap around one another with a right-handed twist. The resulting three-stranded molecule is referred to as tropocollagen. Tropocollagen molecules self-assemble into collagen fibrils and are packed together to form collagen fibers. There are metabolic and genetic disorders which result from collagen abnormalities. Scurvy, osteogenesis imperfecta (Chapter 28) and Ehlers–Danlos syndrome result from defects in collagen synthesis and/or crosslinking. Lens dislocation in homocysteinuria (incidence: 1 in 350,000).

Tertiary structure of proteins

The tertiary structure of a protein is determined by interactions between side chain functional groups, including disulfide bonds, hydrogen bonds, salt bridges, and hydrophobic interactions

The three-dimensional, folded and biologically active conformation of a protein is referred to as its tertiary structure. This structure reflects the overall shape of the molecule and generally consists of several smaller folded units termed **domains**. The tertiary structure of proteins is determined by X-ray crystallography and nuclear magnetic resonance spectroscopy.

The tertiary structure of a protein is stabilized by interactions between side chain functional groups: covalent disulfide bonds, hydrogen bonds, salt bridges, and hydrophobic interactions (Fig. 2.8). The side chains of tryptophan and arginine serve as hydrogen donors, whereas asparagine, glutamine, serine, and threonine can serve as both hydrogen donors and acceptors. Lysine, aspartic acid, glutamic acid, tyrosine, and histidine also can serve as both donors and acceptors in the formation of ion pairs (salt bridges). Two opposite-charged amino acids, such as glutamate with a γ-carboxyl group and lysine with an ε-amino group, may form a salt bridge, primarily on the surface of proteins (see Fig. 2.8).

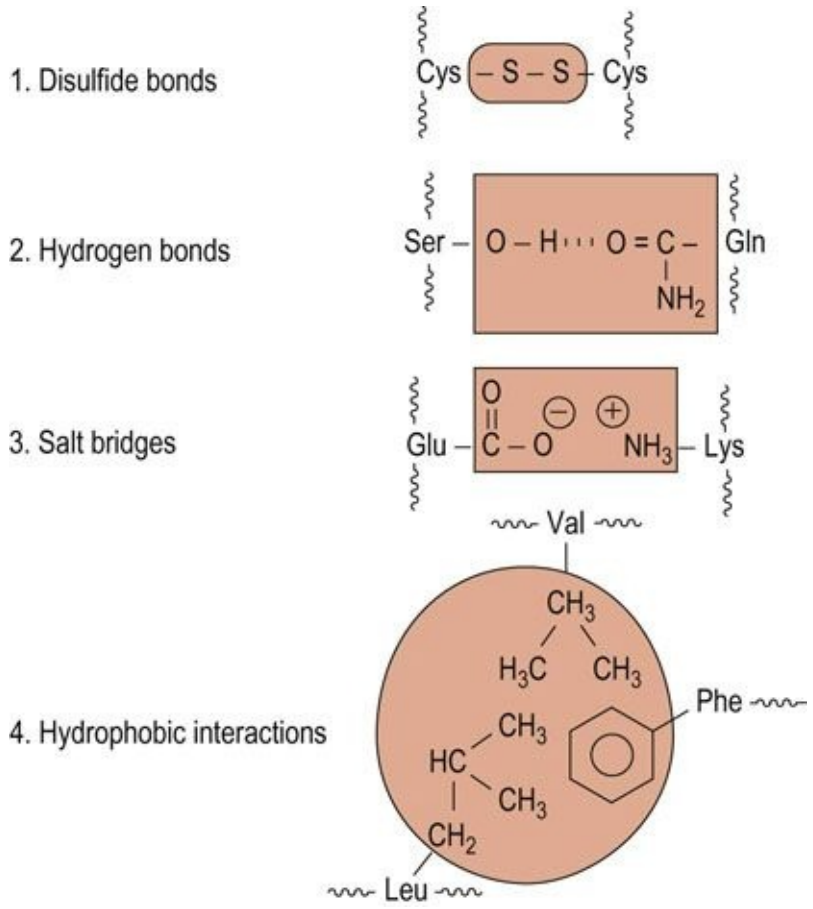


FIG. 2.8 Elements of tertiary structure of proteins.
Examples of amino acid side-chain interactions contributing to tertiary structure.

Compounds such as urea and guanidine hydrochloride cause denaturation or loss of secondary and tertiary structure when present at high concentrations for example, 8 mol/L urea. These reagents are called **denaturants** or **chaotropic agents**.

Advanced concept box Lens dislocation in HOMOCYSTINURIA (incidence: 1 in 200,000)

The most common ocular manifestation of homocystinuria, a defect in sulfur amino acid metabolism, (Chapter 19) is lens dislocation occurring around age 10 years. Fibrillin, found in the fibers that support the lens, is rich in cysteine residues. Disulfide bonds between these residues are required for the crosslinking and stabilization of protein and lens structure. Homocysteine, a metabolic intermediate and homolog of cysteine, can disrupt these bonds by homocysteine-dependent disulfide exchange.

Another equally rare sulfur amino acid disorder – sulfite oxidase deficiency – is also associated with lens dislocation by a similar mechanism (usually presenting at birth with early refractory convulsions). Marfan's syndrome, also associated with lens dislocation, is associated with mutations in the fibrillin gene (Chapter 29).

Quaternary structure of proteins is formed by interactions between peptide chains

The quaternary structure of multisubunit proteins is determined by covalent and noncovalent interactions between the subunit surfaces

Quaternary structure refers to a complex or an assembly of two or more separate

peptide chains that are held together by noncovalent or, in some cases, covalent interactions. In general, most proteins larger than 50 kDa consist of more than one chain and are referred to as dimeric, trimeric or multimeric proteins. Many multisubunit proteins are composed of different kinds of **functional subunits, such as the regulatory and catalytic subunits**. Hemoglobin is a tetrameric protein (Chapter 5), and beef heart mitochondrial ATPase has 10 protomers (Chapter 9). The smallest unit is referred to as a monomer or subunit. Figure 2.9 illustrates the structure of the dimeric protein Cu, Zn-superoxide dismutase. Figure 2.10 is an overview of the primary, secondary, tertiary, and quaternary structures of a tetrameric protein.

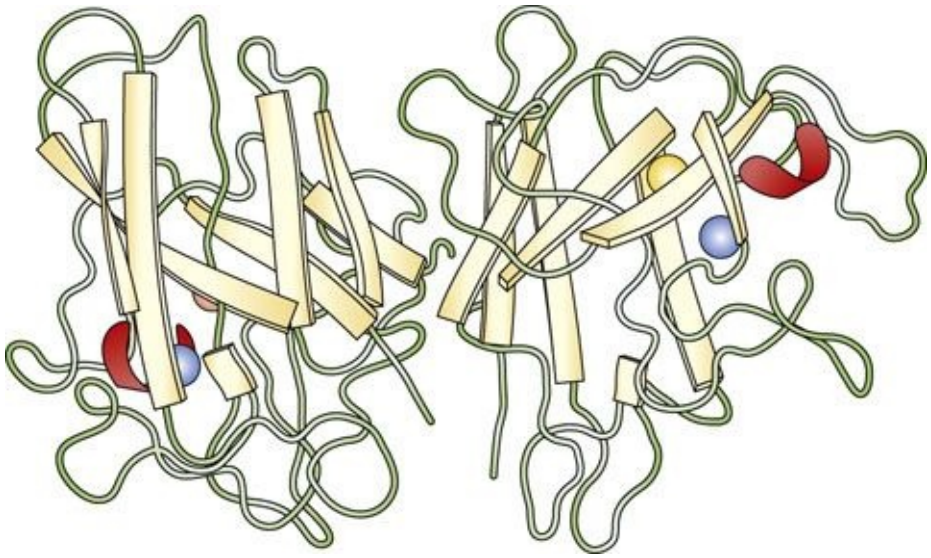


FIG. 2.9 Three-dimensional structure of a dimeric protein.
Quaternary structure of Cu,Zn-superoxide dismutase from spinach. Cu,Zn-superoxide dismutase has a dimeric structure, with a monomer molecular mass of 16,000 Da. Each subunit consists of eight antiparallel β-sheets called a β-barrel structure, in analogy with geometric motifs found on native American and Greek weaving and pottery. Red arc = intrachain disulfide bond. Courtesy of Dr Y. Kitagawa.

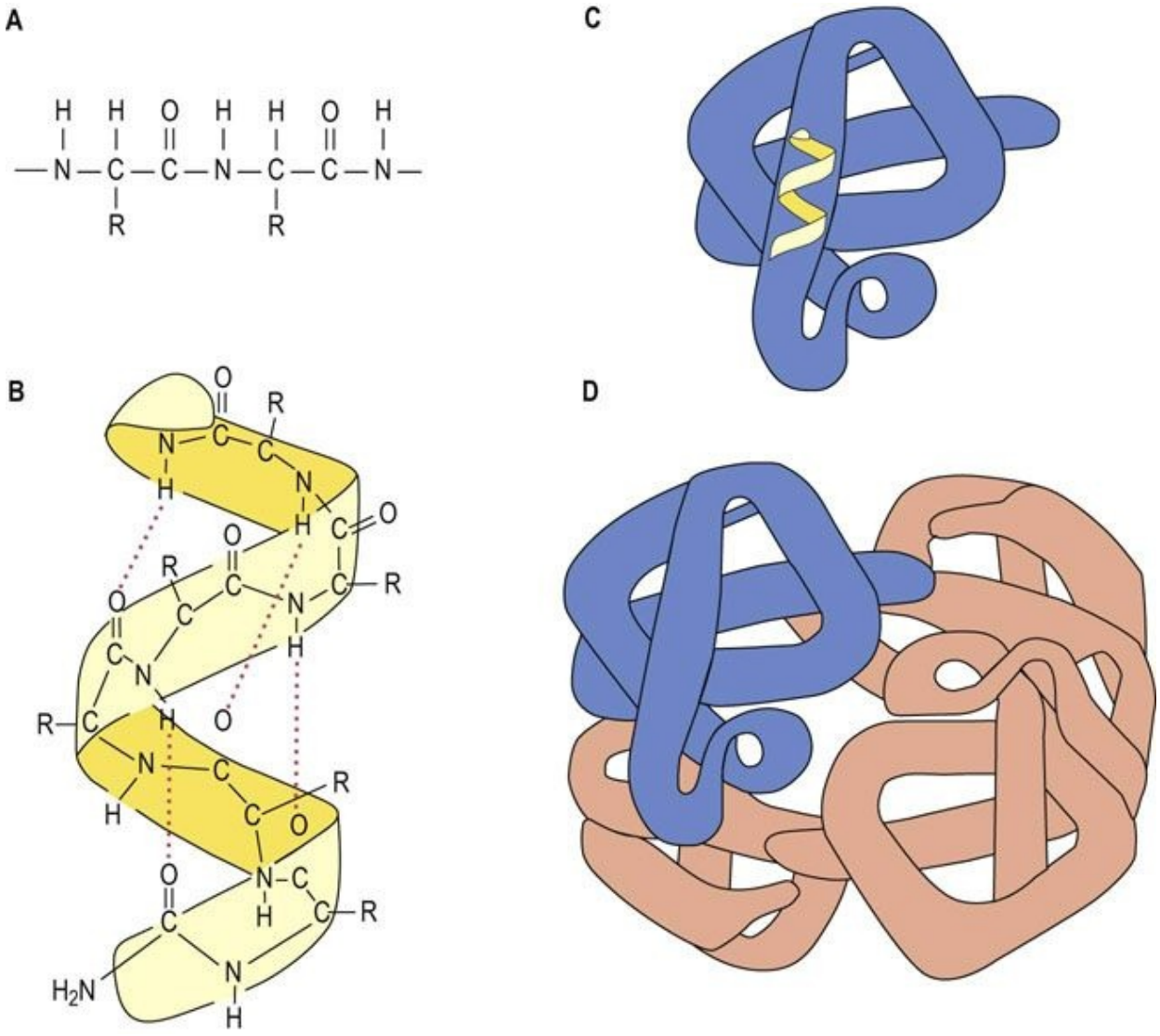


FIG. 2.10 Primary, secondary, tertiary, and quaternary structures.
(A) The primary structure is composed of a linear sequence of amino acid residues of proteins. **(B)** The secondary structure indicates the local spatial arrangement of polypeptide backbone yielding an extended α-helical or β-pleated sheet structure as depicted by the ribbon. Hydrogen bonds between the 'backbone' amide NH and C=O groups stabilize the helix. **(C)** The tertiary structure illustrates the three-dimensional conformation of a subunit of the protein, while the quaternary structure **(D)** indicates the assembly of multiple polypeptide chains into an intact, tetrameric protein.

Purification and characterization of proteins

Protein purification is a multi-step process, based on protein size, charge, solubility and ligand binding

Protein purification procedures take advantage of separations based on charge, size, binding properties, and solubility. The complete characterization of the protein requires an understanding of its amino acid composition, its complete primary, secondary and tertiary structure and, for multimeric proteins, their quaternary structure.

In order to characterize a protein, it is first necessary to purify the protein by separating it from other components in complex biological mixtures. The source of the proteins is commonly blood or tissues, or microbial cells such as bacteria and yeast. First, the cells or tissues are disrupted by grinding or homogenization in buffered isotonic solutions, commonly at physiologic pH and at 4°C to minimize protein denaturation during purification. The 'crude extract' containing organelles such as nuclei, mitochondria, lysosomes, microsomes, and cytosolic fractions can then be fractionated by high-speed centrifugation or ultracentrifugation. Proteins that are tightly bound to the other biomolecules or membranes may be solubilized using organic solvent or detergent.

Advanced concept box

PostTranslational modifications of proteins

Most proteins undergo some form of enzymatic modification after the synthesis of the peptide chain. The 'posttranslational' modifications are performed by processing enzymes in the endoplasmic reticulum, Golgi apparatus, secretory granules, and extracellular space. The modifications include proteolytic cleavage, glycosylation, lipation and phosphorylation. Mass spectrometry is a powerful tool for detecting such modifications, based on differences in molecular mass (see Chapter 35).

Salting out (ammonium sulfate fractionation) and adjustment of pH

The solubility of a protein is dependent on the concentration of dissolved salts

The solubility of a protein may be increased by the addition of salt at a low concentration (salting in) or decreased by high salt concentration (salting out). When ammonium sulfate, one of the most soluble salts, is added to a solution of a protein, some proteins precipitate at a given salt concentration while others do not. Human serum immunoglobulins are precipitable by 33–40% saturated $(NH_4)_2SO_4$, while albumin remains soluble. Saturated ammonium sulfate is about 4.1 mol/L. Most proteins will precipitate from an 80% saturated $(NH_4)_2SO_4$ solution.

Proteins may also be precipitated from solution by adjusting the pH. Proteins are generally least soluble at their isoelectric point (pI). At this pH, the protein has no net charge or charge-charge repulsion between subunits. Hydrophobic interactions between protein surfaces may lead to aggregation and precipitation of the protein.

Separation on the basis of size

Dialysis and ultrafiltration

Small molecules, such as salts, can be removed from protein solutions by dialysis or ultrafiltration

Dialysis is performed by adding the protein–salt solution to a semipermeable membrane tube (commonly a nitrocellulose or collodion membrane). When the tube is immersed in a dilute buffer solution, small molecules will pass through and large protein molecules will be retained in the tube, depending on the pore size of the dialysis membrane. This procedure is particularly useful for removal of $(NH_4)_2SO_4$ or other salts during protein purification, since the salts will interfere with the purification of proteins by ion exchange chromatography (below). Figure 2.11 illustrates the dialysis of proteins.

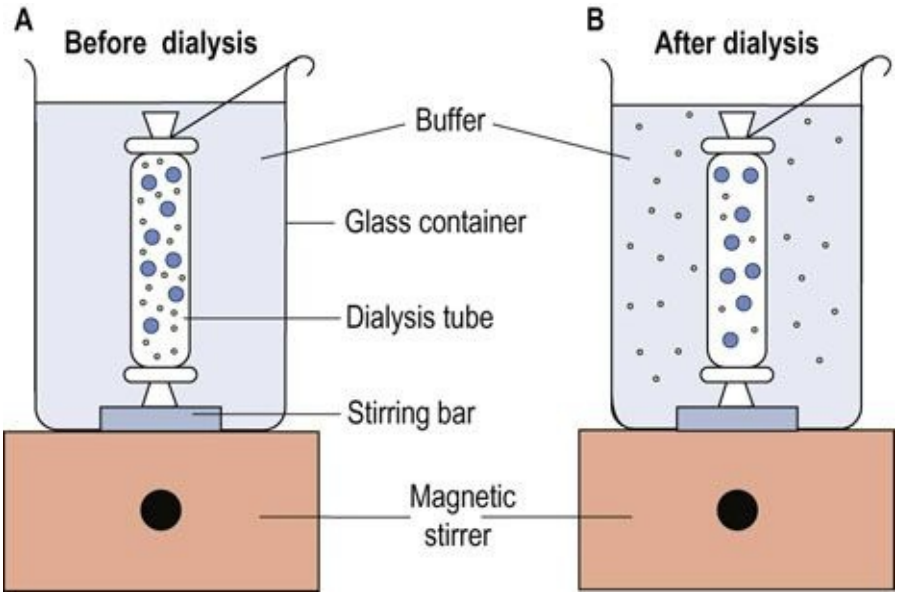


FIG. 2.11 Dialysis of proteins.
Protein and low-molecular-mass compounds are separated by dialysis on the basis of size. (A) A protein solution with salts is placed in a dialysis tube in a beaker and dialyzed with stirring against an appropriate buffer. (B) The protein is retained in the dialysis tube, whereas salts will exchange through the membrane. By use of a large volume of external buffer, with occasional buffer replacement, the protein will eventually be exchanged into the external buffer solution.

Ultrafiltration has largely replaced dialysis for purification of proteins. This technique uses pressure to force a solution through a semipermeable membrane of defined, homogeneous pore size. By selecting the proper molecular weight cut-off value (pore size) for the filter, the membranes will allow solvent and lower molecular weight solutes to permeate the membrane, forming the filtrate, while retaining higher molecular weight proteins in the retentate solution. Ultrafiltration can be used to concentrate protein solutions or to accomplish dialysis by continuous replacement of buffer in the retentate compartment.

Gel filtration (molecular sieving)

Gel filtration chromatography separates proteins on the basis of size

Gel filtration, or gel permeation, chromatography uses a column of insoluble but highly hydrated polymers such as dextrans, agarose or polyacrylamide. Gel filtration chromatography depends on the differential migration of dissolved

solutes through gels that have pores of defined sizes. This technique is frequently used for protein purification and for desalting protein solutions. Figure 2.12 describes the principle of gel filtration. There are commercially available gels made from carbohydrate polymer beads designated as dextran (Sephadex series), polyacrylamide (Bio-Gel P series), and agarose (Sepharose series), respectively. The gels vary in pore size and one can choose the gel filtration materials according to the molecular weight fractionation range desired.

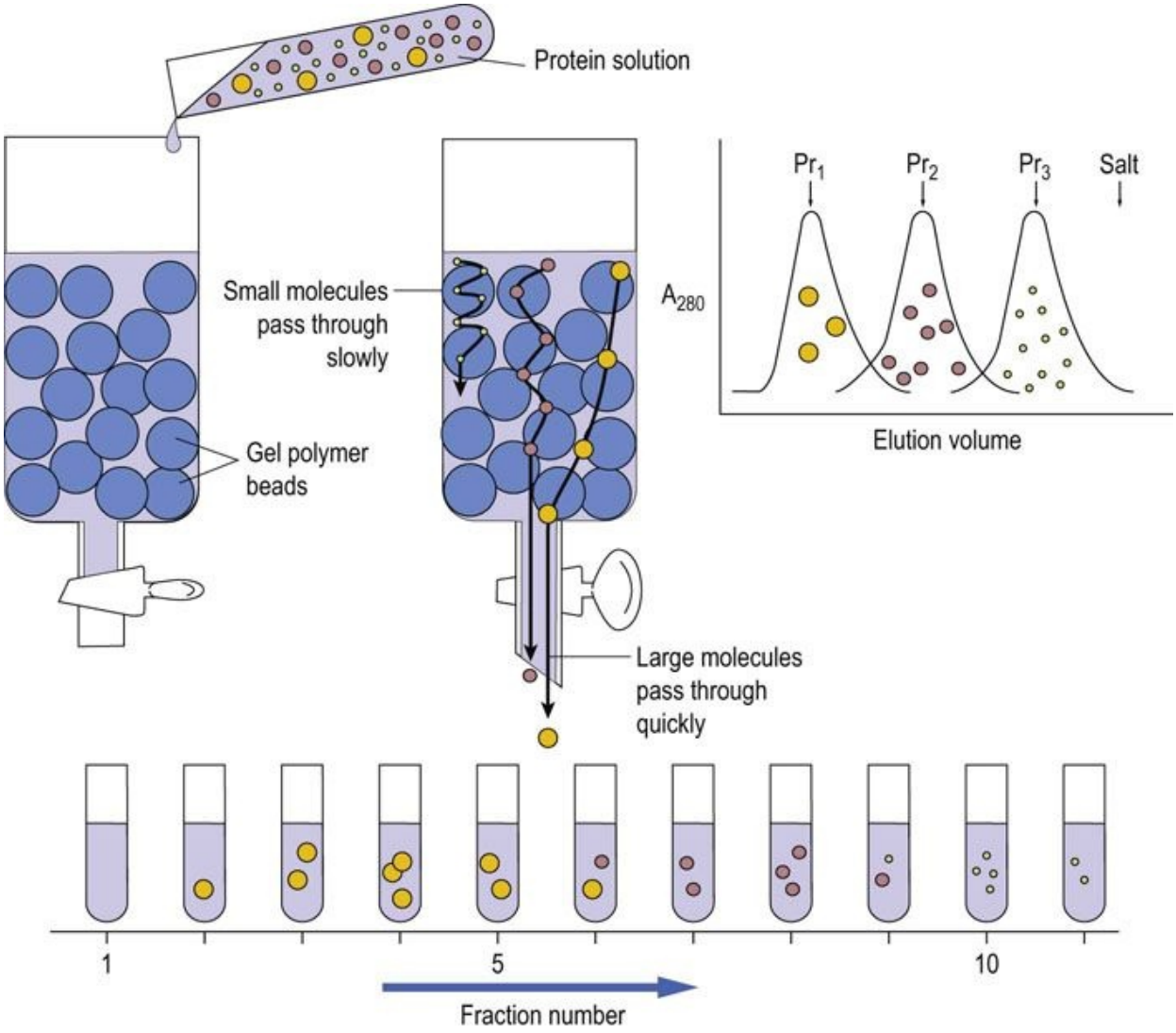


FIG. 2.12 Fractionation of proteins by size: gel filtration chromatography of proteins. Proteins with different molecular sizes are separated by gel filtration based on their relative size. The smaller the protein, the more readily it exchanges into polymer beads,

whereas larger proteins may be completely excluded. Larger molecules flow more rapidly through this column, leading to fractionation on the basis of molecular size. The chromatogram on the right shows a theoretical fractionation of three proteins, Pr_1–Pr_3 of decreasing molecular weight.

Ion exchange chromatography

Proteins bind to ion exchange matrices based on charge-charge interactions

When a charged ion or molecule with one or more positive charges exchanges with another positively charged component bound to a negatively charged immobilized phase, the process is called cation exchange. The inverse process is called anion exchange. The cation exchanger, carboxymethylcellulose (—O—CH_2—COO^-), and anion exchanger, diethylaminoethyl (DEAE) cellulose [—O—C_2H_4—$NH^+(C_2H_5)_2$], are frequently used for the purification of proteins. Consider purifying a protein mixture containing albumin and immunoglobulin. At pH 7.5, albumin, with a pI of 4.8, is negatively charged; immunoglobulin with a pI ~8 is positively charged. If the mixture is applied to a DEAE column at pH 7, the albumin sticks to the positively charged DEAE column whereas the immunoglobulin passes through the column. Figure 2.13 illustrates the principle of ion exchange chromatography. As with gel permeation chromatography, proteins can be separated from one another, based on small differences in their pI. **Adsorbed proteins are commonly eluted with a gradient formed from two or more solutions with different pH and/or salt concentrations.** In this way, proteins are gradually eluted from the column and are well resolved based on their pI.

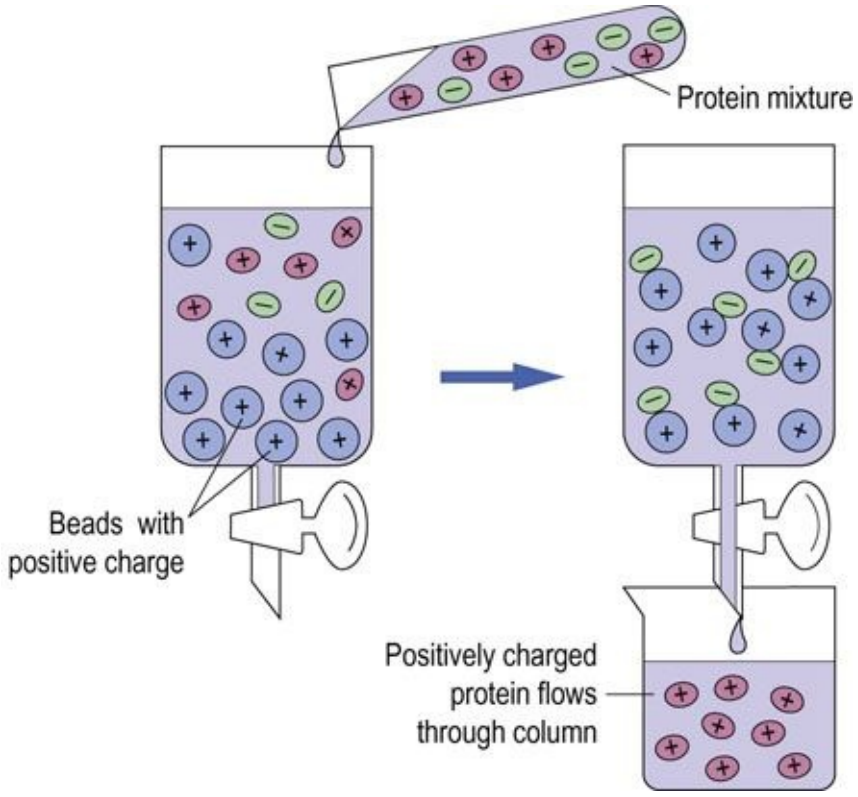


FIG. 2.13 Fractionation of proteins by charge: ion exchange chromatography. Mixtures of proteins can be separated by ion exchange chromatography according to their net charges. Beads that have positively charged groups attached are called anion exchangers, whereas those having negatively charged groups are cation exchangers. This figure depicts an anion exchange column. Negatively charged protein binds to positively charged beads, and positively charged protein flows through the column.

Affinity chromatography

Affinity chromatography purifies proteins based on ligand interactions

Affinity chromatography is a convenient and specific method for purification of proteins. A porous chromatography column matrix is derivatized with a ligand that interacts with, or binds to, a specific protein in a complex mixture. The protein of interest will be selectively and specifically bound to the ligand while the others wash through the column. The bound protein can then be eluted by a high salt concentration, mild denaturation or by a soluble form of the ligand or ligand analogs (see Chapter 6).

Determination of purity and molecular weight of proteins

Polyacrylamide gel electrophoresis in sodium dodecylsulfate can used to separate proteins, based on charge

Electrophoresis can be used for the separation of a wide variety of charged molecules, including amino acids, polypeptides, proteins, and DNA. When a current is applied to molecules in dilute buffers, those with a net negative charge at the selected pH migrate toward the anode and those with a net positive charge toward the cathode. A porous support, such as paper, cellulose acetate or polymeric gel, is commonly used to minimize diffusion and convection.

Like chromatography, electrophoresis may be used for preparative fractionation of proteins at physiologic pH. Different soluble proteins will move at different rates in the electrical field, depending on their charge-to-mass ratio. A denaturing detergent, sodium dodecyl sulfate (SDS), is commonly used in a polyacrylamide gel electrophoresis (PAGE) system to separate and resolve protein subunits according to molecular weight. The protein preparation is usually treated with both SDS and a thiol reagent, such as β-mercaptoethanol, to reduce disulfide bonds. Because the binding of SDS is proportional to the length of the peptide chain, each protein molecule has the same mass-to-charge ratio and the relative mobility of the protein is proportional to the molecular mass of the polypeptide chain. Varying the state of crosslinking of the polyacrylamide gel provides selectivity for proteins of different molecular weights. A purified protein preparation can be readily analyzed for homogeneity on SDS-PAGE by staining with sensitive and specific dyes, such as Coomassie Blue, or with a silver staining technique, as shown in Figure 2.14.

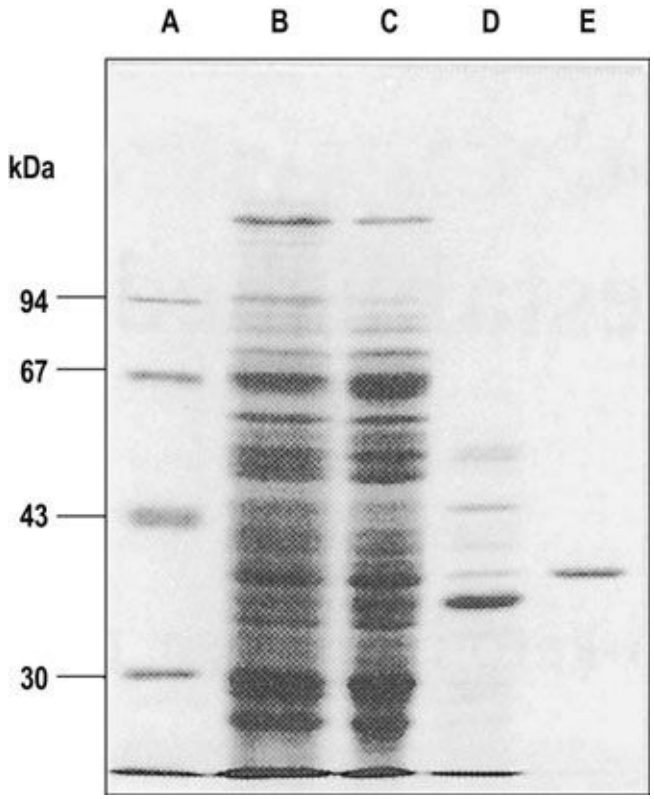


FIG. 2.14 SDS-PAGE.
Sodium dodecylsulfate-polyacrylamide gel electrophoresis is used to separate proteins on the basis of their molecular weights. Larger molecules are retarded in the gel matrix, whereas the smaller ones move more rapidly. Lane **A** contains standard proteins with known molecular masses (indicated in kDa on the left). Lanes **B, C, D,** and **E** show results of SDS-PAGE analysis of a protein at various stages in purification: **B** = total protein isolate; **C** = ammonium sulfate precipitate; **D** = fraction from gel permeation chromatography; **E** = purified protein from ion exchange chromatography.

Advanced concept box High-performance liquid chromatography (HPLC)

HPLC is a powerful chromatographic technique for high-resolution separation of proteins, peptides, and amino acids. The principle of the separation may be based on the charge, size or hydrophobicity of proteins. The narrow columns are packed with a noncompressible matrix of fine silica beads coated with a thin layer of a stationary phase. A protein mixture is applied to the column, and then the components are eluted by either isocratic or

gradient chromatography. The eluates are monitored by ultraviolet absorption, refractive index or fluorescence. This technique gives high-resolution separation.

Isoelectric focusing (IEF)

Isoelectric focusing resolves proteins based on their isoelectric point

Isoelectric focusing (IEF) is conducted in a microchannel or gel containing a stabilized pH gradient. A protein applied to the system will be either positively or negatively charged, depending on its amino acid composition and the ambient pH. Upon application of a current, the protein will move towards either the anode or cathode until it encounters that part of the system which corresponds to its pI, where the protein has no charge and will cease to migrate. **IEF is used in conjunction with SDS-PAGE for two-dimensional gel electrophoresis** (Fig. 2.15). This technique is particularly useful for the fractionation of complex mixtures of proteins for proteomic analysis.

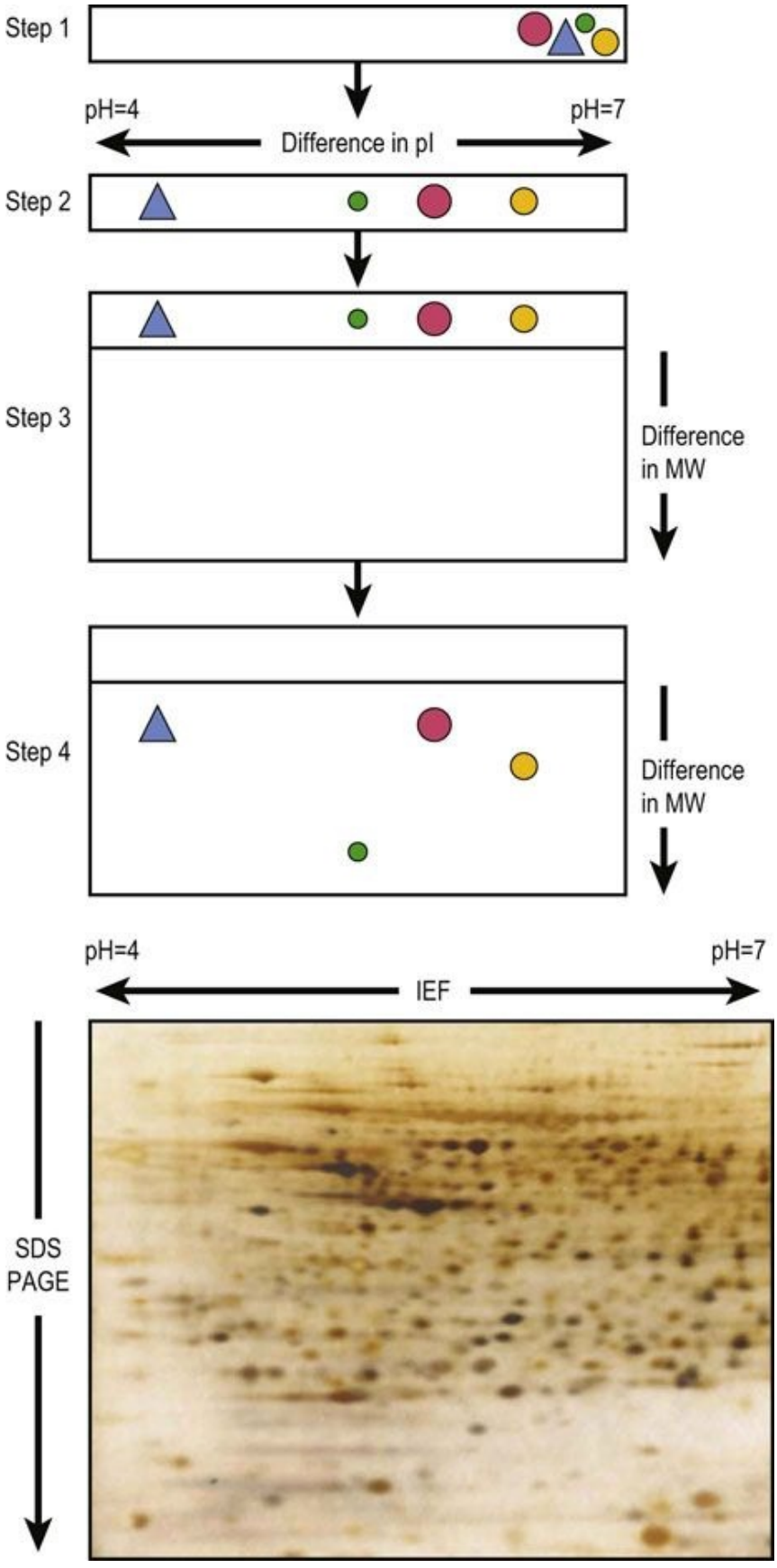

Step 1
pH=4
pH=7
Difference in pI
Step 2
Step 3
Difference
in MW
Step 4
Difference
in MW
pH=4
pH=7
IEF
SDS
PAGE

FIG. 2.15 Two-dimensional gel electrophoresis.
(top) **Step 1:** Sample containing proteins is applied to a cylindrical isoelectric focusing gel within the pH gradient. **Step 2:** Each protein migrates to a position in the gel corresponding to its isoelectric point (pI). **Step 3:** The IEF gel is placed horizontally on the top of a slab gel. **Step 4:** The proteins are separated by SDS-PAGE according to their molecular weight. (bottom) Typical example of 2D-PAGE. A rat liver homogenate was fractionated by 2D-PAGE and proteins were detected by silver staining.

Analysis of protein structure

The typical steps in the purification of a protein are summarized in Figure 2.16. Once purified, for the determination of its amino acid composition, a protein is subjected to hydrolysis, commonly in 6 mol/L HCl at 110°C in a sealed and evacuated tube for 24–48 h. Under these conditions, tryptophan, cysteine and most of the cystine are destroyed, and glutamine and asparagine are quantitatively deaminated to give glutamate and aspartate, respectively. Recovery of serine and threonine is incomplete and decreases with increasing time of hydrolysis.

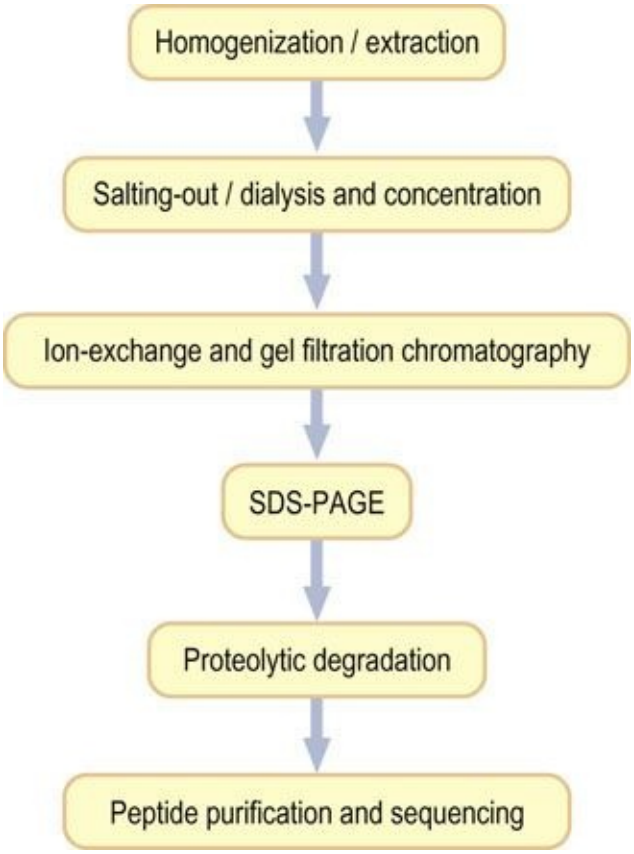


FIG. 2.16 Strategy for protein purification.
Purification of a protein involves a sequence of steps in which contaminating proteins are removed, based on difference in size, charge, and hydrophobicity. Purification is monitored by SDS-PAGE (see Fig. 2.14). The primary sequence of the protein may be determined by automated Edman degradation of peptides (see Fig. 2.18). The three-dimensional structure of the protein may be determined by X-ray crystallography.

Alternative hydrolysis procedures may be used for measurement of tryptophan, while cysteine and cystine may be converted to an acid-stable cysteic acid prior to hydrolysis. Following hydrolysis, the free amino acids are separated on an automated amino acid analyzer using an ion exchange column or, following pre-column derivatization with colored or fluorescent reagents, by reversed-phase high-performance liquid chromatography (HPLC). The free amino acids fractionated by ion exchange chromatography are detected by reaction with a chromogenic or fluorogenic reagent, such as ninhydrin or dansyl chloride, Edman's reagent (see below) or *o*-phthalaldehyde. These techniques allow the measurement of as little as 1 pmol of each amino acid. A typical elution pattern of amino acids in a purified protein is shown in Figure 2.17.

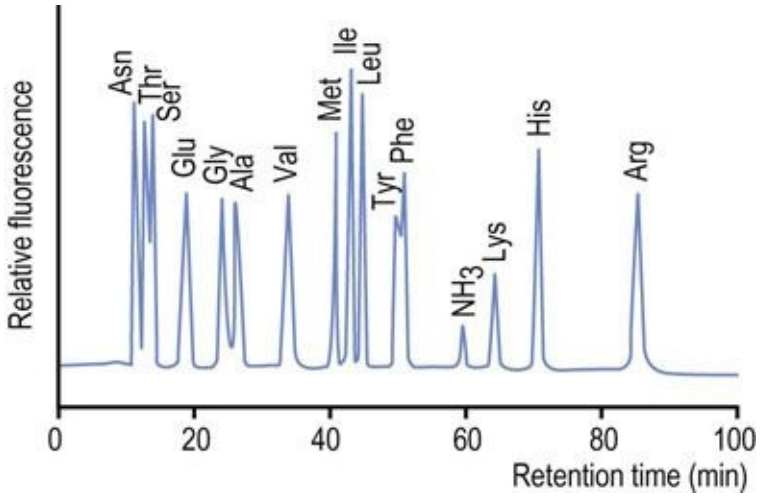


FIG. 2.17 Chromatogram from an amino acid analysis by cation-exchange chromatography.
A protein hydrolysate is applied to the cation exchange column in a dilute buffer at acidic pH (~3.0), at which all amino acids are positively charged. The amino acids are then eluted by a gradient of increasing pH and salt concentrations. The most anionic (acidic) amino acids elute first, followed by the neutral and basic amino acids. Amino acids are derived by post-column reaction with a fluorogenic compound, such as *o*-phthalaldehyde.

Advanced concept box The proteome

A proteome is defined as the full complement of proteins

produced by a particular genome. Changes in cellular and tissue proteomes occur in response to hormonal signaling during development, and environmental stresses. Proteomics is defined as the qualitative and quantitative comparison of proteomes under different conditions. In one approach to analyze the proteome of a cell, proteins are extracted and subjected to two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). Individual protein spots are identified by staining, then extracted and digested with proteases. Small peptides from such a gel are sequenced by mass spectrometry, permitting the identification of the protein. A typical analysis of a rat liver extract is shown in Figure 2.15. In 2D-differential gel electrophoresis (DIGE), two proteomes may be compared by labeling their proteins with different fluorescent dyes, *e.g.* red and green. The labeled proteins are mixed, then fractionated by 2D-PAGE. Proteins present in both proteomes will appear as yellow spots, while unique proteins will be red or green, respectively (see Chapter 36).

Determination of the primary structure of proteins

Historically, analysis of protein sequence was carried out by chemical methods; today, both sequence analysis and protein identification are performed by mass spectrometry

Information on the primary sequence of a protein is essential for understanding its functional properties, the identification of the family to which the protein belongs, as well as characterization of mutant proteins that cause disease. A protein may be cleaved first by digestion by specific endoproteases, such as trypsin (Chapter 6), V8 protease or lysyl endopeptidase, to obtain peptide fragments. Trypsin cleaves peptide bonds on the C-terminal side of arginine and lysine residues, provided the next residue is not proline. Lysyl endopeptidase is also frequently used to cleave at the C-terminal side of lysine. Cleavage by chemical reagents such as cyanogen bromide is also useful. Cyanogen bromide

cleaves on the C-terminal side of methionine residues. Before cleavage, proteins with cysteine and cystine residues are reduced by 2-mercaptoethanol and then treated with iodoacetate to form carboxymethylcysteine residues. This avoids spontaneous formation of inter-or intramolecular disulfides during analyses.

The cleaved peptides are then subjected to reverse-phase HPLC to purify the peptide fragments, and then sequenced on an automated protein sequencer, using the **Edman degradation** technique (Fig. 2.18). The sequence of overlapping peptides is then used to obtain the primary structure of the protein. The Edman degradation technique is largely of historical interest. Mass spectrometry is more commonly used today to obtain both the molecular mass and sequence of polypeptides simultaneously (Chapter 36). Both techniques can be applied directly to proteins or peptides recovered from SDS-PAGE or two-dimensional electrophoresis (IEF plus SDS-PAGE).

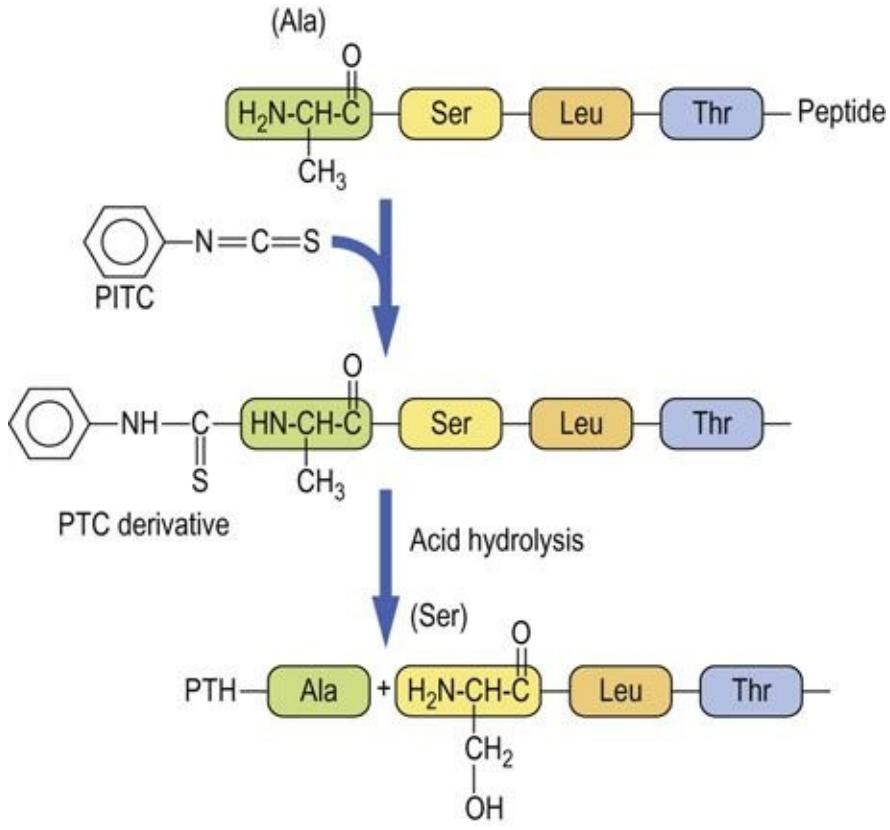


FIG. 2.18 Steps in Edman degradation.
The Edman degradation method sequentially removes one residue at a time from the amino end of a peptide. Phenyl isothiocyanate (PITC) converts the N-terminal amino group of the immobilized peptide to a phenylthiocarbamyl derivative (PTC amino acid) in

alkaline solution. Acid treatment removes the first amino acid as the phenylthiohydantoin (PTH) derivative, which is identified by HPLC.

Protein sequencing and identification can also be done by electrospray ionization liquid chromatography tandem mass spectrometry (HPLC-ESI-MS/MS) (Chapter 36). This technique is sufficiently sensitive that proteins separated by 2D-PAGE (see Fig. 2.15) can be recovered from the gel for analysis. As little as 1 µg of protein per spot, can be digested with trypsin in situ, then extracted from the gel and identified, based on their amino acid sequence. This technique, as well as a complementary technique called matrix-assisted laser desorption ionization-time of flight (MALDI-TOF) MS/MS (Chapter 36), can be applied for determination of the molecular weight of intact proteins, as well as for sequence analysis of peptides, leading to unambiguous identification of a protein.

Determination of the three-dimensional structure of proteins

X-ray crystallography and NMR spectroscopy are usually used for determination of the three-dimensional structure of proteins

X-ray crystallography depends on the diffraction of X-rays by the electrons of the atoms constituting the molecule. However, since the X-ray diffraction caused by an individual molecule is weak, the protein must exist in the form of a well-ordered crystal, in which each molecule has the same conformation in a specific position and orientation on a three-dimensional lattice. Based on diffraction of a collimated beam of electrons, the distribution of the electron density, and thus the location of atoms, in the crystal can be calculated to determine the structure of the protein. For protein crystallization, the most frequently used method is the hanging drop method which involves the use of a simple apparatus that permits a small portion of a protein solution (typically 10 µL droplet containing 0.5–1 mg/protein) to evaporate gradually to reach the saturating point at which the protein begins to crystallize. NMR spectroscopy is usually used for structural analysis of small organic compounds, but high-field NMR is also useful for determination of the structure of a protein in solution and complements information obtained by X-ray crystallography.

Advanced concept box Protein folding

For proteins to function properly, they must fold into the correct shape. Proteins have evolved so that one fold is more favorable than all others – the native state. Numerous proteins assist other proteins in the folding process. These proteins, termed **chaperones,** include 'heat shock' proteins, such as HSP 60 and HSP 70, and protein disulfide isomerases. A protein folding disease is a disease that is associated with abnormal conformation of a protein. This occurs in chronic, age-related diseases, such as Alzheimer's disease, amyotrophic lateral sclerosis, and Parkinson's disease.

Clinical box Creutzfeldt–jakob disease

A 56-year-old male cattle rancher presented with epileptic cramp and dementia and was diagnosed as having Creutzfeldt–Jakob disease, a human prion disease. The **prion diseases**, also known as transmissible spongiform encephalopathies, are neurodegenerative diseases that affect both humans and animals. This disease in sheep and goats is designated as scrapie, and in cows as spongiform encephalopathy (mad cow disease). The diseases are characterized by the accumulation of an abnormal isoform of a host-encoded protein, prion protein-cellular form (PrPC), in affected brains.

Comment.

Prions appear to be composed only of PrPSc (scrapie form) molecules, which are abnormal conformers of the normal, host-encoded protein. PrPC has a high α-helical content and is devoid of β-pleated sheets, whereas PrPSc has a high β-pleated sheet content. The conversion of PrPC into PrPSc involves a profound

conformational change. The progression of infectious prion diseases appears to involve an interaction between PrPC and PrPSc, which induces a conformational change of the α-helix-rich PrPC to the β-pleated sheet-rich conformer of PrPSc. PrPSc-derived prion disease may be genetic or infectious. The amino acid sequences of different mammalian PrPCs are similar, and the conformation of the protein is virtually the same in all mammalian species.

Summary

■ A total of 20 alpha-amino acids are the building blocks of proteins. The side chains of these amino acids contribute charge, polarity and hydrophobicity to protein.
■ Proteins are macromolecules formed by polymerization of L-α-amino acids by peptide bonds. The linear sequence of the amino acids constitutes the primary structure of the protein.
■ Proteins are macromolecules formed by polymerization of L-α-amino acids. There are 20 different amino acids in proteins, linked by peptide bonds. The linear sequence of the amino acids is the primary structure of the protein.
■ The higher-order structure of a protein is the product of its secondary, tertiary, and quaternary structure.
■ These higher order structures are formed by hydrogen bonds, hydrophobic interactions, salt bridges and covalent bonds between the side chains of amino acids.
■ Purification and characterization of proteins are essential for elucidating their structure and function. By taking advantage of differences in their size, solubility, charge and ligand-binding properties, proteins can be purified to homogeneity using various chromatographic and electrophoretic techniques. The molecular mass and purity of a protein, and its subunit composition, can be determined by SDS-PAGE.
■ Deciphering the primary and three-dimensional structures of a protein by chemical methods, mass spectrometry, X-ray analysis and NMR spectroscopy leads to an understanding of structure–function relationships in proteins.

Active learning

1. Mass spectrometry analysis of blood, urine and tissues is now being applied for clinical diagnosis. Discuss the merits of this technique with respect to specificity, sensitivity, through-put and breadth of analysis, including proteomic analysis for diagnostic purposes.
2. Review the importance of protein misfolding and deposition in tissues in age-related chronic diseases.

Further reading

Aguzzi, A, Falsig, J. Prion propagation, toxicity and degradation. *Nat Neurosci*. 2012; 15:936–939.

Dominguez, DC, Lopes, R, Torres, ML. Proteomics: clinical applications. *Clin Lab Sci*. 2007; 20:245–248.

Griffin, MD, Gerrard, JA. The relationship between oligomeric state and protein function. *Adv Exp Med Biol*. 2012; 747:74–90.

Kovacs, GG, Budka, H. Prion diseases: from protein to cell pathology. *Am J Pathol*. 2008; 172:555–565.

Marouga, R, David, S, Hawkins, E. The development of the DIGE system: 2D fluorescence difference gel analysis technology. *Anal Bioanal Chem*. 2005; 382:669–678.

Matt, P, Fu, Z, Ru, Q, Van Eyk, JE. Biomarker discovery: proteome fractionation and separation in biological samples. *J Physiol Genomics*. 2008; 14:12–17.

Shkundina, IS, Ter-Avanesyan, MD. Prions. *Biochemistry (Moscow)*. 2007; 72:1519–1536.

Sułkowska, JI, Rawdon, EJ, Millett, KC, *et al.* Conservation of complex knotting and slipknotting patterns in proteins. *Proc Natl Acad Sci U S A*. 2012; 109:E1715–1723.

Walsh, CT. *Posttranslational modification of proteins: expanding nature's inventory*, ed 3. Colorado: Roberts & Co.; 2007.

Websites

Protein Data Ban. www.rcsb.org [– Use Search Box, then select a structure and view protein in Jmol].

www.ncbi.nlm.nih.gov/Structure [– National Center for Biotechnology Information, National Library of Medicine. Several databases, including protein structure].

http://us.expasy.org [– Bioinformatics resource portal].

CHAPTER 3

Carbohydrates and Lipids

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

- Describe the structure and nomenclature of carbohydrates.
- Identify the major carbohydrates in the human body and in our diet.
- Distinguish between reducing and nonreducing sugars.
- Describe various types of glycosidic bonds in oligosaccharides and polysaccharides.
- Identify the major classes of lipids in the human body and in our diet.
- Describe the types of bonds in lipids and their sensitivity to saponification.
- Explain the general role of triglycerides, phospholipids and glycolipids in the body.
- Outline the general features of the fluid mosaic model of the structure of biological membranes.

Introduction

Carbohydrates and lipids are major sources of energy and are stored in the body as glycogen and triglycerides

This chapter describes the structure of carbohydrates and lipids found in the diet and in tissues. These two classes of compounds differ significantly in physical and chemical properties. Carbohydrates are hydrophilic; the smaller carbohydrates, such as milk sugar and table sugar, are soluble in aqueous solution, while polymers such as starch or cellulose form colloidal dispersions or are insoluble. Lipids vary in size, but rarely exceed 2 kDa in molecular mass; they are insoluble in water but soluble in organic solvents. Both carbohydrates and lipids may be bound to proteins and have important structural and regulatory functions, which are elaborated in later chapters. This chapter ends with a description of the **fluid mosaic model** of biological membranes, illustrating how protein, carbohydrates and lipids are integrated into the structure of biological membranes that surround the cell and intracellular compartments.

Carbohydrates

Nomenclature and structure of simple sugars

The classic definition of a carbohydrate is a polyhydroxy aldehyde or ketone

The simplest carbohydrates, having two hydroxyl groups, are glyceraldehyde and dihydroxyacetone (Fig. 3.1). These three-carbon sugars are trioses; the suffix ‘ose’ designates a sugar. Glyceraldehyde is an **aldose**, and dihydroxyacetone a **ketose** sugar. Prefixes and examples of longer-chain sugars are shown in Table 3.1.

Table 3.1

Classification of carbohydrates by length of the carbon chain

Number of carbons	Name	Examples in human biology
Three	Triose	Glyceraldehyde, dihydroxyacetone
Four	Tetrose	Erythrose
Five	Pentose	Ribose, ribulose*, xylose, xylulose*, deoxyribose
Six	Hexose	Glucose, mannose, galactose, fucose, fructose
Seven	Heptose	Sedoheptulose*
Eight	Octose	None
Nine	Nonose	Neuraminic (sialic) acid

*The syllable ‘ul’ indicates that a sugar is ketose; the formal name for fructose would be ‘gluculose’. As with fructose, the keto group is located at C-2 of the sugar, and the remaining carbons have the same geometry as the parent sugar.

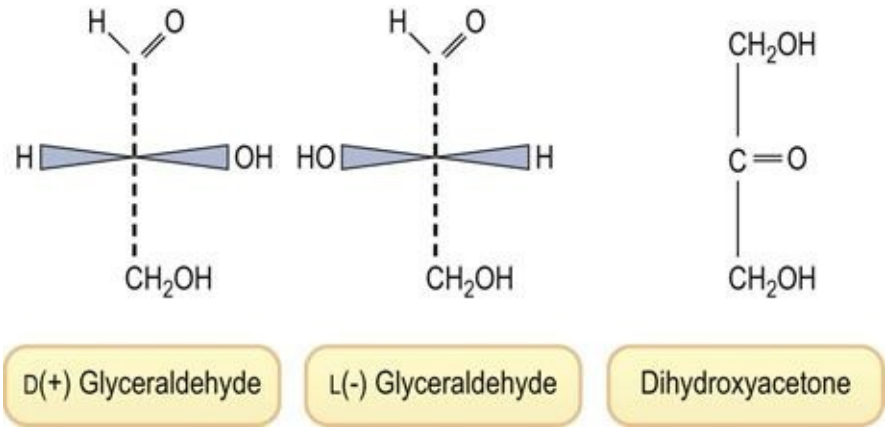


FIG. 3.1 Structures of the trioses: D- and L-glyceraldehyde (aldoses) and dihydroxyacetone (a ketose).

Numbering of the carbons begins from the end containing the aldehyde or ketone functional group. Sugars are classified into the D or L family, based on the configuration around the highest numbered asymmetric center (Fig. 3.2). In contrast to the L-amino acids, nearly all sugars found in the body have the D configuration.

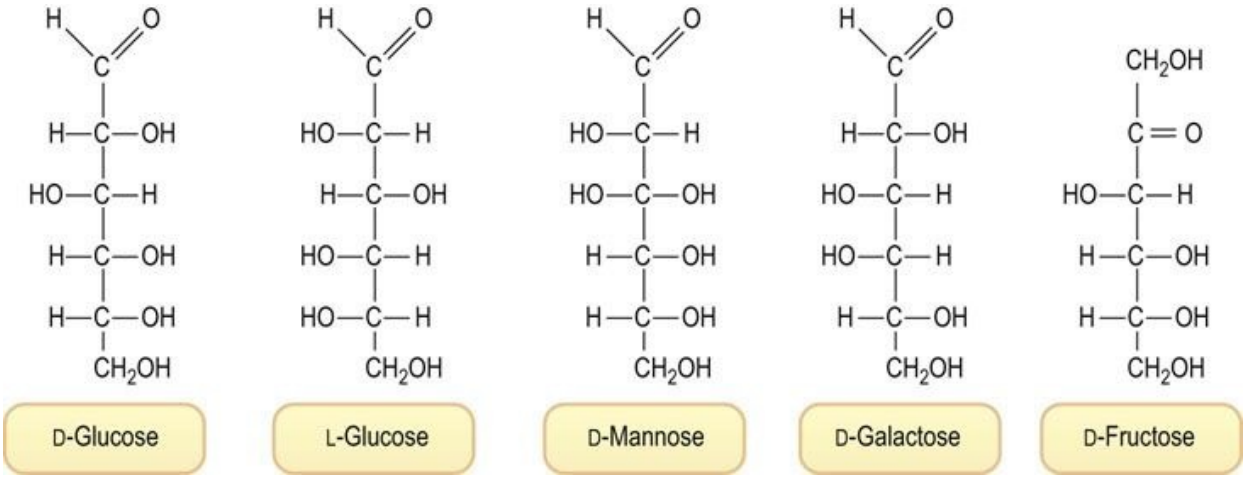


FIG. 3.2 Structures of hexoses: D- and L-glucose, D-mannose, D-galactose and D-fructose.
The D and L designations are based on the configuration at the highest numbered asymmetric center, C-5 in the case of hexoses. Note that L-glucose is the mirror image of D-glucose, *i.e.* the geometry at all of the asymmetric centers is reversed. Mannose is the C-2 epimer, and galactose the C-4 epimer of glucose. These linear projections of carbohydrate structures are known as Fischer projections.

An aldohexose, such as glucose, contains four asymmetric centers, so that there are 16 (2^4) possible stereoisomers, depending on whether each of the four carbons has the D or L configuration (see Fig. 3.2). Eight of these aldohexoses are D-sugars. Only three of these are found in significant amounts in the body: glucose (blood sugar), mannose and galactose (see Fig. 3.2). Similarly, there are four possible epimeric D-ketohexoses; fructose (fruit sugar) (see Fig. 3.2) is the only ketohexose present at significant concentration in our diet or in the body.

Because of their asymmetric centers, sugars are optically active compounds. The rotation of plane polarized light may be dextrorotatory (+) or levorotatory (−). This designation is also commonly included in the name of the sugar; thus

D(+)-glucose or D(−)-fructose indicates that the D form of glucose is dextrorotatory, while the D form of fructose is levorotatory.

Cyclization of sugars

The linear sugar structures shown in Figure 3.2 imply that aldose sugars have a chemically reactive, easily oxidizable, electrophilic, aldehyde residue. Aldehydes such as formaldehyde or glutaraldehyde react rapidly with amino groups in protein to form Schiff base (imine) adducts and crosslinks during fixation of tissues. However, glucose is relatively resistant to oxidation and does not react rapidly with protein. As shown in Figure 3.3, glucose exists largely in nonreactive, inert, cyclic hemiacetal conformations, 99.99% in aqueous solution at pH 7.4 and 37°C. Of all the D-sugars in the world, D-glucose exists to the greatest extent in these cyclic conformations, making it the least oxidizable and least reactive with protein. It has been proposed that the relative chemical inertness of glucose is the reason for its evolutionary selection as blood sugar.

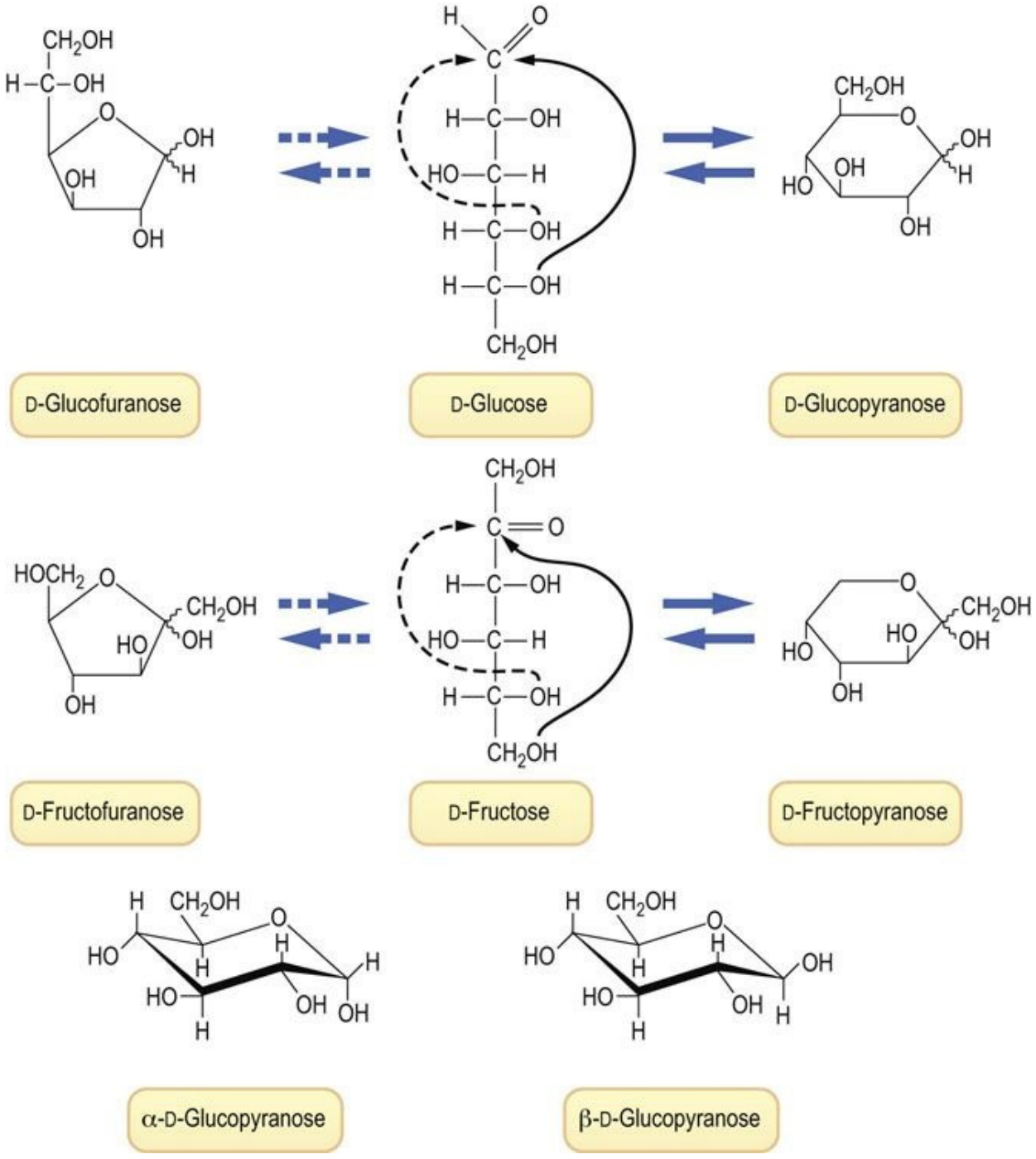
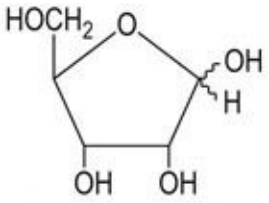


FIG. 3.3 Linear and cyclic representations of glucose and fructose.
(Top) There are four cyclic forms of glucose, in equilibrium with the linear form: α-and β-glucopyranose and α-and β-glucofuranose. The pyranose forms account for over 99% of total glucose in solution. These cyclic conformations are known as Haworth projections; by convention, groups to the right in Fischer projections are shown below the ring, and groups to the left, above the ring. The squiggly bonds to H and OH from *C*-1, the anomeric carbon, indicate indeterminate geometry and represent either the α or the β anomer. (Middle) The linear and cyclic forms of fructose. The ratio of pyranose : furanose

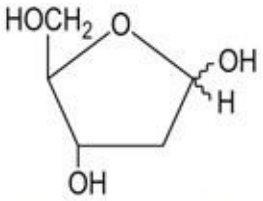
forms of fructose in aqueous solution is ~3 : 1. The ratio shifts as a function of temperature, pH, salt concentration and other factors. (Bottom) Stereochemical representations of the chair forms of α-and β-glucopyranose. The preferred structure in solution, β-glucopyranose, has all of the hydroxyl groups, including the anomeric hydroxyl group, in equatorial positions around the ring, minimizing steric interactions.

When glucose cyclizes to a hemiacetal, it may form a **furanose** or **pyranose** ring structure, named after the 5-and 6-carbon cyclic ethers, furan and pyran (see Fig. 3.3). Note that the cyclization reaction creates a new asymmetric center at *C*-1, which is known as the **anomeric carbon**. The preferred conformation for glucose is the β-anomer (~65%) in which the hydroxyl group on *C*-1 is oriented equatorial to the ring. The β-anomer is the most stable form of glucose because all of the hydroxyl groups, which are bulkier than hydrogen, are oriented equatorially, in the plane of the ring. The α-and β-anomers of glucose can be isolated in pure form by selective crystallization from aqueous and organic solvents. They have different optical rotations, but equilibrate over a period of hours in aqueous solution to form the equilibrium mixture of 65 : 35 β : α anomer. These differences in structure may seem unimportant, but in fact some metabolic pathways use one anomer but not the other, and vice versa. Similarly, while the fructopyranose conformations are the primary forms of fructose in aqueous solution, most of fructose metabolism proceeds from the furanose conformation.

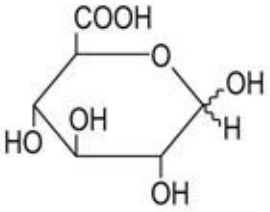
In addition to the basic sugar structures discussed above, a number of other common sugar structures are presented in Figure 3.4. These sugars, deoxysugars, aminosugars and sugar acids, are found primarily in oligosaccharide or polymeric structures in the body, *e.g.* ribose in RNA and deoxyribose in DNA, or they may be attached to proteins or lipids to form glycoconjugates (glycoproteins or glycolipids, respectively). **Glucose is the only sugar found to a significant extent as a free sugar (blood sugar) in the body**.



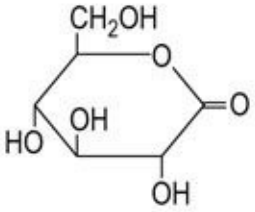
Ribose



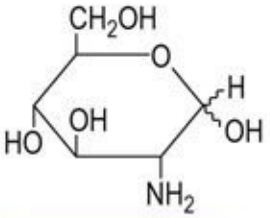
2-Deoxyribose



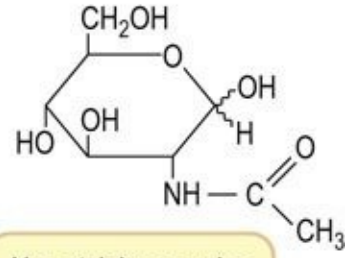
Glucuronic acid



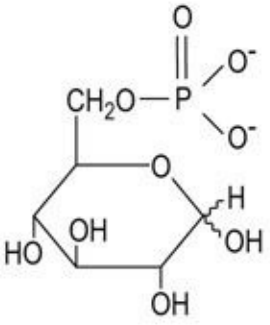
Gluconic acid
(lactone form)



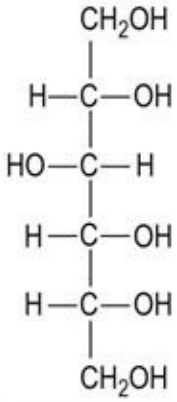
Glucosamine



N-acetylglucosamine



Glucose-
6-phosphate



Sorbitol

FIG. 3.4 Examples of various types of sugars found in human tissues.
Ribose, the pentose sugar in ribonucleic acid (RNA); 2-deoxyribose, the deoxypentose in DNA; glucuronic acid, an acidic sugar formed by oxidation of *C*-6 of glucose; gluconic acid, an acidic sugar formed by oxidation of *C*-1 of glucose, shown in the δ-lactone form; glucosamine, an amino sugar; *N*-acetylglucosamine, an acetylated amino sugar; glucose-6-phosphate, a phosphate ester of glucose, an intermediate in glucose metabolism; sorbitol, a polyol formed on reduction of glucose.

Disaccharides, oligosaccharides and polysaccharides

Sugars are linked to one another by glycosidic bonds to form complex glycans

Carbohydrates are commonly coupled to one another by glycosidic bonds to form disaccharides, trisaccharides, oligosaccharides and polysaccharides. Saccharides composed of a single sugar are termed homoglycans, while saccharides with complex composition are termed heteroglycans. The name of the more complex structures includes not only the name of the component sugars but also the ring conformation of the sugars, the anomeric configuration of the linkage between sugars, the site of attachment of one sugar to another, and the nature of the atom involved in the linkage, usually an oxygen or *O*-glycosidic bond, sometimes a nitrogen or *N*-glycosidic bond. Figure 3.5 shows the structure of several common disaccharides in our diet: **lactose** (milk sugar), **sucrose** (table sugar), **maltose** and isomaltose, which are products of digestion of starch, cellobiose, which is obtained on hydrolysis of cellulose, and **hyaluronic acid**.

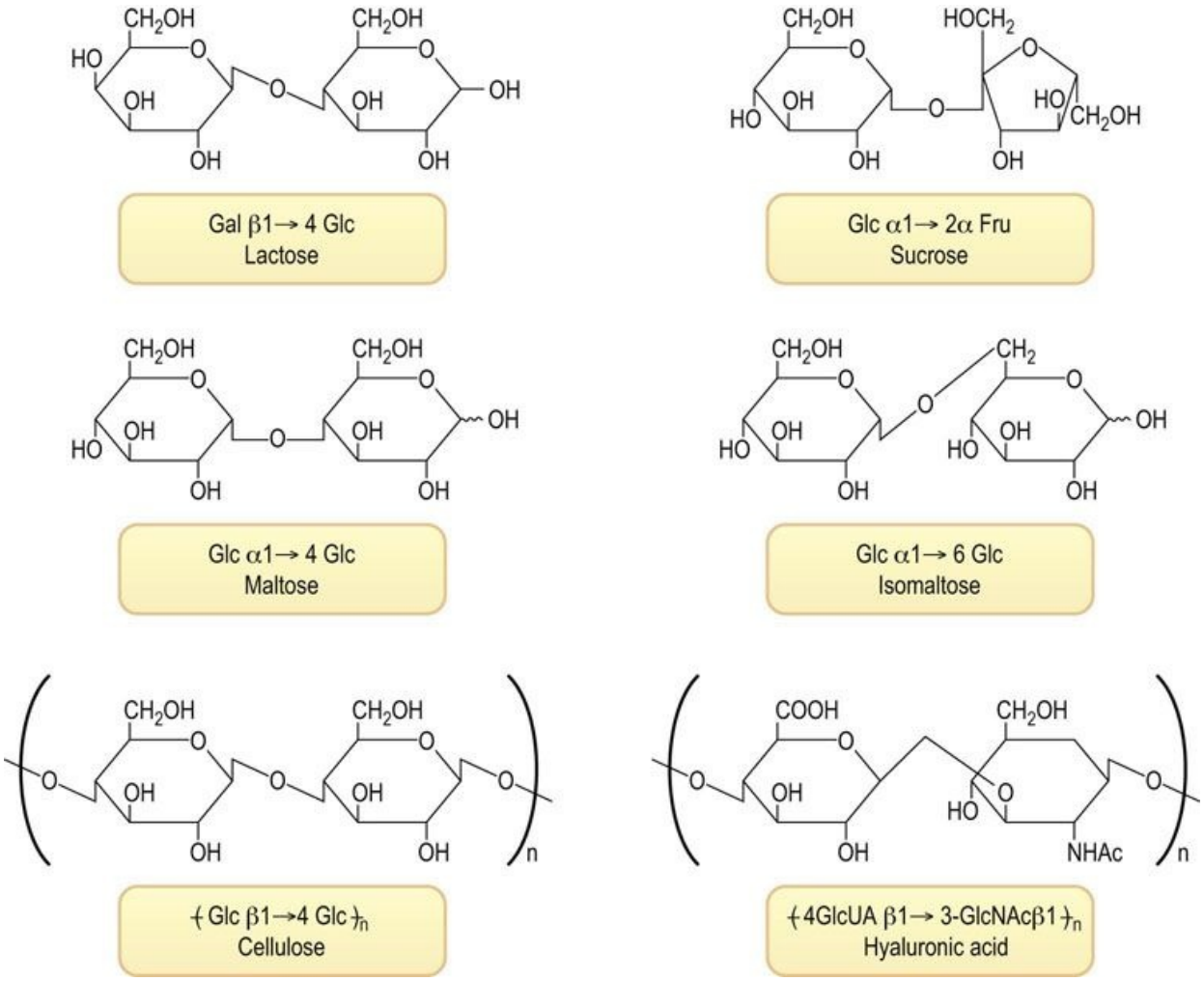


FIG. 3.5 Structures of common disaccharides and polysaccharides. Lactose (milk sugar); sucrose (table sugar); maltose and isomaltose, disaccharides formed on degradation of starch; and repeating disaccharide units of cellulose (from wood) and hyaluronic acid (from vertebral disks). Fru, fructose; Gal, galactose; Glc, glucose; GlcNAc, *N*-acetylglucosamine; GlcUA, glucuronic acid.

Differences in linkage of sugars make a big difference in metabolism and nutrition

Amylose, a component of **starch**, is an α-1→4-linked linear glucan, while **cellulose** is a β-1→4-linked linear glucan. These two polysaccharides differ only in the anomeric linkage between glucose subunits, but they are very different molecules. Starch is soluble in water, cellulose is insoluble; starch is pasty, cellulose is fibrous; starch is digestible, while cellulose is indigestible by humans; starch is a food, rich in calories, while cellulose is roughage.

Advanced concept box The information content of complex glycans

Sugars are attached to each other in **glycosidic linkages** between hemiacetal carbon of one sugar and a hydroxyl group of another sugar. Two glucose residues can be linked in many different linkages (i.e. α1,2; α1,3; α1,4; α1,6; β1,2; β1,3; β1,4; β1,6; α,α1,1; α,β1,1; β,β1,1) to give 11 different disaccharides, each with different chemical and biological properties. Two different sugars, such as glucose and galactose, can be linked either glucose → galactose or galactose → glucose and these two disaccharides can have a total of 20 different isomers.

In contrast, two identical amino acids, such as two alanines, can only form one dipeptide, alanyl-alanine. And two different amino acids, *i.e.* alanine and glycine, can only form two dipeptides, alanyl-glycine and glycyl-alanine. As a result, sugars have the potential to provide a great deal of chemical information. As outlined in Chapters 27–29, carbohydrates bound to proteins and lipids in cell membranes can serve as recognition signals for both cell–cell and cell–pathogen interactions.

Clinical test box Reducing sugar assay for blood glucose

The original assays for blood glucose measured the reducing activity of blood.

These assays work because glucose, at 5 mM concentration, is the major reducing substance in blood. The Fehling and Benedict assays use alkaline cupric salt solutions. With heating, the glucose decomposes oxidatively, yielding a complex mixture of organic acids and aldehydes. Oxidation of the sugar reduces cupric ion (blue-green color) to cuprous ion (orange-red color) in solution.

The color yield produced is directly proportional to the glucose content of the sample.

Reducing sugar assays do not distinguish glucose from other reducing sugars, such as fructose or galactose. In diseases of fructose and galactose metabolism, such as hereditary fructose intolerance or galactosemia (Chapter 27), these assays could yield positive results, creating the false impression of diabetes.

Lipids

Lipids are found primarily in three compartments in the body: plasma, adipose tissue and biological membranes

This introduction will focus on the structure of **fatty acids** (the simplest form of lipids, found primarily in plasma), **triglycerides** (the storage form of lipids, found primarily in adipose tissue), and **phospholipids** (the major class of membrane lipids in all cells). Steroids, such as cholesterol, and (glyco)sphingolipids will be mentioned in the context of biological membranes, but these lipids and others, such as the eicosanoids, will be addressed in detail in later chapters.

Fatty acids

Fatty acids exist in free form and as components of more complex lipids

As summarized in Table 3.2, they are long, straight-chain alkanoic acids, most commonly with 16 or 18 carbons. They may be saturated or unsaturated, the latter containing 1–5 double bonds, all in *cis* geometry. The double bonds are not conjugated, but separated by methylene groups.

Table 3.2

Structure and melting point of naturally occurring fatty acids in the body

Carbon atoms		Chemical formula	Systematic name	Common name	Melting point (°C)
Saturated fatty acids					
12	12:0	$CH_3(CH_2)_{10}COOH$	*n*-dodecanoic	Lauric	44
14	14:0	$CH_3(CH_2)_{12}COOH$	*n*-tetradecanoic	Myristic	54
16	16:0	$CH_3(CH_2)_{14}COOH$	*n*-hexadecanoic	Palmitic	63
18	18:0	$CH_3(CH_2)_{16}COOH$	*n*-octadecanoic	Stearic	70
20	20:0	$CH_3(CH_2)_{18}COOH$	*n*-eicosanoic	Arachidic	77
Unsaturated fatty acids					
16	16:1; ω-7, Δ^9	$CH_3(CH_2)_5CH{=}CH(CH_2)_7COOH$		Palmitoleic	–0.5
18	18:1; ω-9, Δ^9	$CH_3(CH_2)_7CH{=}CH(CH_2)_7COOH$		Oleic	13
18	18:2; ω-6, $\Delta^{9,12}$	$CH_3(CH_2)_4CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$		Linoleic	–5
18	18:3; ω-3, $\Delta^{9,12,15}$	$CH_3CH_2CH{=}CHCH_2CH{=}CHCH_2CH{=}CH(CH_2)_7COOH$		Linolenic	–11
20	20:4; ω-6, $\Delta^{5,8,11,14}$	$CH_3(CH_2)_4CH{=}CHCH_2CH{=}CHCH_2CH{=}CHCH_2CH{=}CH(CH_2)_3COOH$		Arachidonic	–50

For unsaturated fatty acids, the ω designation indicates the location of the first double bond from the methyl end of the molecule; the Δ superscripts indicate the location of the double bonds from the carboxyl end of the molecule. Unsaturated fatty acids account for about two-thirds of all fatty acids in the body; oleate and palmitate account for about one half and one quarter of total fatty acids.

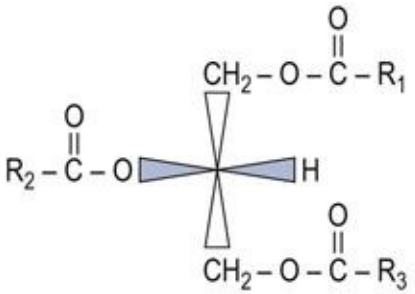
Fatty acids with a single double bond are described as monounsaturated, while those with two or more double bonds are described as polyunsaturated fatty acids. The polyunsaturated fatty acids are commonly classified into two groups, **ω-3 and ω-6 fatty acids**, depending on whether the first double bond appears three or six carbons from the terminal methyl group. The melting point of fatty acids, as well as that of more complex lipids, increases with the chain length of the fatty acid, but decreases with the number of double bonds. The ***cis*-double bonds** place a kink in the linear structure of the fatty acid chain, interfering with close packing, therefore requiring a lower temperature for freezing, *i.e.* they have a lower melting point.

Triacylglycerols (triglycerides)

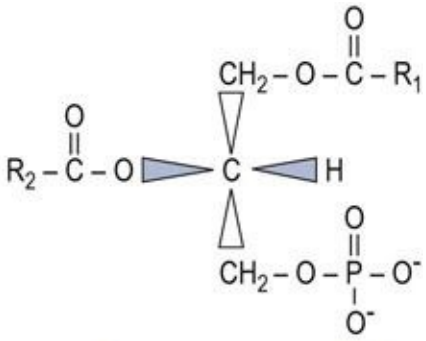
Triglycerides are the storage form of lipids in adipose tissue

Fatty acids in plant and animal tissues are commonly esterified to glycerol, forming a triacylglycerol (triglyceride) (Fig. 3.6), either oils (liquid) or fats (solid). In humans, triglycerides are stored in solid form (fat) in adipose tissue. They are degraded to glycerol and fatty acids in response to hormonal signals, then released into plasma for metabolism in other tissues, primarily muscle and liver. The ester bond of triglycerides and other glycerolipids is also readily

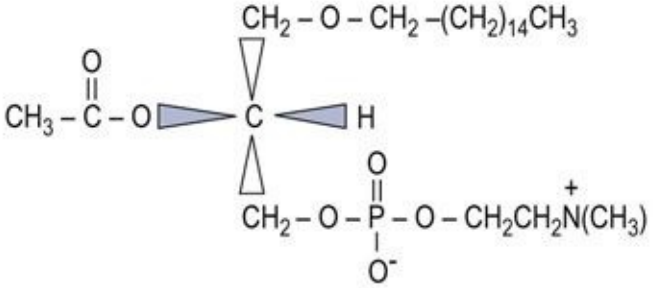
hydrolyzed ex vivo by a strong base, such as NaOH, forming glycerol and free fatty acids. This process is known as **saponification**; one of the products, the sodium salt of the fatty acid, is soap.



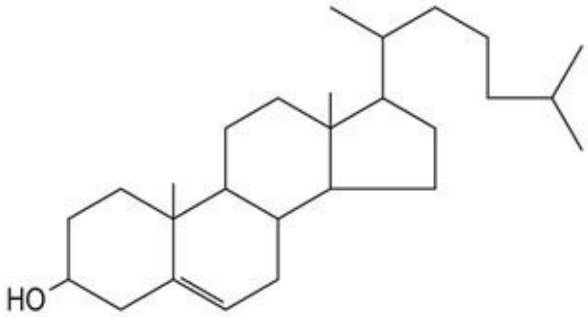
Triglyceride



Phosphatidic acid



Platelet-activating factor (PAF)



Cholesterol

FIG. 3.6 Structure of four lipids with significantly different biological functions. Triglycerides are storage fats. Phosphatidic acid is a metabolic precursor of both triglycerides and phospholipids (see Fig. 3.7). Platelet-activating factor, a mediator of inflammation, is an unusual phospholipid, with a lipid alcohol rather than an esterified lipid at the *sn*-1 position, an acetyl group at *sn*-2, and phosphorylcholine esterified at the *sn*-3 position. Cholesterol is less polar than phospholipids; the hydroxyl group tends to be on the membrane surface, while the polycyclic system intercalates between the fatty acid chains of phospholipids.

Glycerol itself does not have a chiral carbon, but the numbering is standardized using the stereochemical numbering (*sn*) system, which places the hydroxyl group of C-2 on the left; thus all glycerolipids are derived from L-glycerol (see Fig. 3.6). Triglycerides isolated from natural sources are not pure compounds, but mixtures of molecules with different fatty acid composition, *e.g.* 1-palmitoyl, 2-oleyl, 3-linoleoyl-L-glycerol, where the distribution and type of fatty acids vary from molecule to molecule.

Advanced concept box Butter or margarine?

There is continuing debate among nutritionists about the health benefits of butter versus margarine in foods.

Butter is rich in both cholesterol and triglycerides containing saturated fatty acids, which are dietary risk factors for atherosclerosis. Margarine contains no cholesterol and is richer in unsaturated fatty acids.

However, the unsaturated fatty acids in margarine are mostly the unnatural *trans*-fatty acids formed during the partial hydrogenation of vegetable oils. *Trans*-fatty acids affect plasma lipids in the same fashion as saturated fatty acids, suggesting that there are comparable risks associated with the consumption of butter or margarine. The resolution of this issue is complicated by the fact that various forms of margarine, for example soft-spread and hard-block types, vary significantly in their content of *trans*-fatty acids. Partially hydrogenated oils are more stable than the natural oils during heating; when used for deep-frying, they need to be changed less frequently. Despite the additional expense, the food and food-service industries have gradually shifted to the use of natural oils, rich in unsaturated fatty acids and without *trans*-fatty

acids, for cooking and baking.

Phospholipids

Phospholipids are the major lipids in biological membranes

Phospholipids are polar lipids derived from phosphatidic acid (1,2-diacyl-glycerol-3-phosphate) (see Fig. 3.6). Like triglycerides, the glycerophospholipids contain a spectrum of fatty acids at the *sn*-1 and *sn*-2 position, but the *sn*-3 position is occupied by phosphate esterified to an amino compound. The phosphate acts as a bridging diester, linking the diacylglyceride to a polar, nitrogenous compound, most frequently choline, ethanolamine or serine (Fig. 3.7). Phosphatidylcholine (**lecithin**), for example, usually contains palmitic acid or stearic acid at its *sn*-1 position and an 18-carbon, unsaturated fatty acid (e.g. oleic, linoleic or linolenic) at its *sn*-2 position. Phosphatidylethanolamine (cephalin) usually has a longer-chain polyunsaturated fatty acid at the *sn*-2 position, such as arachidonic acid. These complex lipids contribute charge to the membrane: phosphatidylcholine and phosphatidylethanolamine are zwitterionic at physiologic pH and have no net charge, while phosphatidylserine and phosphatidylinositol are anionic. A number of other phospholipid structures with special functions will be introduced in later chapters.

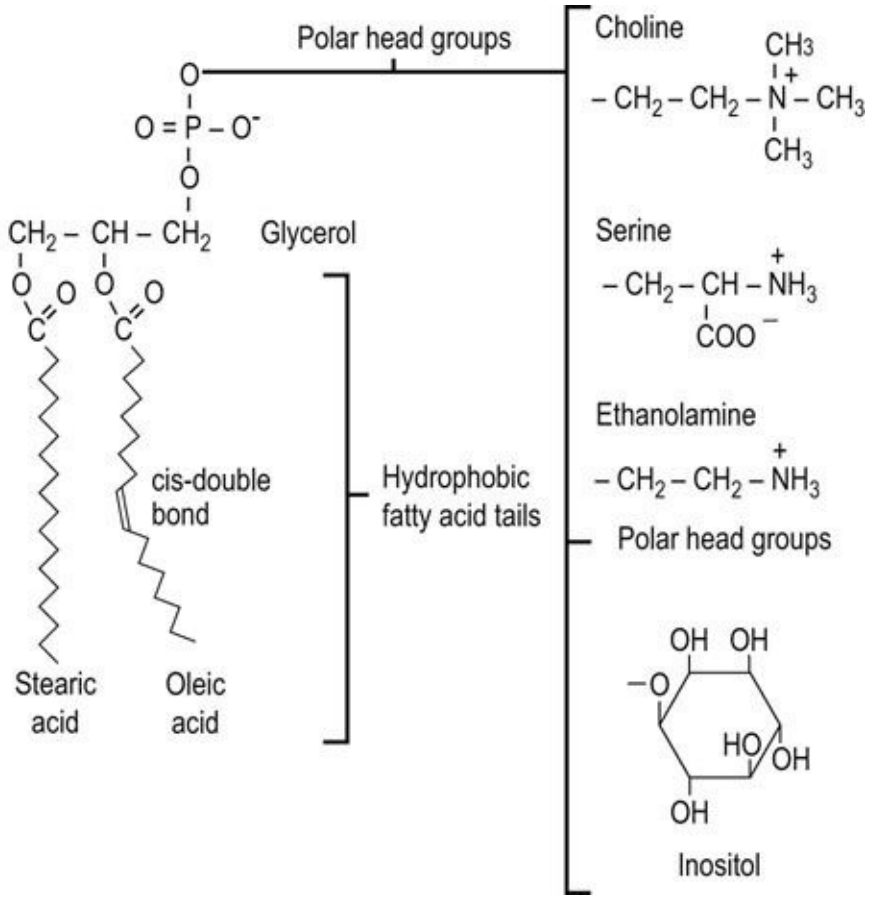


FIG. 3.7 Structure of the major phospholipids of animal cell membranes. Phosphatidylcholine, phosphatidylserine, phosphatidylethanolamine and phosphatidylinositol (see also Chapter 28).

When dispersed in aqueous solution, phospholipids spontaneously form lamellar structures and, under suitable conditions, they organize into extended bilayer structures – not only lamellar structures but also closed vesicular structures termed **liposomes**. The liposome is a model for the structure of a biological membrane, a bilayer of polar lipids with the polar faces exposed to the aqueous environment and the fatty acid side chains buried in the oily, hydrophobic interior of the membrane. The liposomal surface membrane, like its component phospholipids, is a pliant, mobile and flexible structure at body temperature.

Biological membranes also contain another important amphipathic lipid, cholesterol, a flat, rigid hydrophobic molecule with a polar hydroxyl group (see Fig. 3.6). Cholesterol is found in all biomembranes and acts as a modulator of membrane fluidity. At lower temperatures it interferes with fatty acid chain associations and increases fluidity, but at higher temperatures it tends to limit

disorder and decrease fluidity. Cholesterol–phospholipid mixtures have properties intermediate between the gel and liquid crystalline states of the pure phospholipids; they form stable but supple membrane structures.

Advanced concept box Platelet-activating factor and hypersensitivity

Platelet-activating factor (PAF; see Fig. 3.6.) contains an acetyl group at C-2 of glycerol and a saturated 18-carbon alkyl ether group linked to the hydroxyl group at C-1, rather than the usual long-chain fatty acids of phosphatidylcholine. It is a major mediator of hypersensitivity reactions, acute inflammatory reactions, and anaphylactic shock, and affects the permeability properties of membranes, increasing platelet aggregation and causing cardiovascular and pulmonary changes, including edema and hypotension.

In allergic persons, cells involved in the immune response become coated with immunoglobulin E (IgE) molecules that are specific for a particular antigen or allergen, such as pollen or insect venom. When these individuals are reexposed to that antigen, antigen–IgE complexes form on the surface of the inflammatory cells and activate the synthesis and release of PAF.

Structure of biomembranes

Eukaryotic cells have a plasma membrane, as well as a number of intracellular membranes that define compartments with specialized functions

Cellular and organelle membranes differ significantly in protein and lipid composition (Table 3.3). In addition to the major phospholipids described in Figure 3.7, other important membrane lipids include cardiolipin, sphingolipids (sphingomyelin and glycolipids), and cholesterol, which are described in detail in later chapters. Cardiolipin (diphosphatidyl glycerol) is a significant component of the mitochondrial inner membrane, while sphingomyelin, phosphatidylserine and cholesterol are enriched in the plasma membrane (see Table 3.3). Some lipids are distributed asymmetrically in the membrane, *e.g.* phosphatidylserine and phosphatidylethanolamine are enriched on the inside, and phosphatidylcholine and sphingomyelin on the outside, of the red blood cell membrane. The protein to lipid ratio also differs among various biomembranes, ranging from about 80% (dry weight) lipid in the myelin sheath that insulates nerve cells, to about 20% lipid in the inner mitochondrial membrane. Lipids affect the structure of the membrane, the activity of membrane enzymes and transport systems, and membrane function in processes such as cellular recognition and signal transduction. Exposure of phosphatidylserine in the outer leaflet of the erythrocyte plasma membrane increases the cell's adherence to the vascular wall and is a signal for macrophage recognition and phagocytosis. Both of these recognition processes contribute to the natural process of red cell turnover in the spleen.

Table 3.3

Phospholipid composition of organelle membranes from rat liver

	Mitochondria	Microsomes	Lysosomes	Plasma membrane	Nuclear membrane	Golgi membrane
Cardiolipin	18	1	1	1	4	1
Phosphatidylethanolamine	35	22	14	23	13	20
Phosphatidylcholine	40	58	40	39	55	50
Phosphatidylinositol	5	10	5	8	10	12
Phosphatidylserine	1	2	2	9	3	6
Phosphatidic acid	–	1	1	1	2	<1
Sphingomyelin	1	1	20	16	3	8
Phospholipids (mg/mg protein)	0.18	0.37	0.16	0.67	0.50	0.83
Cholesterol (mg/mg protein)	<0.01	0.01	0.04	0.13	0.04	0.08

This table shows the phospholipid composition (%) of various organelle membranes together with weight ratios of phospholipids and cholesterol to protein.

The fluid mosaic model

The fluid mosaic model portrays cell membranes as flexible lipid bilayers with embedded proteins

The generally accepted model of biomembrane structure is the fluid mosaic model proposed by Singer & Nicolson in 1972. This model represents the membrane as a fluid-like phospholipid bilayer into which other lipids and proteins are embedded (Fig. 3.8). As in liposomes, the polar head groups of the phospholipids are exposed on the external surfaces of the membrane, with the fatty acyl chains oriented to the inside of the membrane. Whereas membrane lipids and proteins easily move on the membrane surface (lateral diffusion), ‘flip-flop’ movement of lipids between the outer and inner bilayer leaflets rarely occurs without the aid of the membrane enzyme flippase.

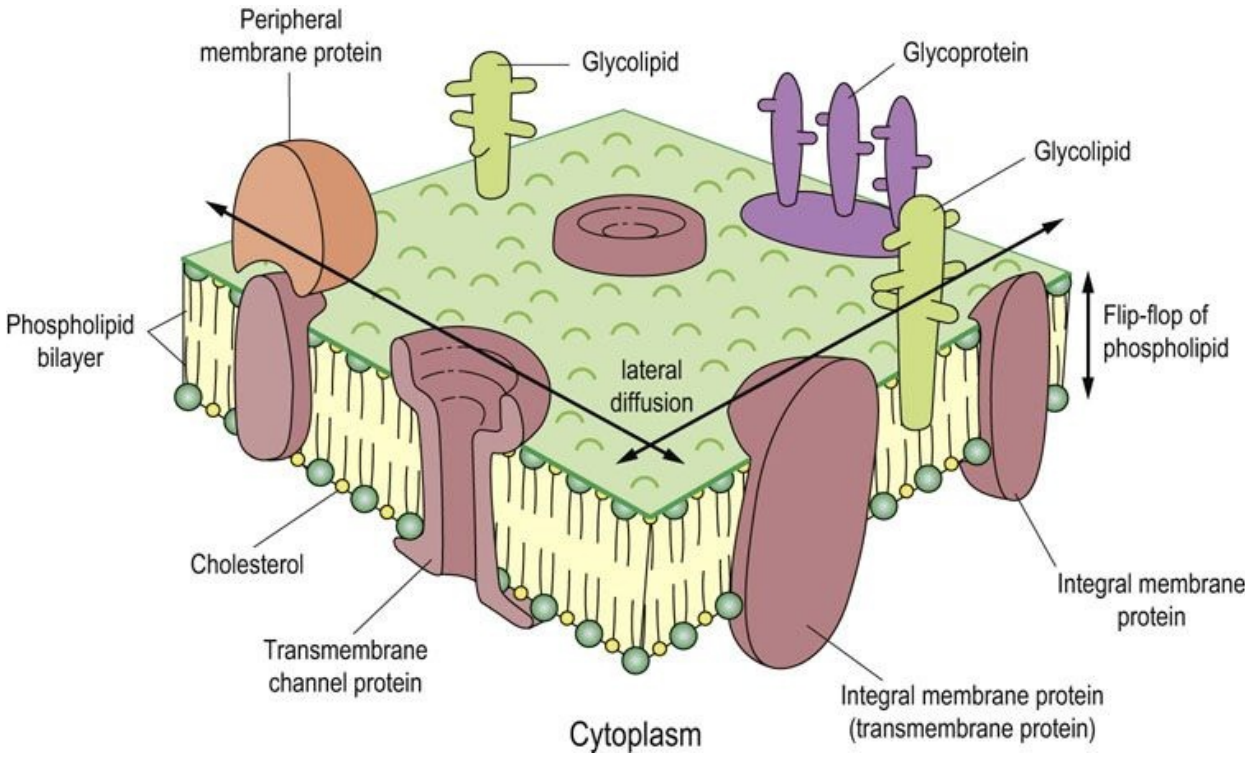


FIG. 3.8 Fluid mosaic model of the plasma membrane.
In this model, proteins are embedded in a fluid phospholipid bilayer; some are on one surface (peripheral) and others span the membrane (transmembrane). Carbohydrates, covalently bound to some proteins and lipids, are not found on all subcellular membranes, *e.g.* mitochondrial membranes. On the plasma membrane, they are located almost exclusively on the outer surface of the cell (see Chapters 8 and 28).

Membrane proteins are classified as integral (intrinsic) or peripheral (extrinsic) membrane proteins. The former are embedded deeply in the lipid bilayer and some of them traverse the membrane several times (**transmembrane proteins**) and have both internal and external polypeptide segments that participate in regulatory processes. In contrast, peripheral membrane proteins are bound to membrane lipids and/or integral membrane proteins (see Fig. 3.8); they can be removed from the membrane by mild denaturing agents, such as urea, or mild detergent treatment without destroying the integrity of the membrane. In contrast, transmembrane proteins can be removed from the membrane only by treatments that dissolve membrane lipids and destroy the integrity of the membrane. Most of the transmembrane segments of integral membrane proteins form α-helices. They are composed primarily of amino acid residues with nonpolar side chains – about 20 amino acid residues forming six to seven α-helical turns are enough to traverse a membrane of 5 nm (50 Å) thickness. The

transmembrane domains interact with one another and with the hydrophobic tails of the lipid molecules, often forming complex structures, such as channels involved in ion transport processes (see Fig. 3.8 and Chapter 8).

Membranes maintain the structural integrity, cellular recognition and the transport functions of the cell

There is growing evidence that many membrane proteins have limited mobility and are anchored in place by attachment to cytoskeletal proteins. Membrane substructures, described as lipid rafts, also demarcate regions of membranes with specialized composition and function. Specific phospholipids are also enriched in regions of the membrane involved in endocytosis and junctions with adjacent cells. However, the fluidity is essential for membrane function and cell viability. For example, when bacteria are transferred to lower temperature, they respond by increasing the content of unsaturated fatty acids in membrane phospholipids, thereby maintaining membrane fluidity at low temperature. The membrane also mediates the transfer of information and molecules between the outside and inside of the cell, including cellular recognition, signal transduction processes and metabolite and ion transport; fluidity is essential for these functions. Overall, cell membranes, which are often viewed by microscopy as static, are well-organized, flexible and responsive structures. In fact, the microscope picture is like a high-speed stop-action photo of a sporting event; it may look peaceful and still but there's a lot of action going on.

Advanced concept box Membrane patches

Although the fluid mosaic model is basically correct, it is recognized that there are membrane regions with unique protein and lipid compositions. **Caveolae**, 50–100 nm plasma membrane invaginations, and lipid rafts are plasma membrane patches (microdomains) important for signal transduction and endocytosis. These patches are enriched in cholesterol and sphingolipids, and the interaction of the long saturated fatty acid tails of sphingolipids with cholesterol results in the stabilization of the fluid environment.

The patches are detergent insoluble and show high buoyant density on sucrose density gradient centrifugation. Pathogens such

as viruses, parasites, bacteria and even bacterial toxins may enter into the host cells through binding to specific components of caveolae. Classic examples of patches enriched in a particular protein are the purple membrane of *Halobacterium halobium* containing bacteriorodopsin, and gap junctions containing connexin. Bacteriorodopsin is a light driven-proton pump which generates a H^+-concentration gradient across the bacterial membrane, providing energy for nutrient uptake for bacterial growth. **Gap junctions** between uterine muscle cells increase significantly during the late stages of pregnancy. They provide high-capacity channels between cells and permit coordinated contraction of the uterus during labor.

Advanced concept box Membrane perturbation by amphipathic compounds

Amphipathic compounds have distinct polar and nonpolar moieties. They include many anesthetics and tranquilizers. The pharmacologic activities of these compounds are dependent on their ability to interact with membranes and perturb membrane structure. A number of antibiotics and natural products, such as bile salts and fatty acids, are also amphipathic. While effective at therapeutic concentrations, some of these drugs exhibit detergent-like action at moderate to high concentrations and disrupt the bilayer structure, resulting in membrane leakage.

Advanced concept box Membrane anchoring proteins

Lateral movements of some membrane proteins are restricted by their tethering to macromolecular assemblies inside (cytoskeleton) and/or outside (extracellular matrix) the cell and, in some cases, to membrane proteins of adjacent cells, *e.g.* in tight junctions between epithelial cells.

Lateral diffusion of erythrocyte integral membrane proteins, band 3 (an anion transporter) and glycophorin, is limited by indirect interaction with spectrin, a cytoskeletal protein, through ankyrin and band 4.1 protein, respectively. Such interactions are so strong that they limit lateral diffusion of band 3. Genetic defects of spectrin cause hereditary spherocytosis and elliptocytosis, diseases characterized by altered red cell morphology. Ankyrin mutation affects the localization of plasma membrane proteins in cardiac muscle, causing cardiac arrhythmia, a risk factor for sudden cardiac death.

Summary

Following the previous chapter on amino acids and proteins, this chapter provides a broader foundation for further studies in biochemistry, by introducing the basic structural features and physical and chemical properties of two major building blocks – carbohydrates and lipids.

■ Carbohydrates are polyhydroxyaldehydes and ketones; they exist primarily in cyclic forms, which are linked to one another by glycosidic bonds.

■ Glucose is the only monosaccharides that exists in the body in free form.

■ Lactose and sucrose are important dietary disaccharides.

■ Starch, cellulose and glycogen are important homoglucan polymers of glucose.

■ Carbohydrates may be linked to proteins and lipids to form glycoconjugates, known as glycolipids and glycoproteins.

■ Lipids are hydrophobic compounds, commonly containing fatty acids esterified to glycerol.

■ Fatty acids are long chain alkanoic acids; unsaturated fatty acids contain one or more *cis*-double bonds, which decrease the melting (freezing) point of lipids.

■ Triglycerides (triacylglycerols) are the storage form of lipids in adipose tissue.

■ Phospholipids are amphipathic lipids found in biological membranes; they contain a phosphodiester at C-3 of glycerol, linking a diglyceride to an amino compound, most frequently choline, ethanolamine or serine.

■ The Fluid Mosaic Model describes the essential role of phospholipids, integral and membrane proteins and other lipids in the structure and function of biological membranes.

■ Biological membranes compartmentalize cellular functions, and also mediate ion and metabolite transport, cellular recognition, signal transduction, and electrochemical processes involved in bioenergetics, nerve transmission and muscle contraction.

Active learning

1. Compare the caloric value of starch and cellulose. Explain the difference.
2. Explain why disaccharides such as lactose, maltose and isomaltose are reducing sugars, but sucrose is not.
3. What does the iodine number of a lipid indicate about its structure?

4. Review the industrial process for making soaps.
5. Review the history of models for biological membranes. What are the limitations of the original Singer–Nicolson model?

Further reading

Brand-Miller, J, Buyken, AE. The glycemic index issue. *Curr Opin Lipidol*. 2012; 23:62–67.

Kar, S, Webel, R. Fish oil supplementation and coronary artery disease: does it help? *Mo Med*. 2012; 109:142–145.

Kusumi, A, Fujiwara, TK, Chadda, R, *et al.* Dynamic organizing principles of the plasma membrane that regulate signal transduction: commemorating the fortieth anniversary of Singer and Nicolson's fluid-mosaic model. *Annu Rev Cell Dev Biol*. 2012; 28:215–250.

Shanmugam, T, Banerjee, R. Nanostructured self-assembled lipid materials for drug delivery and tissue engineering. *Ther Deliv*. 2011; 2:1485–1516.

Singer, SJ. Some early history of membrane molecular biology. *Annu Rev Physiol*. 2004; 66:1–27.

Taubes, G. *Good calories, bad calories: fats, carbs, and the controversial science of diet and health*. New York: Anchor Books; 2008.

Zhang, YM, Rock, CO. Membrane lipid homeostasis in bacteria. *Nat Rev Microbiol*. 2008; 6:222–233.

Websites

History of membrane models. www1.umn.edu/ships/9-2/membrane.htm.

Soaps and saponification. www.kitchendoctor.com/articles/soap.html.

CHAPTER 4

Blood and Plasma Proteins

Marek H. Dominiczak and William D. Fraser

Learning objectives

After reading this chapter you should be able to:

- Describe the major components of blood.
- Explain the difference between plasma and serum.
- Define the roles of plasma proteins and their broad classification.
- Identify diseases associated with deficiency of specific proteins.
- Discuss the structure and function of the immunoglobulins.
- Appreciate the pathologic significance of monoclonal gammopathies.
- Define the acute phase response and the change it induces in the concentrations of circulating plasma proteins.

Introduction

For a clinician, plasma is an important ‘window’ on metabolism

Blood is the transport and distribution medium for the body. It delivers essential nutrients to tissues and removes waste products. It is an aqueous solution containing molecules of varying sizes, and a number of cellular elements. Some of the components of blood perform important roles in the body's defense against external insult and in the repair of damaged tissues. For a clinician, plasma is also an important ‘window’ on metabolism. Because it is easy to collect, most of the diagnostic laboratory tests in biochemistry, hematology and immunology are performed on plasma samples.

Plasma is the natural environment of blood cells, but many chemical measurements are done with serum

The formed elements of blood are suspended in an aqueous solution: the plasma. **Plasma** is the supernatant obtained after centrifugation of blood collected into a test tube containing **anticoagulant** to prevent clotting. Several anticoagulants are used in laboratory practice, the most common being lithium heparinate and ethylenediaminetetraacetic acid (EDTA). Heparinate prevents clotting by binding to thrombin. EDTA and citrate bind Ca^{2+} and Mg^{2+}, thus interfering with the action of calcium and magnesium-dependent enzymes involved in the clotting cascade (Chapter 7). Citrate is used as an anticoagulant when blood is collected for transfusion.

Serum, on the other hand, is the supernatant obtained after a blood sample has been allowed to clot spontaneously (this usually requires 30–45 minutes). During clotting, fibrinogen is converted to fibrin as a result of proteolytic cleavage by thrombin, and so a major difference between plasma and serum is the absence of fibrinogen in serum.

Throughout this book, when we describe physiologic or pathologic mechanisms, we refer to plasma; *e.g.* we would say that albumin binds many drugs present in *plasma*. We may mention serum when we specifically refer to the results of laboratory tests specifically performed on serum; *e.g.* we would say that patient's *serum* albumin was 40 mg/dL.

Formed Elements of Blood

There are three major cellular components of blood: the red blood cells (erythrocytes), the white blood cells (leukocytes) and the blood platelets (thrombocytes).

Erythrocytes are not complete cells, as they do not possess nuclei and intracellular organelles

The erythrocytes are cellular remnants, containing specific proteins and ions, which can be present in high concentrations. They are the end-product of erythropoiesis in the bone marrow, which is under the control of erythropoietin produced by the kidney (Fig. 4.1). Hemoglobin is synthesized in the erythrocyte precursor cells (erythroblasts and reticulocytes) under a tight control dictated by the concentration of heme (Chapter 30). The main functions of erythrocytes are the transport of oxygen and the removal of carbon dioxide and hydrogen ions; as they lack cellular organelles, they are not capable of protein synthesis and repair (Chapters 5 and 25). As a result, erythrocytes have a finite life span of approximately 120 days before being trapped and broken down in the spleen.

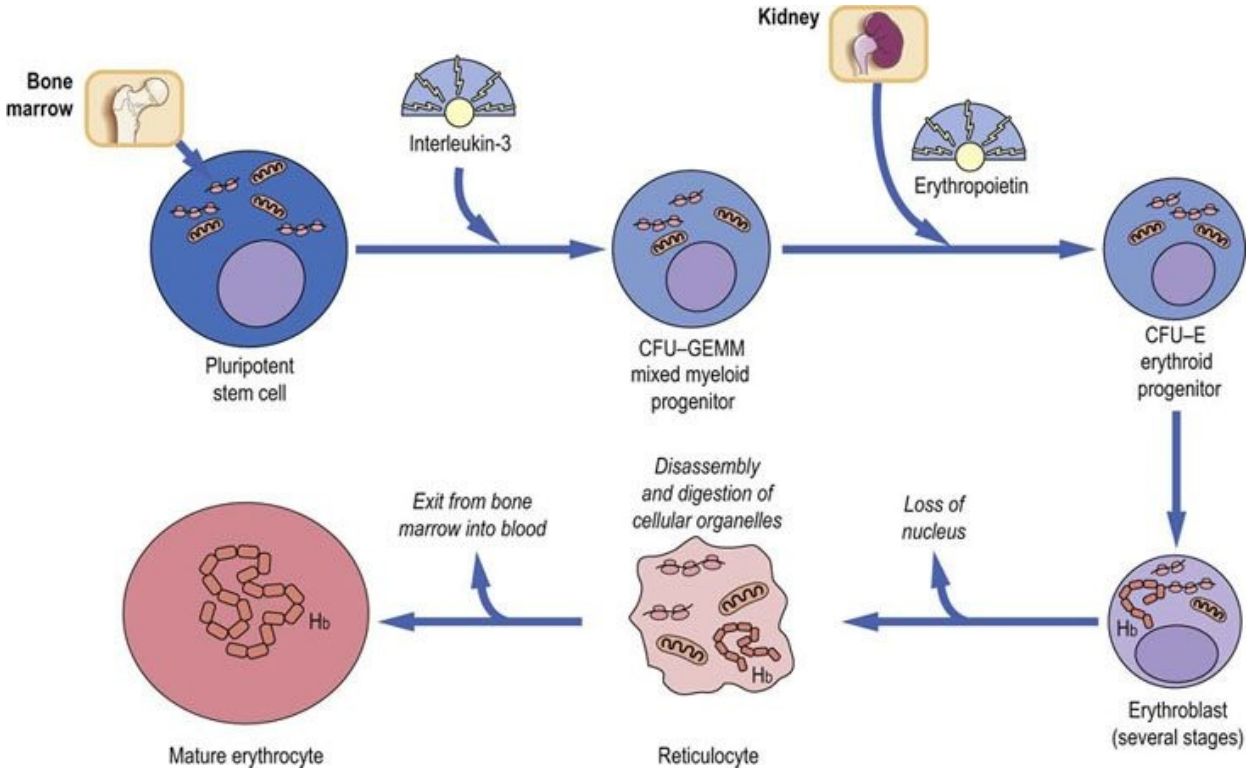


FIG. 4.1 Simplified scheme of the formation of erythrocytes.
In an average day, 10^{11} erythrocytes are formed. Hemoglobin is synthesized in the erythrocyte and reticulocyte before the loss of ribosomes and mitochondria. CFU-GEMM, colony-forming unit: granulocyte, erythroid, monocyte, megakaryocyte; CFU-E, colony-forming unit erythroid.

Leukocytes are cells, the main function of which is to protect the body from infection (Chapter 38)

Most leukocytes are produced in the bone marrow, some are produced in the thymus, and others mature within several tissues (Fig. 4.2) (Chapter 38). Leukocytes can control their own synthesis by secreting into the blood signal peptides that subsequently act on the bone marrow stem cells. In order to function correctly, leukocytes have the ability to migrate out of the bloodstream into surrounding tissues.

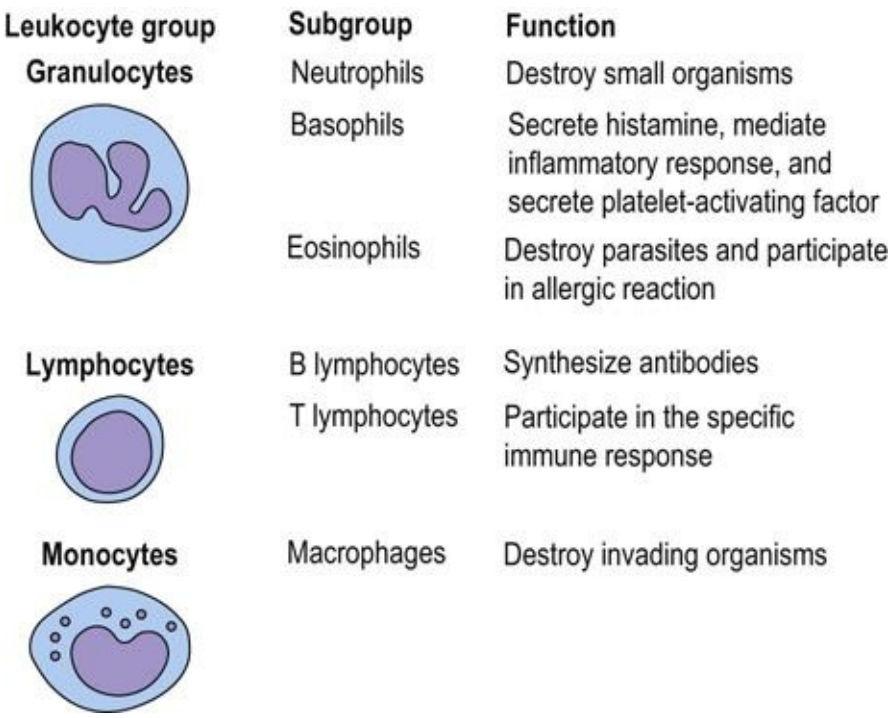


FIG. 4.2 Leukocytes.
Classification and functions of leukocytes (see also Chapter 38).

Thrombocytes are not true cells, but are membrane-bound fragments derived from megakaryocytes

They reside in the bone marrow. They play a key role in blood clotting (Chapter 7).

Plasma Proteins

Plasma proteins can be broadly classified into two groups: those, including albumin, that are synthesized by the liver, and the immunoglobulins, which are produced by plasma cells of the bone marrow, usually as part of the immune response.

A number of plasma proteins have the ability to bind certain molecules (their ligands) with a high affinity and specificity. These proteins can then act as a reservoir for the ligand and help control its distribution and availability by transporting it to tissues. Binding to a protein can also render a toxic substance less harmful to the tissues. Major binding proteins and their ligands are shown in Table 4.1.

Table 4.1

Transport proteins and their ligands

Proteins	Ligands
Cation binding	
Albumin	Divalent and trivalent cations, e.g. Cu^{2+}, Fe^{3+}
Ceruloplasmin	Cu^{2+}
Transferrin	Fe^{3+}
Hormone binding	
Thyroid-binding globulin (TBG)	Thyroxine (T4), tri-iodothyronine (T3)
Cortisol-binding globulin (CBG)	Cortisol
Sex hormone-binding globulin (SHBG)	Androgens (testosterone), estrogens (estradiol)
Hemoglobin/protoporphyrin binding	
Albumin	Heme, bilirubin, biliverdin
Haptoglobin	Hemoglobin dimers
Fatty acid binding	
Albumin	Nonesterified fatty acids, steroids

Albumin

Albumin serves as an osmotic regulator and is a major transport protein

Albumin is the predominant plasma protein. It has no known enzymatic or hormonal activity, and accounts for approximately 50% of the protein found in human plasma. Its normal concentration is 35–45 g/L. With a molecular weight of about 66 kDa, albumin has highly polar nature and dissolves easily in water. At pH 7.4, it is an anion with 20 negative charges per molecule; this gives it a

high capacity for non-selective binding of many ligands. It also plays a critical role in maintaining colloid osmotic pressure of the plasma.

The rate of albumin synthesis (14–15 g daily) depends on nutritional status, especially on the extent of amino acid deficiencies. Its half-life is about 20 days. Importantly, although the albumin level reflects the nutritional status in the longer term, in hospitalized patients the short-term changes in plasma albumin concentration are usually due to changes in hydration (Chapter 24).

Albumin is the primary plasma protein responsible for the transport of hydrophobic fatty acids, bilirubin, and drugs

Albumin can bind (and thus solubilize) a range of substances that include the long-chain fatty acids, sterols, and several synthetic compounds. The transport of long-chain fatty acids underpins much of the body's distribution of energy-rich substrates. There are numerous fatty acid binding sites on the albumin molecule, with variable affinities. The highest affinity sites are believed to lie in the globular segments within specialized clefts of the albumin molecule (Fig. 4.3).

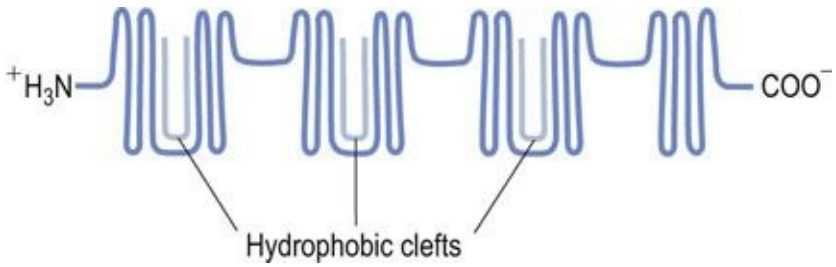


FIG. 4.3 Molecular model of human albumin.
The hydrophobic clefts are globular segments of albumin that bind fatty acids with high affinity.

In addition to binding fatty acids, albumin binds unconjugated bilirubin (Chapter 30). The sites within the albumin molecule are also capable of binding a variety of drugs, including salicylates, barbiturates, sulfonamides, penicillin and warfarin. This is of great pharmacologic relevance. Such interactions are weak and the ligands become easily displaced by substances competing for a binding site. Somewhat surprisingly, albumin is not essential for human survival and rare congenital defects have been described where there is hypoalbuminemia or complete absence of albumin (analbuminemia).

Proteins that transport metal ions

Transferrin transports iron

The binding of ferric ions (Fe^{3+}) to transferrin protects against the toxic effects of these ions. In inflammatory reactions, the iron–transferrin complex is degraded by the reticuloendothelial system without a corresponding increase in the synthesis of either of its components; this results in low plasma concentrations of transferrin and iron (Chapter 11).

Ferritin is the major iron storage protein found in almost all cells of the body

Ferritin acts as the reserve of iron in the liver and bone marrow. The concentration of ferritin in plasma is proportional to the amount of stored iron; therefore measurement of plasma ferritin is one of the best indicators of iron deficiency.

Ceruloplasmin is the major transport protein for copper

Ceruloplasmin helps export copper from the liver to peripheral tissues, and is essential for the regulation of the oxidation reduction reactions, transport, and utilization of iron (Fig. 4.4). Increased concentrations of ceruloplasmin occur in active liver disease and in tissue damage.

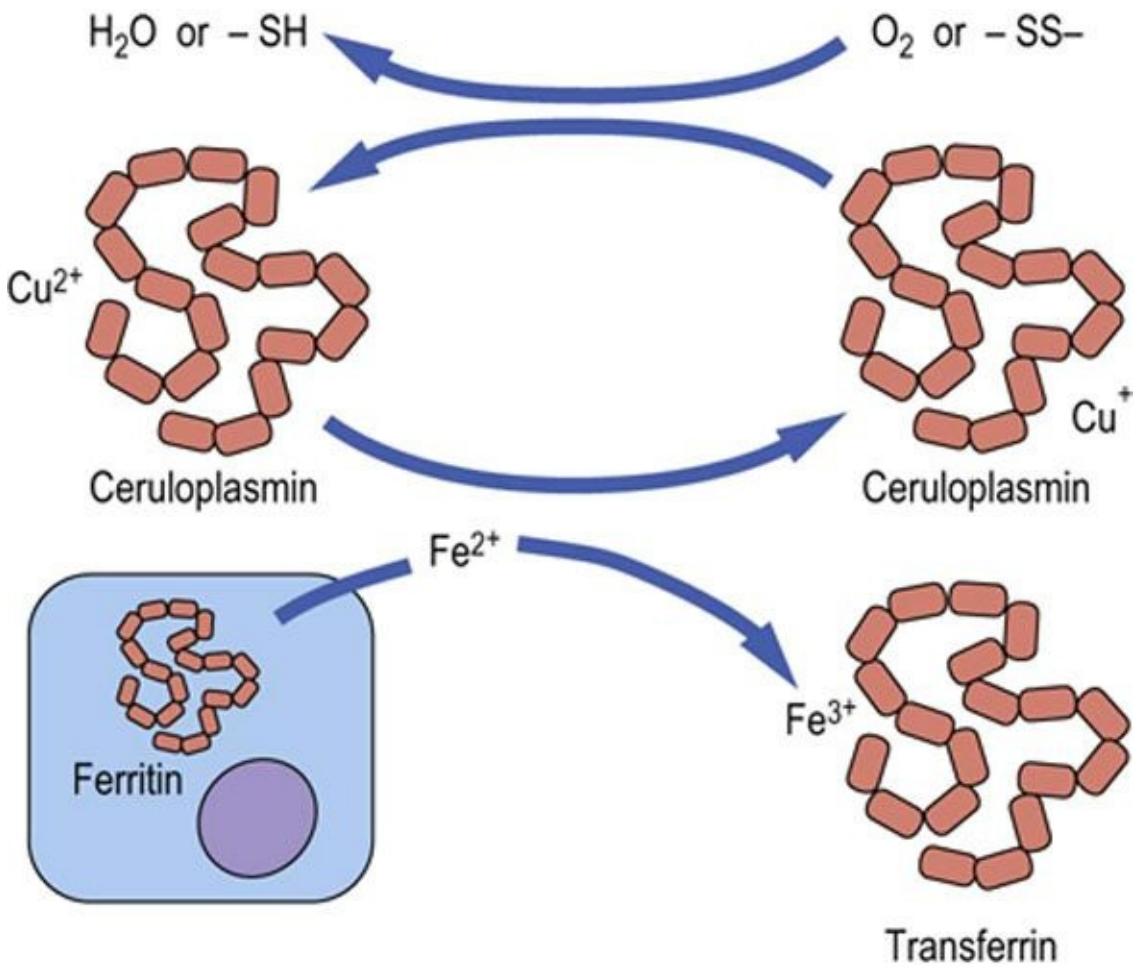


FIG. 4.4 Plasma ferroxidase activity of ceruloplasmin.
Oxidation of Fe^{2+} by ceruloplasmin permits the binding and transport of iron by plasma transferrin. The cuprous ion (Cu^{2+}) bound to ceruloplasmin is regenerated by reaction with oxygen or with oxidized thiol groups.

Advanced concept box Hemolysis and free hemoglobin

When erythrocytes are hemolyzed, hemoglobin is released into the plasma, where it dissociates into dimers that bind to **haptoglobin**. The hemoglobin–haptoglobin complex is metabolized more rapidly than haptoglobin alone, in the liver and reticuloendothelial system, producing an iron–globulin complex and bilirubin. This prevents the loss of iron in the urine. When excessive hemolysis occurs, the plasma haptoglobin concentration can become very low – and as

such it serves as a **marker of hemolysis**. If hemoglobin breaks down into heme and globin, the free heme is bound by hemopexin. Unlike haptoglobin, which is an acute phase protein, hemopexin is not affected by acute phase response. The heme–hemopexin complex is taken up by the liver cells, where iron binds to ferritin. A third complex, called methemalbumin, can form between oxidized heme and albumin. These mechanisms have evolved to allow the body not only to prevent major losses of iron but also to complex the free heme, which is toxic to many tissues.

Clinical box A 14-year-old girl with abdominal pain and enlarged liver

Wilson's disease

A 14-year-old girl was admitted as an emergency. She was jaundiced with abdominal pain and had an enlarged, tender liver. She was drowsy and had asterixis (flapping tremor) due to acute liver failure. Previous history revealed behavior disturbance, difficulty with movement in the recent past, and truancy from school. Her ceruloplasmin concentration was 50 mmol/L (normal range 200–450 mmol/L (20–45 mg/dL)), serum copper was 8 mmol/L (normal range 10–22 µmol/L (65–144 µg/dL)), urinary excretion of copper was 4.2 µmol/24 h (normal range 2–3.9 µmol/24 h (13–25 µg/dL)). A liver biopsy established the diagnosis of Wilson's disease.

Comment.

This case highlights the importance of measurement of ceruloplasmin. **In Wilson's disease, a deficiency of ceruloplasmin leads to low plasma concentrations of copper.** The metabolic defect is in the excretion of copper in bile and its reabsorption in the kidney; copper is thus deposited in liver, brain,

and kidney. Liver symptoms are present in patients of younger age, and cirrhosis and neuropsychiatric problems are predominant in those who are older. Detection of low plasma concentrations of ceruloplasmin and copper, increased urinary excretion of copper, and markedly increased concentrations of copper in the liver confirm the diagnosis.

Immunoglobulins

Immunoglobulins are proteins produced in response to foreign substances (antigens; see Chapter 38)

Immunoglobulins (antibodies) are secreted by the B lymphocytes. They have defined specificity for foreign substances that stimulated their synthesis. Not all foreign substances entering the body can elicit this response, however; those that do are called **immunogens**, whereas any agent that can be bound by an antibody is termed an **antigen**. The immunoglobulins form a uniquely diverse group of molecules, recognizing and reacting with a wide range of specific antigenic structures (**epitopes**) and giving rise to a series of effects that result in the eventual elimination of the presenting antigen. Some immunoglobulins have additional effector functions; for example, IgG is involved in complement activation.

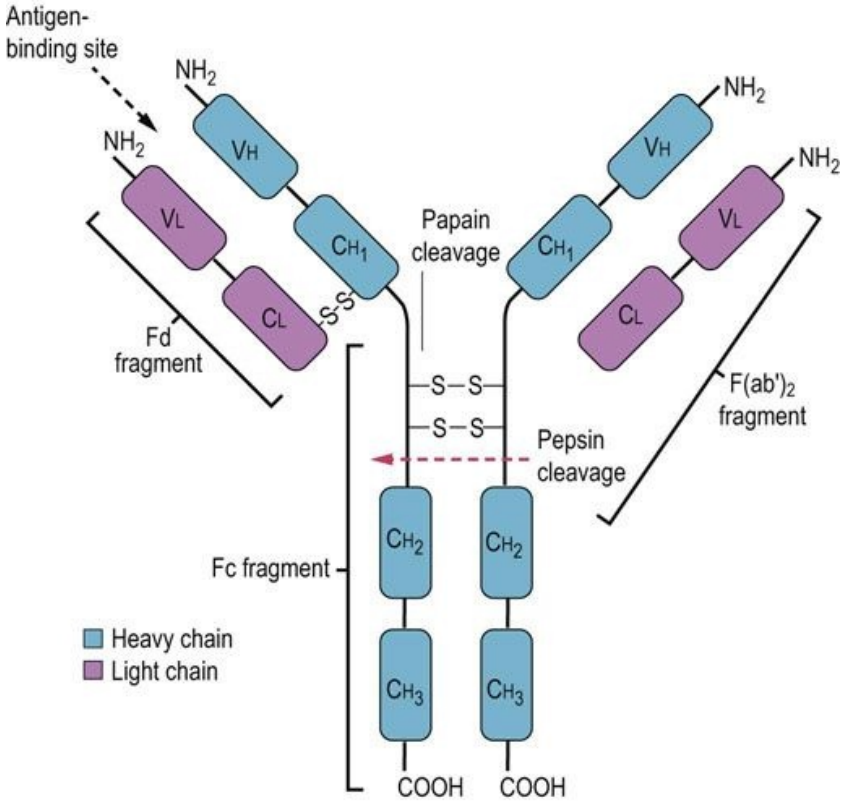
Immunoglobulins share a common Y-shaped structure of two heavy and two light chains

An immunoglobulin is a Y-shaped molecule containing two identical units termed heavy (H) chains and two identical, but smaller, units termed light (L) chains. Several H chains exist, and the nature of the H chain determines the class of immunoglobulin: IgG, IgA, IgM, IgD and IgE are characterized by α, γ, δ, μ, and ε heavy chains, respectively. L chains are of only two types, κ and λ. Each polypeptide chain within the immunoglobulin is characterized by a series of globular regions, which have considerable sequence homology and, in evolutionary terms, are probably derived from protogene duplication.

The *N*-terminal domains of both H and L chains contain a region of variable

amino acid sequence (the V region); together, these regions determine antigenic specificity. Both H and L chains are required for full antibody activity, as the physically apposed V regions in the L and H chains form a functional pocket into which the epitope fits; this is termed the antibody recognition ($F(ab'_2)$) region. The domain immediately adjacent to the V region is much less variable, in both H and L chains. The remainder of the H chain consists of a further constant region (Fc region) consisting of a hinge region and two additional domains. The constant region is responsible for immunoglobulin functions other than epitope recognition, such as complement activation (Chapter 38). This basic structure of immunoglobulins is depicted in Figure 4.5. When antigen binds to the immunoglobulin, conformational changes are transmitted through the hinge region of the antibody to the Fc region, which is then said to have become activated.

The immunoglobulin G molecule



Pentameric IgM

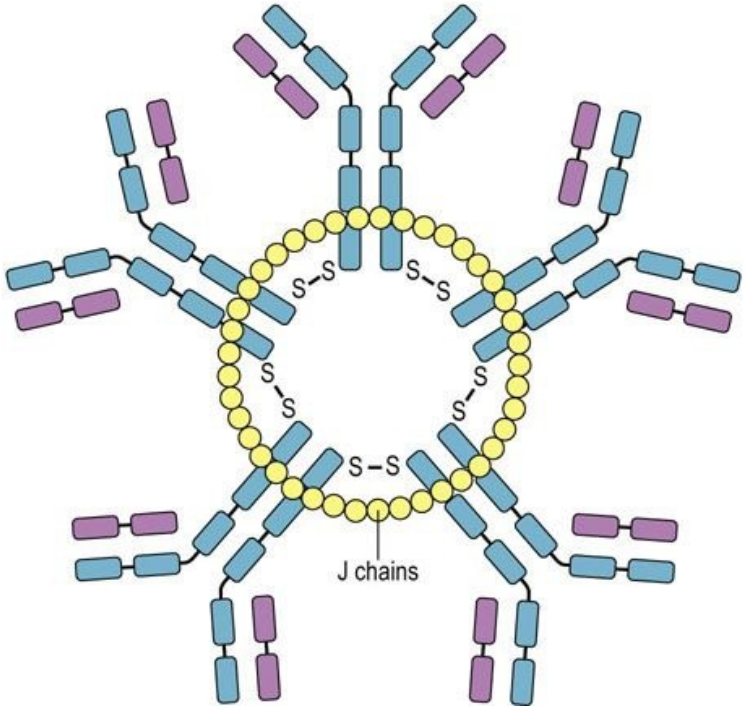


FIG. 4.5 The structure of immunoglobulins.
Diagrammatic representation of the basic structure of a monomeric immunoglobulin and that of pentameric immunoglobulin (IgM). V, variable region; C, constant region; H, heavy chain; L, light chain; J chain, joining chain; $F(ab')_2$, fragment generated by pepsin cleavage of the molecule; Fc, Fd, fragments generated by papain proteolysis.

Major immunoglobulins

IgG is the most common immunoglobulin that protects tissue spaces and freely crosses the placenta

IgG, with an overall molecular mass of 160 kDa, consists of the basic 2H2L immunoglobulin subunit joined by a variable number of disulfide bonds. The γ H chains have several antigenic and structural differences, allowing classification of IgG into a number of subclasses according to the type of H chain present; however, functional differences between the subclasses are minor.

IgG circulates in high concentrations in the plasma, accounting for 75% of immunoglobulin present in adults, and has a half-life of 22 days. It is present in all extracellular fluids, and appears to eliminate small, soluble antigenic proteins through aggregation and enhanced phagocytosis by the reticuloendothelial system. From weeks 18–20 of pregnancy, IgG is actively transported across the placenta and provides humoral immunity for the fetus and neonate before maturation of the immune system.

IgA is found in secretions and presents an antiseptic barrier, which protects mucosal surfaces

IgA has an H chain similar to the γ chain of IgG, and α chains possess an extra 18 amino acids at its *C*-terminus. The extra peptide sequence enables the binding of a ‘joining’ or J chain. This short (129-residue) acidic glycopeptide, synthesized by plasma cells, allows dimerization of secretory IgA. IgA is often found in noncovalent association with the so-called secretory component, a highly glycosylated 71 kDa polypeptide, synthesized by mucosal cells and capable of protecting IgA against proteolytic digestion.

IgA represents 7–15% of plasma immunoglobulins and has a half-life of 6 days. It is found, in particular in the dimerized form, in parotid, bronchial, and intestinal secretions. It is a major component of colostrum (the first milk from the mother's breasts after the birth of a child). IgA appears to function as the

primary immunologic barrier against pathogenic invasion of mucous membranes. It can promote phagocytosis, cause eosinophilic degranulation, and activate complement via the so-called alternative pathway.

IgM is confined to the intravascular space and helps eliminate circulating antigens and microorganisms

Immunoglobulins belonging to this final major class are polyvalent, with a high molecular mass. IgM has a basic form similar to that of IgA, having the extra H chain domain that allows for J chain binding, and is thus capable of polymerization. IgM normally circulates as a pentamer, with a molecular mass of 971 kDa, linked by disulfide bonds and the J chain (Fig. 4.5).

IgM accounts for 5–10% of plasma immunoglobulins and has a half-life of 5 days. With its polymeric nature and high molecular mass, most IgM is found confined to the intravascular space, although lesser amounts may be found in secretions, usually in association with secretory component. It is the first antibody to be synthesized after an antigenic challenge.

Minor immunoglobulins

IgD is the surface receptor for antigen in B lymphocytes

IgD differs from the standard immunoglobulin structure chiefly by its high carbohydrate content of numerous oligosaccharide units, resulting in an increased molecular mass of 190 kDa. Its δ chains are characterized by having only a single interconnecting disulfide bridge, and an elongated hinge region that is particularly susceptible to proteolysis.

IgD accounts for less than 0.5% of circulating plasma immunoglobulin mass. Its role remains elusive although, as a surface component of the mature B cells, it probably has some role in the response to antigens. Rare cases of isolated IgD deficiency seem to be associated with no obvious pathology.

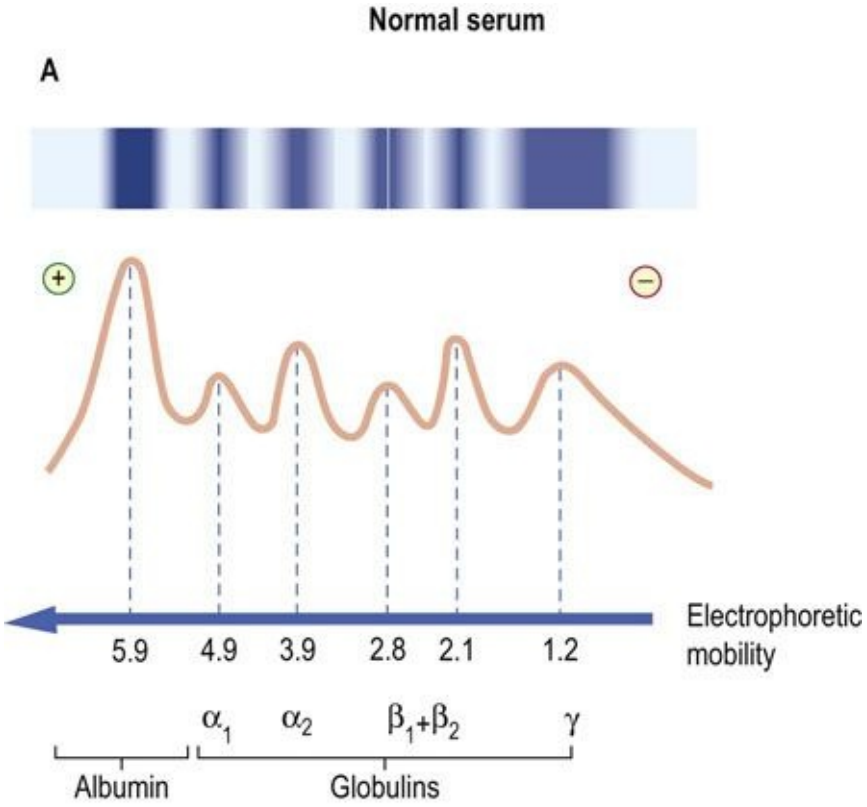
IgE binds antigens and promotes release of vasoactive amines from mast cells

IgE is similar to IgM in its unit structure. It has ε heavy chains that consist of five, rather than four, domains, but J chain binding and polymerization do not occur. The extended H chain helps to explain its high molecular mass of approximately 200 kDa. It is present only in trace amounts in plasma.

IgE has a high affinity for binding sites on mast cells and basophils. Antigenic binding at the Fab_2 region induces crosslinking of the high-affinity receptor, granulation of the cell, and release of vasoactive amines. By this mechanism, IgE plays a major part in allergy/atopy and mediates antiparasitic immunity.

Monoclonal immunoglobulins are the product of a single B cell, and arise from benign or malignant transformations of B cells

Monoclonal immunoglobulins result from the proliferation of a single B cell clone, which thus produces identical antibodies. Usually these are structurally normal molecules but sometimes they may be in some way fragmented or truncated. On gel electrophoresis, monoclonal immunoglobulin forms a single band in the γ region (the **paraprotein band**) (Fig. 4.6).



α_1 band: high-density lipoprotein
α_1-acid glycoprotein
α_1-antitrypsin
α_2 band: α_2-macroglobulin, haptoglobin
β_1 + β_2 band: transferrin and low-density lipoprotein
γ band: immunoglobulins

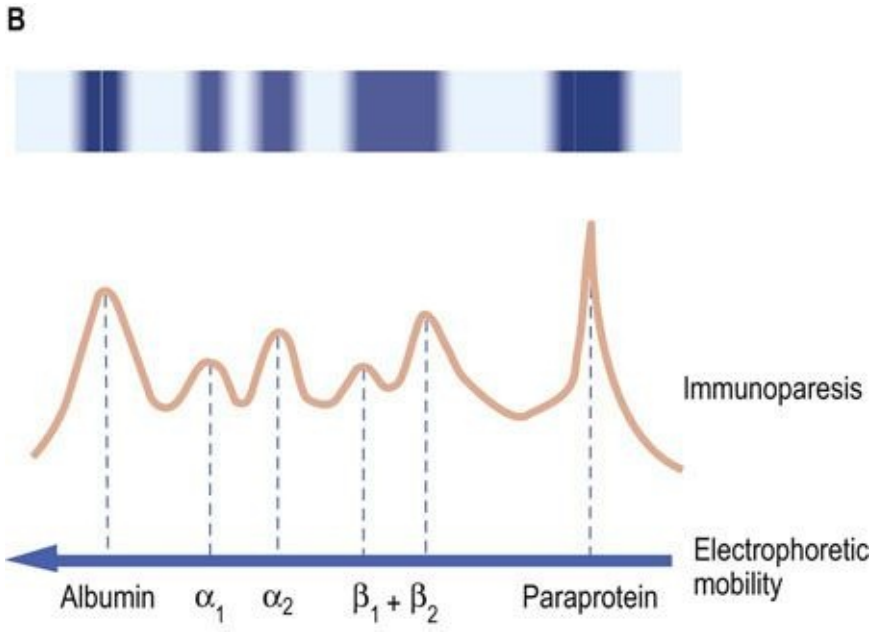


FIG. 4.6 Comparison of gel electrophoretic appearance of normal serum and that containing monoclonal immunoglobulins.
The scanning pattern peaks (solid line) represent the relative concentrations of the separated proteins. **(A)** Normal serum. **(B)** Monoclonal gammopathy: a strongly stained band is present in the γ-globulin region on electrophoresis, and there is an associated reduction of staining in the remainder of the γ-region (immunoparesis).

Monoclonal immunoglobulins are associated with malignant pathologies such as **myeloma** and **Waldenström's macroglobulinemia**, and also with more benign transformations that are known as monoclonal gammopathies of uncertain significance (MGUS).

Clinical box A man with a sudden onset of back pain: multiple myeloma

A 65-year-old man presented with a sudden onset of low back pain. Radiography revealed a crush fracture of the second lumbar vertebra, and discrete so-called ‘punched out’ lesions in the skull. Serum electrophoresis demonstrated the presence of a monoclonal immunoglobulin. This proved to be an IgG and, on electrophoresis, excess free κ-chains (**Bence–Jones protein**) were found in the patient's urine.

Comment.

Multiple myeloma affects men and women with equal incidence

and presents mostly after the age of 50 years. The clinical features are due to both the malignant proliferation of monoclonal plasma cells and the synthesis and secretion of antibody by these cells. Bone lesions affect the skull, vertebrae, ribs, and pelvis. There may be generalized osteoporosis and pathologic fractures. In up to 20% of cases, no monoclonal plasma protein is detected, although Bence–Jones proteins are present in urine. Such cases are commonly associated with suppression of the production of other immunoglobulins (immunoparesis). The presence of excess light chains may cause renal failure as a result of the deposition of Bence–Jones proteins in the renal tubules, or amyloidosis. Other common findings include anemia and hypercalcemia.

The Acute Phase Response and C-Reactive Protein

The acute phase response is a nonspecific response to tissue injury or infection; it affects several organs and tissues

During the acute phase response, there is a characteristic marked increase in the synthesis of some proteins (predominantly in the liver), along with a decrease in the plasma concentration of some others (Fig. 4.7). An increase in the synthesis of proteins such as proteinase inhibitors (α_1-antitrypsin), coagulation proteins (fibrinogen, prothrombin), complement proteins, and C-reactive protein is of obvious clinical benefit. The synthesis of albumin, transthyretin (prealbumin), and transferrin decreases during the acute phase response, and they are thus termed the 'negative acute phase reactants'.

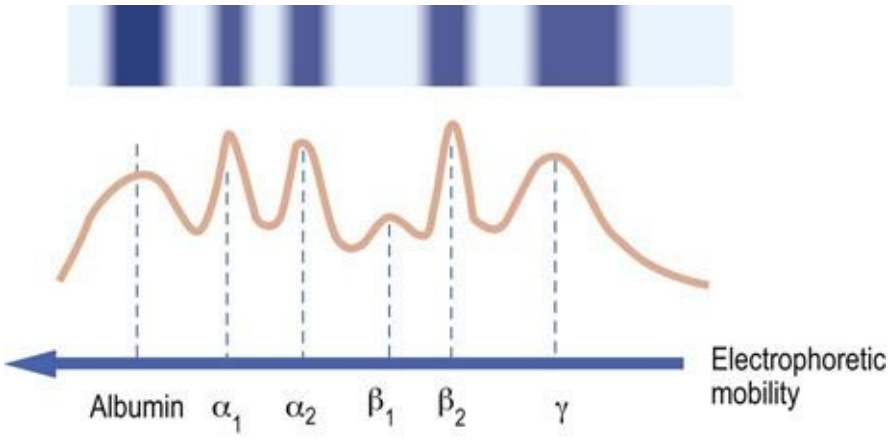


FIG. 4.7 Acute phase response.
Gel electrophoretic pattern observed in serum during the acute phase response. Albumin is decreased, the sum of α_1- and α_2-globulins is increased, β_1-globulins are decreased, β_2-globulins are increased and there is a mild increase in γ-globulins. Compare this to the normal pattern shown in Fig. 4.6A.

C-reactive protein (CRP) is a major component of the acute

phase response and a marker of bacterial infection

CRP is synthesized in the liver and is constructed of five polypeptide subunits. It has a molecular mass of around 130 kDa. It is present in only minute quantities (<1 mg/L in normal serum) and is believed to bind to phospholipids, and some proteins on invading microorganisms or damaged cells, facilitating activation of the complement via the classic pathway (Chapter 38). Measurement of CRP concentration in plasma is an essential laboratory test in diagnosis and monitoring of infection and sepsis (Fig. 4.8).

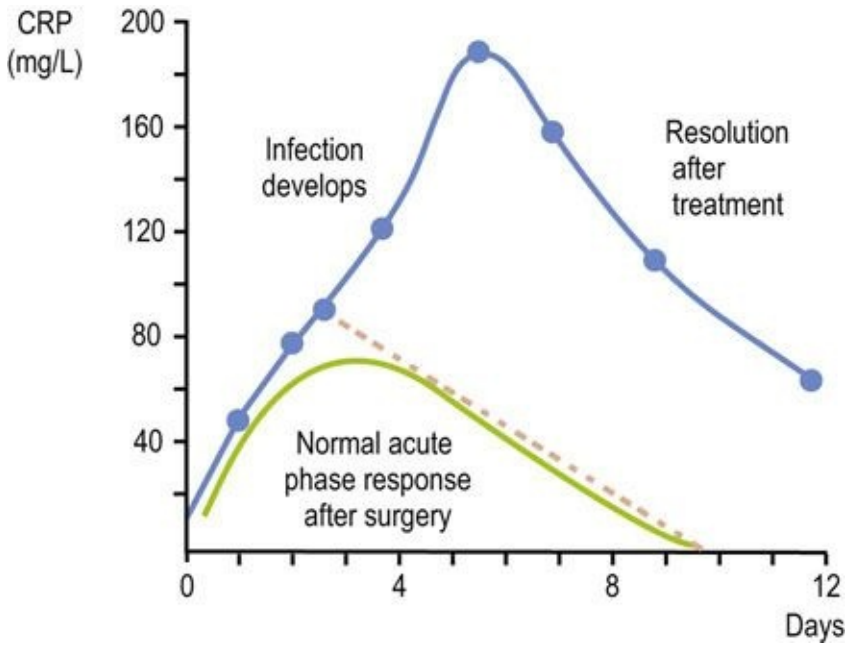


FIG. 4.8 C-reactive protein (CRP) and the postoperative acute phase reaction. The concentration of CRP increases as part of the acute phase response to surgical trauma, and a further increase may be observed if recovery is complicated by infection (the dotted line represents a response to uncomplicated surgery).

High sensitivity CRP assay is used in the assessment of cardiovascular risk

An assay for CRP, which is approximately 100 times more sensitive than the conventional CRP measurement method, detects minimal fluctuations in the concentration of this protein in infection-free individuals. Very small increases in CRP concentration seem to reflect a state of **chronic low-grade**

inflammation associated, for instance, with an increased risk of cardiovascular disease (Chapter 18). Other inflammatory conditions such as inflammatory bowel disease, type 2 diabetes and the metabolic syndrome have been associated with these minute increases in serum CRP concentration.

Clinical box A 44-year-old woman with edema: nephrotic syndrome

A 44-year-old woman was admitted to hospital because of weakness, anorexia, recurrent infections, bilateral leg edema, and breathlessness. Her plasma albumin concentration was 19 g/L (normal range 36–52 g/L) and her urinary protein excretion 10 g/24 h (normal value 0.15 g/24 h). There was microscopic hematuria. Renal biopsy confirmed the diagnosis as membranoproliferative glomerulonephritis.

Comment.

This woman had the classic triad of the nephrotic syndrome: hypoalbuminemia, proteinuria, and edema. The nephritis has resulted in damage to the glomerular basement membrane, with resultant leak of albumin. In the nephrotic syndrome, the continued loss of albumin exceeds the synthetic capacity of the liver, and results in hypoalbuminemia; consequently, the capillary osmotic pressure is significantly reduced. This leads to both peripheral (leg) edema and pulmonary edema (breathlessness). With increasing glomerular damage, proteins of larger molecular mass, such as immunoglobulins and complement, are lost in urine.

Clinical box A woman admitted after a road traffic accident: acute phase response

A 45-year-old woman suffered severe lower limb injuries in a road traffic accident. After admission to hospital, biochemical profiling

revealed slightly decreased concentrations of total serum protein (58 g/L; normal 60–80 g/L) and serum albumin (38 g/L; normal 36–52 g/L). Serum electrophoresis revealed an increase in the α_1 and α_2 protein fractions. Four days after her operation, the patient's condition deteriorated and she developed an increased temperature, sweating, and confusion. An acute infection was diagnosed and treatment with appropriate antibiotics was commenced. CRP concentrations peaked 5 days after the operation (Fig. 4.8).

Comment.

Increased concentrations of α_1 and α_2 proteins (which include α_1-antitrypsin, α_1-acid glycoprotein, and haptoglobin), together with a decrease in serum albumin concentration, suggest an acute phase response. This response is also associated with an increase in CRP, the erythrocyte sedimentation rate (ESR), and increased plasma viscosity. Patient's response to treatment of infection is associated with a decrease in plasma CRP concentration.

Clinical Tests: Biomarkers

Clinical laboratories perform a large number of biochemical analyses on body fluids to provide answers to specific clinical questions

The majority of specimens received by the laboratory are blood and urine samples. Whereas some measurements are performed on whole blood, serum or plasma are preferred for most analyses of metabolites and ions. In general, the time devoted to the analysis of each sample is short, but the entire process from a request for analysis to receipt of a result involves many steps. Throughout the process, constant checking and quality assurance are performed to ensure that the produced results are analytically and clinically valid. An outline of the laboratory workflow is shown in Figure 4.3.

Today's clinical laboratories rely, much like modern genetics research labs, on high throughput-automated instruments and robotic technologies

A high degree of automation and the underlying information technology allow us to perform large numbers of analyses, while enabling customized sets of hematology and biochemistry tests, (see Appendix 1). It is the logistics of specimen transport to the laboratory, rather than analysis itself, which in many instances becomes the factor that limits the workflow.

Discovering new biomarkers

A biomarker is an indicator of biological state, or a characteristic that is measured as an indicator or normal or pathologic process

Such a marker can be used for screening and risk assessment, diagnosis and detection of clinical conditions, or for monitoring of treatment and its side effects. The process of discovery of new biomarkers is now driven by the application of -omic technologies: genomics, transcriptomics, proteomics and metabolomics (Chapter 36). Blood and urine samples remain the most important

materials for biomarker discovery.

Metabolomics explores patterns of small molecules

Metabolomics is the study of all metabolites generated in the organism, including the metabolites of drugs, food-derived compounds and substances generated by the microbial flora

The Human Metabolome Database (version 2.5) contains around 7900 metabolites. The exploration of metabolites is a process located downstream from genomics and transcriptomics – the study of genes and their expression, respectively (Chapter 36). Metabolomic methodologies allow exploration of entire pathways, producing patterns of metabolite concentrations. Thus they provide a dynamic picture of the entire metabolite pattern in any given condition. The drawback of such a complex picture is that it can be confounded by compounds derived from ingested food or by metabolites of drugs and other foreign substances (xenobiotics).

The key process is the comparison of such patterns between the reference and affected populations using techniques such as mass spectrometry combined with gas chromatography (GCMS) or liquid chromatography (LC-MS). Individual metabolites can subsequently be identified using nuclear magnetic resonance (NMR).

Validation of the potential markers requires studies on large cohorts of nonaffected and affected individuals. The required population sizes usually exceed the capabilities of a single study. Therefore **meta-analysis** (comparison of different studies) is extensively used. Such meta-analysis can be simply a comparison of several published studies, or can involve a combined analysis of group data from different studies, or – the most laborious but also most reliable method – it can constitute the analysis of individual participant data from different studies combined into a ‘new’ very large group.

Summary

■ The formed elements of blood are erythrocytes, leukocytes and platelets. They are suspended in an aqueous solution (plasma) and have several specialized functions such as transport of oxygen, destruction of external agents, and the clotting of blood. Plasma allowed to clot yields serum. Most biochemical tests are done on serum. To obtain plasma, blood must be taken into a test tube containing an anticoagulant.

■ Plasma contains many proteins broadly classified into albumin and globulins (predominantly immunoglobulins). Albumin functions as a determinant of the osmotic pressure and a major transport protein for trace metals, hormones, bilirubin, and free fatty acids.

■ Other binding proteins bind specific ligands, *e.g.* ceruloplasmin binds Cu^{2+} and thyroxine-binding globulin (TBG) binds thyroid hormones.

■ Immunoglobulins participate in the defense against antigens that may enter or attempt to enter the body. They have a common structure: five classes of immunoglobulin exist and perform different protective functions.

■ Changes in the concentration of plasma proteins give important clinical information. A characteristic pattern with decreased albumin, transthyretin and transferrin and increased α_1-antitrypsin, fibrinogen and C-reactive protein indicates the acute phase response.

■ Serum and urine protein electrophoresis is an important way of identifying the presence of monoclonal immunoglobulins.

Active learning

1. Compare and contrast plasma and serum and discuss the different types of blood samples taken for laboratory tests.
2. Discuss the transport role of serum albumin.
3. Describe the core structure of immunoglobulins and different roles played in immunity by different classes of immunoglobulins.
4. How does the acute phase reaction affect the results of blood tests?
5. Characterize Wilson's disease.
6. What happens to hemoglobin when erythrocytes become disrupted?

Further reading

Abu Aboud, O, Weiss, RH. New opportunities from the cancer metabolome. *Clin Chem*. 2013; 59:138–146.

Anderson, NL. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. *Clin Chem*. 2010; 56:177–185.

El-Youssef, M. Wilson's disease. *Mayo Clin Proc*. 2003; 78:1126–1136.

Gilstrap, LG, Wang, TJ. Biomarkers and cardiovascular risk assessment for primary prevention: an update. *Clin Chem*. 2012; 58:72–82.

Nature Outlook. Multiple myeloma. *Nature*. 2011; 480:833–858.

Pavlou, MP, Diamandis, EP, Blasutig, IM. The long journey of cancer biomarkers from the bench to the clinic. *Clin Chem*. 2013; 59:147–157.

Suhre, K, Shin, S-Y, Petersen, A-K, *et al*. Human metabolic individuality in biomedical and pharmaceutical research. *Nature*. 2011; 477:54–60.

Zimmermann, MA, Selzman, CH, Cothren, C. Diagnostic implications of C-reactive protein. *Arch Surg*. 2003; 138:220–224.

Websites

Biomarkers. www.niehs.nih.gov/health/topics/science/biomarkers/index.cfm.

National Cancer Institute, NCI NIH. NCI Dictionary of Cancer Term. www.cancer.gov/dictionary?cdrid=45618

CHAPTER 5

Oxygen Transport

Norma Frizzell and George M. Helmkamp Jr.

Learning objectives

After reading this chapter you should be able to:

- Describe the mechanism of oxygen binding to myoglobin and hemoglobin.
- Describe conformational differences between deoxygenated and oxygenated hemoglobins.
- Define the concept of cooperativity in oxygen binding to hemoglobin.
- Describe the Bohr effect and its role in modulating the binding of oxygen to hemoglobin.
- Explain how 2,3-bisphosphoglycerate interacts with hemoglobin and influences oxygen binding.
- Summarize the processes by which carbon dioxide is transported from peripheral tissues to the lungs.
- Describe the major classifications of hemoglobinopathies.
- Describe the molecular basis of sickle cell disease.

Introduction

Vertebrates are aerobic organisms

They have a closed circulatory system and a mechanism for extraction of O_2 from air (or water), and release of carbon dioxide (CO_2) in waste products. Inspired O_2 leads to an efficient utilization of metabolic fuels, such as glucose and fatty acids; expired CO_2 is a major product of cellular metabolism. This utilization of O_2 as a metabolic substrate is accompanied by the generation of reactive oxygen species (ROS) that are capable of damaging virtually all biological macromolecules (Chapter 37). Organisms protect themselves from radical damage in several ways: sequestering O_2, limiting production of ROS, and detoxifying them. **Heme proteins**, interestingly, participate in these protective mechanisms by sequestering and transporting oxygen. The major heme proteins in mammals are myoglobin (Mb) and hemoglobin (Hb). Mb is found primarily in skeletal and striated muscle, and serves to store O_2 in the cytoplasm and deliver it on demand to the mitochondrion. Hb is restricted to erythrocytes where it facilitates the transport of O_2 and CO_2 between the lungs and peripheral tissues. This chapter presents the molecular features of Mb and Hb, the biochemical and physiologic relationships between the structures of Mb and Hb and their interaction with O_2 and other small molecules, and the pathologic aspects of selected Hb mutations.

Properties of oxygen

Most oxygen in the body is bound to a carrier protein containing heme

Photosynthetic organisms release diatomic oxygen into the earth's atmosphere during energy production, contributing to the current level of 21% oxygen in air. In mixtures of gases. Each component makes a specific contribution, known as its partial pressure (Dalton's Law), that is directly proportional to its concentration. It is also customary to use the partial pressure of a gas as a measure of its concentration in physiologic fluids. For atmospheric O_2 at a barometric pressure (sea level) of 760 mmHg or torr (101.3 kP or kPa; 1 atmosphere absolute or ATA), the partial pressure of oxygen, pO_2, is 150–

160 mmHg. The amount of O_2 in solution is, in turn, directly proportional to its partial pressure. Thus, in the arterial blood (37°C, pH 7.4) the pO_2 is 100 mmHg, which produces a concentration of dissolved O_2 of 0.13 mmol/L. This level of dissolved O_2, however, is inadequate to support efficient aerobic metabolism.

Rather, the major fraction of O_2 transported in blood and stored in muscle is complexed with the iron (ferrous, Fe^{2+}) proteins Hb and Mb, respectively. Hb is a tetrameric protein with four O_2-binding sites (heme groups). In arterial blood with a Hb concentration of 150 g/L (2.3 mmol/L) and O_2 saturation of 97.4%, the contribution of protein-bound O_2 is about 8.7 mmol/L. This concentration represents a dramatic 67-fold increase over physically dissolved O_2. The total oxygen-carrying capacity of the arterial blood, in dissolved and protein-bound forms, is 8.8 mmol/L – almost 200 mL of dissolved oxygen per liter of blood.

Characteristics of mammalian globin proteins

Globins constitute an ancient family of soluble metalloproteins

Globins are ubiquitous family of proteins, found in microorganisms, plants, invertebrates, and vertebrates. Present-day globins, with their spectacular diversity of function, are most likely derived from a single ancestral globin. While the extent of amino acid identity among invertebrate and vertebrate globins varies widely and can often appear random, two features are noteworthy: the invariant residues PheCD1 and HisF8 and the characteristic patterns of hydrophobic residues in helical segments (Fig. 5.1). Human Mb consists of a single globin polypeptide (153 amino acid residues, 17,053 Da). Human Hb is a tetrameric assembly of two α-globin polypeptides (141 residues, 15,868 Da) and two β-globin polypeptides (146 residues, 15,126 Da). A single heme prosthetic group is noncovalently associated with each globin apoprotein.

FIG. 5.1 Human globin amino acid sequences are highly conserved.
An alignment of human globins is depicted, with identical amino acid residues in orange. The two residues in green, PheCD1 (F) and HisF8 (H), are absolutely conserved in all metazoan globins. Helical segments in myoglobin are identified by the blue bars. Mb, myoglobin; Hbα, α-globin; Hbβ, β-globin; HbγA, γA-globin; Cygb, cytoglobin; Ngb, neuroglobin.

The secondary structure of mammalian globins is dominated by a high proportion of α-helix, with over 75% of the amino acids associated with eight helical segments. These α-helices are organized into a tightly packed, nearly spherical, tertiary structure, designated the globin fold (Fig. 5.2). So universal is this overall tertiary structure among all globins that the conventional nomenclature for globin residues follows that defined initially for sperm whale Mb, namely helices A, B, C, etc., starting at the *N*-terminus, separated by corners

AB, BC, etc., with residues numbered within each helix and corner. For example, residue A14, an amino acid that participates in electrostatic stabilization between helix A and the GH corner, corresponds to Lys^{15} in insect Hb, Lys^{16} in Mb and α-globin, and Lys^{17} in β-globin (arrow, lower center, Fig. 5.2).

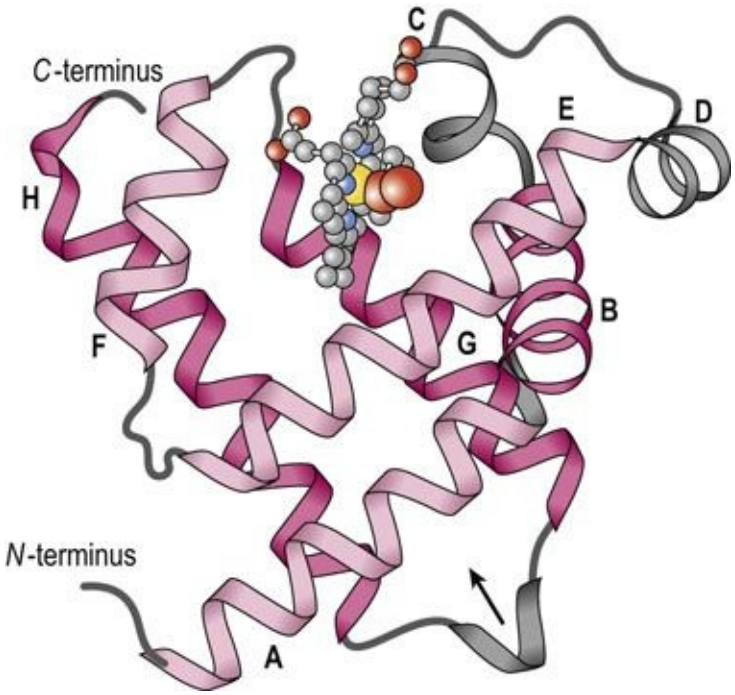


FIG. 5.2 Myoglobin is a compact globular protein.
In this drawing of mammalian Mb only the globin polypeptide backbone is shown, with emphasis on the high proportion of secondary structure (exclusively α-helix). The two-layer, three-over-three arrangement of α-helices is highlighted by the light and dark shades of red. The heme group is illustrated as a gray 'ball-and-stick' structure, with iron (yellow) and a bound oxygen molecule (red).

Polar amino acids are located almost exclusively on the exterior surface of globin polypeptides and contribute to the remarkably high solubility of these proteins (e.g. 370 g/L (37% protein solution; 5.7 mmol/L) Hb in the erythrocyte). Amino acids that are both polar and hydrophobic, such as threonine, tyrosine and tryptophan, are oriented with their polar functions toward the protein's exterior. Hydrophobic residues are buried within the interior, where they stabilize the folding of the polypeptide and form a pocket that accommodates the heme prosthetic group. Notable exceptions to this general

distribution of amino acid residues in globins are the two histidines that play indispensable roles deep within the heme pocket (Fig. 5.3). Their side chains are oriented perpendicular to and lay on either side of the heme prosthetic group. One of the side chain imidazole nitrogens of the invariant proximal histidine (HisF8) is close enough to bond directly to the pentacoordinate Fe^{2+} atom. On the opposite side of the heme plane the distal histidine (HisE7), which is too far from the heme iron for direct bonding, functions to stabilize bound O_2 by hydrogen bonding.

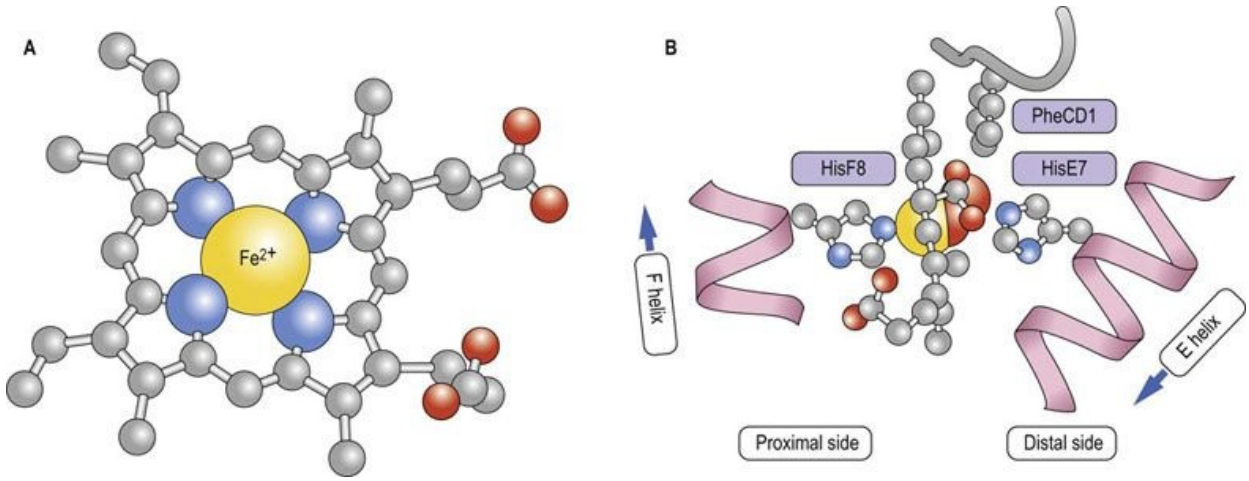


FIG. 5.3 Heme is a complex of porphyrin and iron.
(A) In this view the carbon framework of protoporphyrin IX, a conjugated tetrapyrrole ring, is shown in gray; O_2 molecules are red. Iron (yellow sphere) prefers six ligands in an octahedral coordination geometry; pyrrole nitrogen atoms (blue spheres) provide four of these. PheCD1 makes critical hydrophobic and electrostatic stacking interactions with the porphyrin ring. **(B)** In the oxygenated globin structure, the planar heme is positioned between the proximal and distal histidines (HisF8 and HisE7, respectively); only HisF8 has an imidazole nitrogen (blue sphere) close enough to bond with iron. The α-helices that contain these histidines are shown in pink. In deoxygenated globins, the sixth position remains vacant, leaving a pentacoordinated iron. In the oxygenated state, O_2 occupies the sixth position. Both porphyrin propionate moieties participate in hydrogen and electrostatic bonding interactions with globin side chains and solvent.

Structure of the heme prosthetic group

Heme, the O_2-binding moiety common to Mb and Hb, is a porphyrin molecule to which an iron atom (Fe^{2+}) is

coordinated

The Fe-porphyrin prosthetic group of heme is planar and hydrophobic, with the exception of two propionate groups which are exposed to solvent. Heme becomes an integral component of the holoprotein during polypeptide synthesis; it is heme that gives globins, their characteristic purple-red color – purple in the deoxygenated state in venous blood, red in the oxygenated state in arterial blood.

Globins increase the aqueous solubility of the otherwise poorly soluble, hydrophobic heme prosthetic group. Once sequestered inside a hydrophobic pocket created by the folded globin polypeptide, heme is in a protective environment that minimizes the spontaneous oxidation of Fe^{2+} (ferrous) to Fe^{3+} (ferric: rusting) in the presence of O_2. Such an environment is also essential for globins to bind and release O_2. Should the iron atom become oxidized to the ferric state, heme can no longer interact reversibly with O_2, compromising its function in O_2 storage and transport.

Myoglobin: an oxygen storage protein

Mb binds O2 that has been released from Hb in tissue capillaries and subsequently diffused into tissues

Located in the cytosol of skeletal, cardiac and some smooth muscle cells, Mb stores a supply of O_2 that is readily available to cellular organelles, particularly the mitochondrion, that carry out oxidative metabolism. With its single ligand-binding site, the reversible reaction of Mb with O_2:

$$\mathrm{Mb} + \mathrm{O_2} \rightleftharpoons \mathrm{Mb{\cdot}O_2}$$

may be described by the following equations:

$$K_a = [\mathrm{Mb{\cdot}O_2}] / [\mathrm{Mb}][\mathrm{O_2}]$$

$$Y = [\mathrm{Mb{\cdot}O_2}] / \{[\mathrm{Mb{\cdot}O_2}] + [\mathrm{Mb}]\}$$

where K_a is an affinity or equilibrium constant, and Y is the fractional O_2 saturation. Combining these two equations, expressing the concentration of O_2 in terms of its partial pressure (pO_2), and substituting the term P_{50} for $1/K_a$ yields the equation for the O_2 saturation curve of Mb:

$$Y = pO_2 / \{pO_2 + P_{50}\}$$

By definition, the constant P_{50} is the value of pO_2 at which $Y = 0.5$ or half the ligand sites are occupied by O_2. In a plot of Y versus pO_2, the equation for ligand binding by Mb describes a hyperbola (Fig. 5.4) with a P_{50} of 4 mmHg. The low value of P_{50} reflects a high affinity for O_2. In the capillary beds of muscle tissues, pO_2 values are in the range of 20–40 mmHg. Predictably, working muscles exhibit lower pO_2 values than muscles at rest. With its high affinity for O_2, myocyte Mb readily becomes saturated with O_2 that has entered from the blood. As O_2 is consumed during aerobic metabolism, it dissociates through mass action from Mb and diffuses into mitochondria, the power plants of the muscle cell. Whales and other diving mammals have unusually high concentrations of Mb in their muscle tissue; the Mb is thought to function as a large oxygen reservoir for prolonged periods under water.

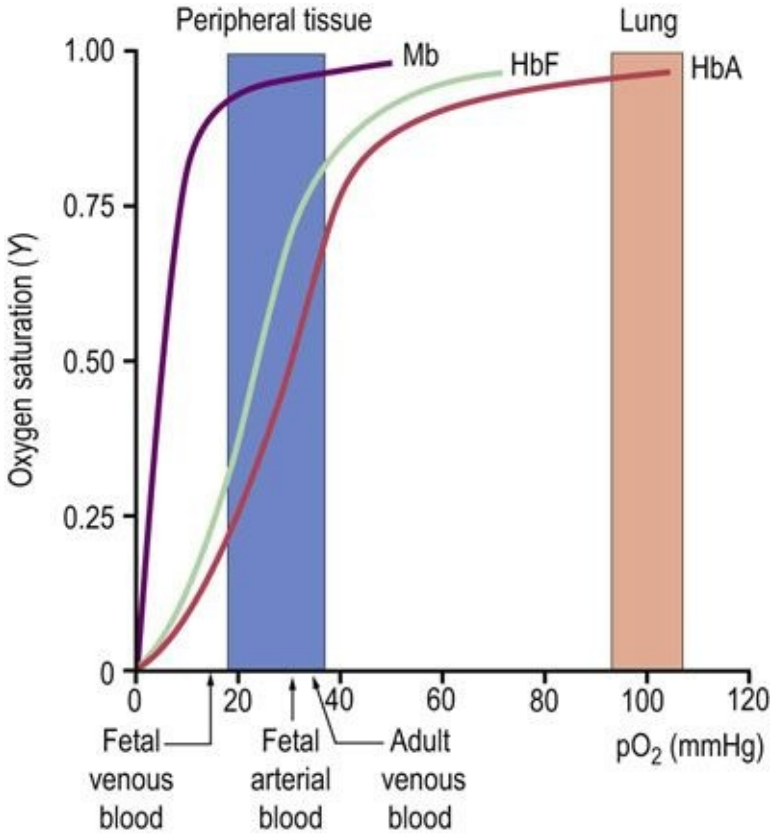


FIG. 5.4 Oxygen saturation curves of myoglobin and hemoglobin.
Mb and Hb have different O_2 saturation curves. The fractional saturation (Y) of O_2-binding sites is plotted against the concentration of O_2 (pO_2 (mmHg)). Curves are shown for Mb, fetal Hb (HbF), and adult Hb (HbA). Also indicated, by arrows and shading, are the normal levels of O_2 measured in various adult and fetal blood samples.

Clinical box Hyperbaric O_2 therapy for acute carbon monoxide poisoning

A 22-year-old pregnant woman, carrying a fetus of 31 weeks gestational age, was brought to the maternity clinic of a hospital for suspected CO poisoning.

The patient was experiencing headache, nausea, and visual abnormalities. She stated that her workplace had been undergoing repairs to the heating and ventilation systems during the past 2 weeks, and on the day of her hospital visit the fire department had evacuated the building after detecting a high level of CO

(200 ppm), compared to a typical urban street level of 10 ppm. Her blood pressure was 116/68 mmHg, pulse rate of 100, and respiratory rate of 24/min. Noteworthy in the patient's evaluation was a carboxy-Hb (COHb) component of 15% of total Hb at time of admission (normal = 3%, but may exceed 10% in heavy smokers). Fetal monitoring indicated a fetal heart rate of 135, with occasional, moderate irregularities. Uterine contractions were occurring every 3–5 min.

The patient was treated in the hospital's hyperbaric O_2 chamber: 30 min at 2.5 ATA, then 60 min at 2.0 ATA. She also received magnesium sulfate intravenously to resolve the premature contractions. The patient was discharged 2 days later. She delivered a healthy female infant at 38 weeks of gestational age who, on examination at birth and at 6 weeks of age, exhibited no apparent sequelae to her in utero exposure to CO or 100% O_2.

Comment.

Carbon monoxide is a normal product of heme catabolism and has a range of physiologic activities in vascular, neuronal and immunologic systems. Like O_2, CO also binds to heme prosthetic groups. Because the affinity of globin-bound heme for CO is about 250 times that for O_2, prolonged exposure of hemoglobin to exogenous CO would be virtually irreversible ($t_{1/2}$ for reversal in blood, 4–8 h) and lead to highly toxic levels of carboxy-Hb. Hyperbaric O_2 is the treatment of choice for severe or complicated CO poisoning.

The administration of 100% O_2 at 2–3 ATA creates arterial and tissue pO_2 values of 2000 mmHg and 400 mmHg, respectively (20 times normal). The immediate result is a reduction in the $t_{1/2}$ of carboxy-Hb to less than 30 min. Hyperbaric O_2 is also used in the treatment of decompression sickness, arterial gas embolism, radiation-induced or ischemic tissue injury, and severe hemorrhage.

Hemoglobin: an oxygen transport protein

Hb is the principal O_2-transporting protein in human blood; it is localized exclusively in erythrocytes

Adult Hb (HbA) is a tetrahedral array of two identical α-globin, and two identical β-globin subunits, a geometry that predicts several types of subunit–subunit interactions in the quaternary structure (Fig. 5.5). Importantly, within the Hb tetrahedron each subunit is in contact with the other three. Experimental analysis of the quaternary structure indicates multiple noncovalent interactions (hydrogen bonds and electrostatic bonds) between each pair of dissimilar subunits, *i.e.* at the α–β interfaces. In contrast, there are fewer and predominantly hydrophobic interactions between identical subunits, at the α_1–α_2 or β_1–β_2 interfaces. The actual number and nature of contacts differ in the presence or absence of O_2. Strong associations within each αβ heterodimer and at the interface between the two heterodimers (see Fig. 5.5) are now recognized as major factors determining O_2 binding and release. Thus, Hb is more appropriately considered a dimer of heterodimers, $(\alpha\beta)_2$, rather than an $\alpha_2\beta_2$ tetramer. Although a solution of HbA is theoretically a dynamic mixture of heterodimers and tetramers, under physiologic conditions (high Hb and neutral pH) the equilibria greatly favor the tetramer: 99% for oxygenated Hb, and 99.9% for deoxygenated Hb.

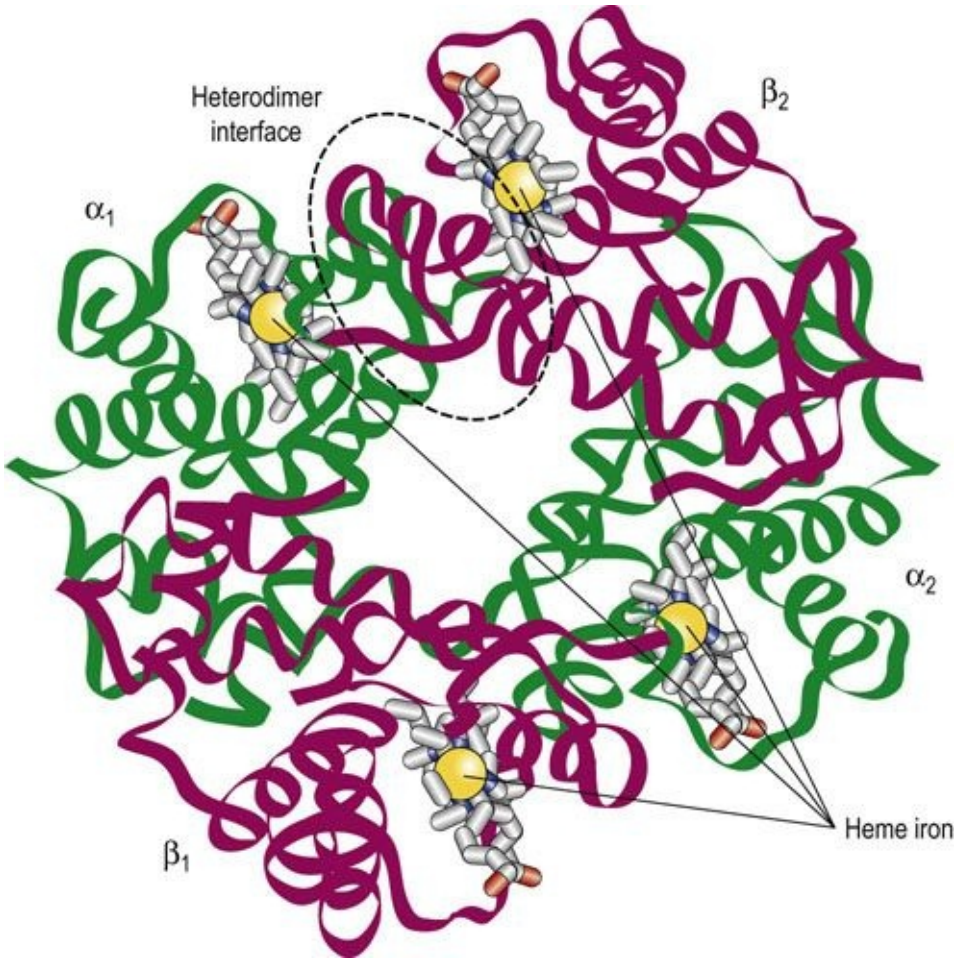


FIG. 5.5 Hemoglobin is a tetramer of four globin subunits.
Hb is a tetrahedral complex of two identical α-globins (α_1 and α_2, greens) and two identical β-globins (β_1 and β_2, reds). With this geometry each globin subunit contacts the other three subunits, creating the interfaces and interactions that define cooperativity. One of the heterodimer interfaces is outlined in a dashed oval.

Interactions of hemoglobin with oxygen

Hb binds oxygen cooperatively, with a Hill coefficient of ~2.7

As a gas delivery vehicle, Hb must be able to bind O_2 efficiently as it enters the lung alveoli during respiration and to release O_2 to the extracellular environment with similar efficiency as erythrocytes circulate through tissue capillaries. This remarkable duality of function is achieved by cooperative interactions among globin subunits. When deoxygenated Hb becomes oxygenated, significant structural changes extend throughout the protein molecule. In the heme pocket,

as a consequence of O_2 coordination to iron and a new orientation of atoms in the heme structure, the proximal histidine and helix F to which it belongs, shift their positions (see Fig. 5.3). This subtle conformational change triggers major structural realignments elsewhere within that globin subunit. In turn, these tertiary structural changes are transmitted, even amplified, in the overall quaternary structure, such that a 12–15° rotation and a 0.10 nm displacement of the $\alpha_1\beta_1$ dimer relative to the $\alpha_2\beta_2$ dimer take place. Because of the inherent asymmetry of the $\alpha_2\beta_2$ tetramer, these combined motions result in quite dramatic changes within and, more importantly, between the αβ heterodimers. Because of structural changes in hemoglobin as a result of binding of oxygen and other effectors, the binding affinity for subsequent molecules of oxygen may be increased (positive cooperativity) or decreased (negative cooperativity).

Hb can bind up to four molecules of O_2 in a cooperative manner

With its multiple ligand-binding sites and structural changes in response to binding, the oxygen affinity and the fractional saturation of Hb are more complex functions than those of Mb. Consequently, the equation for the fractional O_2 saturation curve must be modified to:

$$Y = {pO_2}^n / \left\{ {pO_2}^n + {P_{50}}^n \right\}$$

where n is the Hill coefficient. In a plot of Y versus pO_2 when $n > 1$, the equation for ligand binding describes a sigmoid (S-shaped) curve (see Fig. 5.4). The Hill coefficient, determined experimentally, is a measure of cooperativity among ligand-binding sites, *i.e.* the extent to which the binding of O_2 to one subunit influences the affinity of O_2 to other subunits. For fully cooperative binding, n is equal to the number of sites (four in Hb), an indication that binding at one site maximally enhances binding at other sites in the same molecule. The normal Hill coefficient for adult Hb ($n = 2.7$) reflects strongly cooperative ligand binding. Hb has a considerably lower affinity for O_2, reflected in a P_{50} of 27 ± 2 mmHg, compared to myoglobin (P_{50} = 4 mmHg). In the absence of cooperativity, even with multiple sites, the Hill coefficient would be 1, *i.e.* binding of one molecule of O_2 would not influence the binding of other

molecules. Decreased or absent cooperativity is observed for Hb mutants that have lost functional subunit–subunit contacts (Table 5.1). The steepest slope of the saturation curve for Hb lies in a range of pO_2 that is found in most tissues (see Fig. 5.4). Thus, relatively small changes in pO_2 will result in considerably larger changes in the interaction of Hb with O_2. Accordingly, slight shifts of the curve in either direction will also dramatically influence O_2 affinity.

Table 5.1

Classification and examples of hemoglobinopathies

Classification	Common name	Mutation	Frequency	Biochemical changes	Clinical consequences
Abnormal solubility	HbC	$Glu^{6(\beta)} \rightarrow Lys$	Common	Cellular crystallization of oxygenated protein; increased fragility	Mild hemolytic anemia; splenomegaly (enlarged spleen)
Decreased O_2 affinity	Hb Titusville	$Asp^{94(\alpha)} \rightarrow Asn$	Very rare	Heterodimer interface altered to stabilize T state; decreased cooperativity	Mild cyanosis (blue-purple skin coloration from deoxygenated blood)
Increased O_2 affinity	Hb Helsinki	$Lys^{82(\beta)} \rightarrow Met$	Very rare	Reduced binding of 2,3-BPG in T state	Mild polycythemia (increased erythrocyte count)
Ferric heme (methemoglobin)	HbM Boston	$His^{58(\alpha)} \rightarrow Tyr$	Occasional	Altered heme pocket (mutation of distal His)	Cyanosis of skin and mucous membranes; decreased Bohr effect
Unstable protein	Hb Gun Hill	$\Delta\beta 91–95$	Very rare	Misfolding caused by loss of Leu in heme pocket and shorter helix	Formation of Heinz bodies (inclusions of denatured Hb); jaundice (yellow coloration of integument and sclera); pigmented urine
Abnormal synthesis	Hb Constant Spring	$Tyr^{142(\alpha)} \rightarrow Gln$	Very rare	Loss of termination codon; decreased MRNA stability	α-Thalassemia (hemolytic anemia, splenomegaly and jaundice)

Hemoglobinopathies are usually classified according to the most prominent change to the protein's structure, function, or regulation. Initial identification of a mutation often involves electrophoretic or chromatographic analysis, as shown in Figure 5.9 for HbSC, a double heterozygous genotype associated with a sickle cell disease-like phenotype. Δ = deletion mutant.

Hemoglobin subunits may assume two different conformations that differ in O_2 affinity

The mechanism underlying the cooperativity in oxygen binding by hemoglobin involves a shift between two conformational states of the hemoglobin molecule which differ in oxygen affinity. These two quaternary conformations are known as the T (tense) and R (relaxed) states, respectively. In the T state, interactions between the heterodimers are stronger; in the R state, these noncovalent bonds are, in summation, weaker. O_2 affinity is lower in the T state and higher in the R state. Transition between these states is accompanied by the breaking of existing

noncovalent bonds and formation of new ones at the heterodimer interfaces (Fig. 5.6). Contact between the two αβ heterodimers (see Fig. 5.5) is stabilized by a mixture of hydrogen and electrostatic bonds. Approximately 30 amino acids participate in the noncovalent interactions that characterize the deoxygenated and oxygenated Hb conformations.

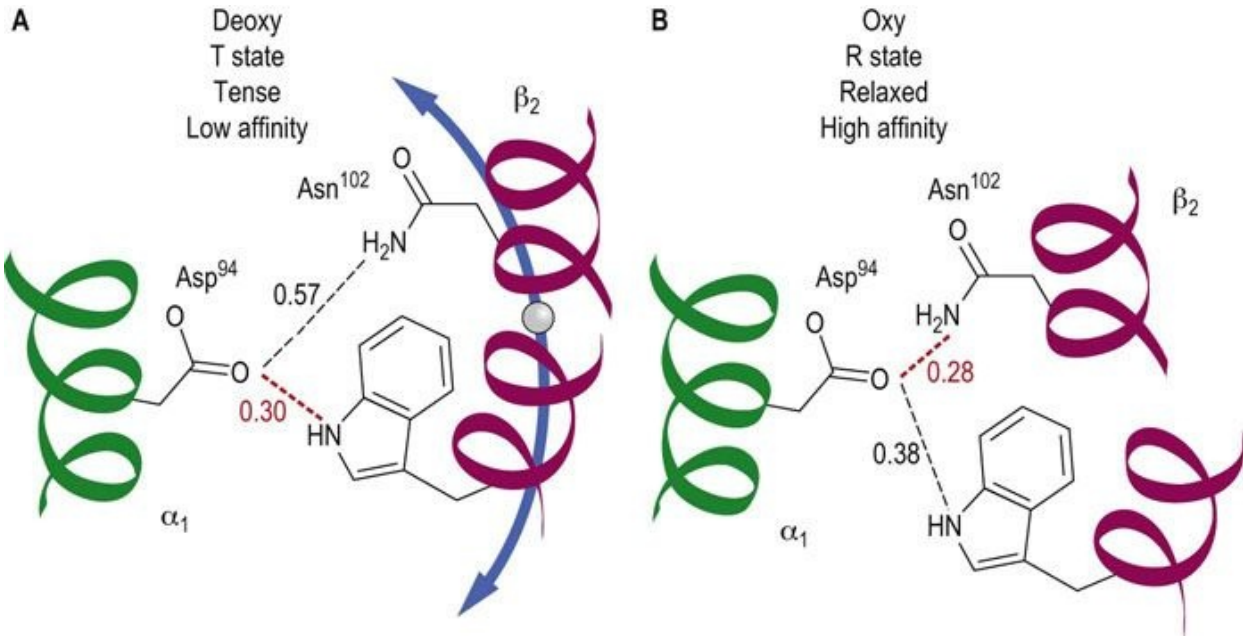


FIG. 5.6 Noncovalent bonds differ in deoxygenated and oxygenated hemoglobin. In the middle of the interface between the two αβ heterodimers are the residues $Asp^{94(\alpha)}$ on the α_1-globin of one heterodimer and $Trp^{37(\beta)}$ and $Asn^{102(\beta)}$ on the β_2-globin of the other heterodimer (see dashed oval in Fig. 5.5). Each has side chain atoms capable of noncovalent interactions. **(A)** In the deoxygenated T state the distance between the Asp and Trp residues favors a hydrogen bond, whereas the distance between Asp and Asn is too great. **(B)** As a result of the conformational changes that accompany the transition to the oxygenated R state, the distance between Asp and Trp is now too large, but that between Asp and Asn is compatible with formation of a new hydrogen bond. Elsewhere along this interface, other bonds are created and broken. An identical alignment of residues and noncovalent interactions is found between the α_2- and β_1-globin monomers. Distances are shown in nm. Hydrogen bonds are commonly 0.27–0.31 nm in length.

Several models have been developed to describe the transition between the T and R states of Hb. At one extreme is a model in which each Hb subunit sequentially responds to O_2 binding with a conformational change, thereby permitting hybrid intermediates of the T and R states. At the opposite extreme is

a model in which all four subunits switch concertedly; hybrid states are forbidden, and binding of O_2 to one subunit shifts the equilibrium of all subunits from the T to the R state simultaneously. The molecular structures of deoxygenated and partially and fully ligated Hb have been studied extensively by a broad range of thermodynamic and kinetic techniques. Yet, progress toward reconciling inconsistencies among classic and more recent models has been slow. These different viewpoints on conformational changes in multi-subunit proteins are discussed further in the section on allosteric enzymes in Chapter 6.

Clinical test box Pulse oximetry

Pulse oximetry ('pulse-ox') is a noninvasive method of estimating the oxygen saturation of arterial Hb. Two physical principles are involved: first, the visible and infrared spectral characteristics of oxy-and deoxy-Hb are different; second, arterial blood flow has a pulsatile component that results from volume changes with each heartbeat. Transmission or reflectance measurements are made in a translucent tissue site with reasonable blood flow, commonly a finger, toe or ear of adults and children, or a foot or hand of infants. The photodetector and microprocessor of the pulse oximeter permit a calculation of oxygen saturation (SpO_2 = saturation of peripheral oxygen) that typically correlates within 4–6% of the value found by arterial blood gas analysis.

Pulse oximetry is used to monitor the cardiopulmonary status during local and general anesthesia, in intensive care and neonatal units, and during patient transport. Body movement, radiated ambient light, elevated bilirubin, artificial or painted fingernails can interfere with pulse oximetry.

Conventional, two-wavelength instruments 'assume' that the optical measurements are associated with oxygenated and deoxygenated hemoglobins; they cannot discriminate among oxy-, carboxy-and metHb. Newer technologies, however, utilize six or eight wavelengths and permit multiple Hb species discrimination with an accuracy of ±2.0% and precision of ±1.0%.

Allosteric Modulation of the Oxygen Affinity of Hemoglobin

Allosteric proteins and effectors

Hb is an allosteric protein; its affinity for O2 is regulated by small molecules

Hb is one of the best studied examples of an allosteric protein, These small bind to proteins at sites that are spatially distinct from the ligand-binding sites, thus their designation as **allosteric (other site) effectors**. Through long-range conformational effects, they alter the ligand or substrate binding affinity of the protein. Allosteric proteins are typically multi-subunit proteins. The O_2 binding affinity of Hb is affected positively by O_2, as well as by a number of chemically diverse allosteric effectors, including H^+, CO_2, and 2,3-bisphosphoglycerate (2,3-BPG) (Fig. 5.7). When an allosteric effector affects its own binding to the protein (at other sites), the process is termed **homotropic**, *e.g.* the effect of binding of O_2 at one site on Hb enhances the affinity for binding of O_2 to other sites on Hb. When the allosteric effector is different from the ligand whose binding is altered, the process is termed **heterotropic**, *e.g.* the effect of H^+ (pH) on the P_{50} for oxygen binding to Hb. These interactions lead to horizontal shifts in the O_2 binding curves (see Fig. 5.7).

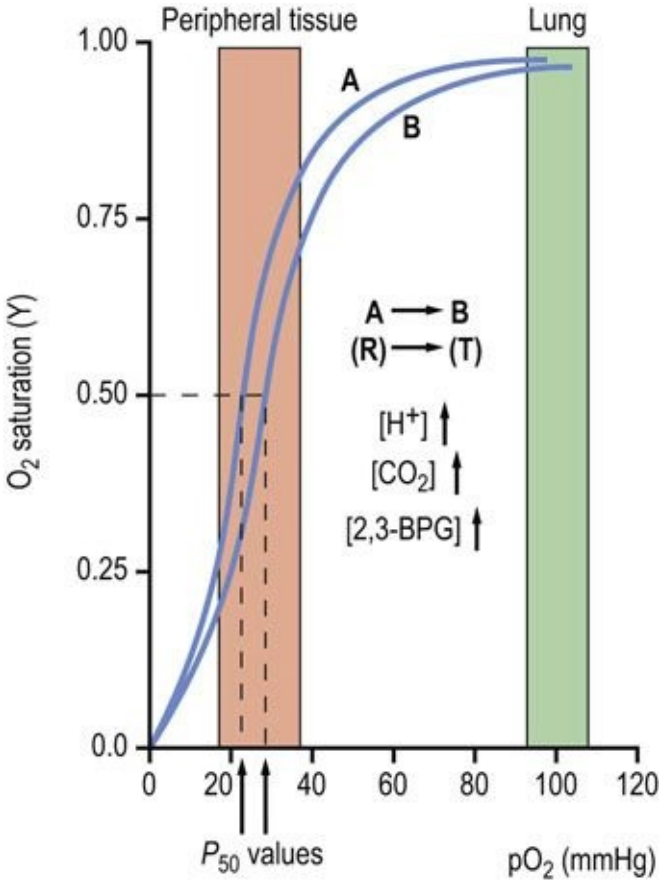


FIG. 5.7 Allosteric effectors decrease the oxygen affinity of hemoglobin.
O_2 interaction with Hb is regulated by allosteric effectors. Under physiologic conditions HbA exhibits a highly cooperative O_2 saturation curve. With an increase in the erythrocyte concentration of any of three allosteric effectors, H^+, CO_2 or 2,3-bisphosphoglycerate (2,3-BPG), the curve shifts to the right **(position B),** indicating a decreased affinity for O_2 (increase in P_{50} value). Actions of the effectors that modulate O_2 affinity appear to be additive. Conversely, a decrease in any of the allosteric effectors shifts the curve to the left **(position A).** Increasing temperature will also shift the curve to the right. The sensitivity of O_2 saturation to H^+ is known as the Bohr effect. Normal ranges of O_2 measured in pulmonary and peripheral tissue capillaries are indicated by shaded areas.

Bohr effect

Acidic pH (protons) decreases the O_2 affinity of Hb

The O_2 affinity of Hb is exquisitely sensitive to pH, a phenomenon known as the Bohr effect. The Bohr effect is most readily described as a right shift in the O_2

saturation curve with decreasing pH. Thus, an increased concentration of H^+ (decreased pH) favors an increased P_{50} (lower affinity) for O_2 binding to Hb, **equivalent to an H^+-dependent shift of Hb from the R to the T state**.

To understand the Bohr effect at the level of protein structure and to appreciate the role of H^+ as a heterotropic allosteric effector, it is important to recall that Hb is a highly charged molecule. The residues that participate in the Bohr effect include the *N*-terminal Val amino group of α-globin and the *C*-terminal His side chain of β-globin. The pK_a values of these weak acids differ sufficiently between the deoxygenated and oxygenated forms of Hb to cause the uptake of 1.2–2.4 protons by the deoxygenated, compared to oxygenated, tetramer.

Identification of specific amino acid residues of the α-and β-globins that participate in the Bohr effect is complicated by differential interactions of other charged solutes with deoxy-and oxy-Hb. Thus, a preferential binding of a given anion, *i.e.* Cl^- and/or organic phosphates, to deoxygenated Hb involves the alteration of the p*K*s of some cationic groups, thereby contributing to the overall observed Bohr effect. For example, there is compelling evidence showing that $Val^{1(\alpha)}$ is relevant to the Bohr effect only in the presence of Cl^-. The pK_a of this group shifts from 8.0 in deoxygenated Hb to 7.25 in oxygenated Hb in the presence of physiologic concentration of Cl^- (≈100 mmol/L). Further, the participation of the $Val^{1(\alpha)}$ groups in the chloride-dependent Bohr effect is strongly modulated by CO_2 because of the formation of CO_2 (carbamino) adducts of Hb (described below).

As Hb binds O_2, protons dissociate from selected weak acid functions. Conversely, in acidic media, protonation of the conjugate bases inhibits O_2 binding

During their circulation between pulmonary alveoli and peripheral tissue capillaries, erythrocytes encounter markedly different conditions of pO_2 and pH. The high pO_2 in the lungs promotes ligand saturation and forces protons from the Hb molecule to stabilize the R state. In the capillary bed, particularly in metabolically active tissues, the pH is slightly lower, due to the production of acidic metabolites, such as lactate. Oxygenated Hb, upon entering this environment, will acquire some ‘excess’ protons and shift toward the T state, promoting release of O_2 for uptake by tissues for aerobic metabolism.

Effects of CO_2 and temperature

Like H^+, CO_2 is is increased in venous capillaries and is a negative allosteric effector of the O_2 affinity of Hb

Closely related to the Bohr effect is the ability of CO_2 to alter the O_2 affinity of Hb. The increase in pCO_2 in venous capillaries decreases the affinity of Hb for O_2. Accordingly, a right shift in the ligand saturation curve occurs as pCO_2 increases. It should be emphasized that the allosteric effector is, in fact, CO_2, not HCO_3^-: CO_2 reacts reversibly with the unprotonated *N*-terminal amino groups of the globin polypeptides to form **carbamino adducts**:

$$\text{Hb} - \text{NH}_2 + \text{CO}_2 \rightleftharpoons \text{Hb} - \text{NHCOO}^- + \text{H}^+$$

This transient covalent chemical modification of Hb is not only a specialized example of allosteric control, resulting in a stabilization of deoxygenated Hb; it also represents one form of transport of CO_2 to the lungs for clearance from the body. Between 5% and 10% of the total CO_2 content of blood exists as carbamino adducts.

There is a strong physiologic correlation between pCO_2 and the O_2 affinity of Hb. CO_2 is a major product of mitochondrial oxidation and, like H^+, is particularly abundant in metabolically active tissues. Upon diffusing into blood, CO_2 can react with oxygenated Hb, shift the equilibrium toward the T state, and thereby promote the dissociation of bound O_2 (see Fig. 5.7). The vast majority of peripheral tissue CO_2, however, is hydrated by erythrocyte **carbonic anhydrase** to carbonic acid (H_2CO_3), a weak acid that dissociates partially to H^+ and HCO_3^-:

$$\text{CO}_2 + \text{H}_2\text{O} \rightleftharpoons \text{H}_2\text{CO}_3 \text{ enzyme-catalyzed reaction}$$

$$\text{H}_2\text{CO}_3 \rightleftharpoons \text{H}^+ + \text{HCO}_3^- \text{ spontaneous dissociation}$$

Interestingly, from both carbamino adduct formation and

hydration/dissociation reactions involving CO_2, an additional pool of protons is generated. These are protons that become available to participate in the Bohr effect and facilitate O_2–CO_2 exchange. During its return to the lungs, blood transports two forms of CO_2: carbamino-Hb and the H_2CO_3/HCO_3^- acid–conjugate base pair. Blood and Hb are now exposed to a low pCO_2, and through mass action the carbamino adduct formation is reversed and binding of O_2 is again favored. Similarly, in the pulmonary capillaries, erythrocyte carbonic anhydrase converts H_2CO_3 to CO_2 and H_2O, which are expired into the atmosphere (see Chapter 25).

Advanced concept box Artificial hemoglobins

The supply-and-demand curves for whole blood and packed red cell availability and utilization point to an impending crisis and the need to develop alternatives. Red cell substitutes are transfusion alternatives that are potentially useful during major surgical procedures and hemorrhagic shock emergencies.

Three types of artificial O_2 carriers have been investigated: Hb-based oxygen carrier (HBOC), liposome-or nanoparticle-encapsulated Hb, and perfluorocarbon emulsions. HBOCs are hemoglobins derived from allogeneic, xenogeneic or recombinant sources that have been modified by polymerization, crosslinkage or conjugation. These modifications facilitate purification and sterilization, and minimize toxicity and immunogenicity. They are also necessary to stabilize the extracellular Hb tetramers; otherwise the hemoglobin dissociates into dimers and monomers in plasma and is excreted in urine. The artificial forms have O_2 affinity (P_{50}) in the range 16–38 mmHg, but usually have diminished cooperativity (n = 1.3–2.1) and Bohr effects.

Several HBOCs have progressed through extensive clinical evaluation and some are used in medical procedures in some countries. Adverse effects are not uncommon with HBOCs. Increased vasoconstriction with subsequent hypertension occurs, a result of increased binding of nitric oxide (NO), an endogenous regulator of vasodilation, by extracellular Hb. Other problems include increased heme oxidation to metHb, elevated iron

deposition in tissues, gastrointestinal distress, neurotoxicity, and interference with diagnostic measurements. Molecular engineering of human Hb, now underway in a number of laboratories, seeks to improve O_2 binding, allosteric properties, and side effects of HBOCs. Packaging of hemoglobin in liposomes or nanocapsules, producing artificial red blood cells, is also a promising technology since this limits the escape of Hb into extravascular spaces.

Working muscles not only produce the allosteric effectors H^+ and CO_2 as byproducts of aerobic metabolism but also liberate heat. Because the binding of O_2 to heme is an exothermic process, **the O_2 affinity of Hb decreases with increasing temperature**. Thus, the microenvironment of an exercising muscle profoundly favors a more efficient release of Hb-bound O_2 to the surrounding tissue.

Effect of 2,3-bisphosphoglycerate

2-3-Bisphoglycerate (2,3-BPG), an intermediate in carbohydrate metabolism, is an important allosteric effector of Hb

2,3-BPG is synthesized in human erythrocytes in a one-step shunt off the glycolytic pathway (Chapter 12). Like H^+ and CO_2, 2,3-BPG is an indispensable negative allosteric effector that, when bound to Hb, causes a marked increase in P_{50} (see Fig. 5.7). Were it not for the high erythrocyte concentration of 2,3-BPG (4.1 mmol/L, nearly equal to that of Hb), the O_2 saturation curve of Hb would approach that of Mb!

At one end of the twofold symmetry axis within the quaternary structure of Hb there is a shallow cleft defined by cationic amino acids of the juxtaposed β-globin subunits (Fig. 5.8). A single molecule of 2,3-BPG binds to this site. **A critical consequence of the conformational differences between the T and R states is that deoxygenated Hb preferentially interacts with the negatively charged 2,3-BPG.** Multiple electrostatic interactions stabilize the complex between the polyanionic effector and deoxygenated Hb. The cleft is too narrow

in fully oxygenated Hb to accommodate 2,3-BPG.

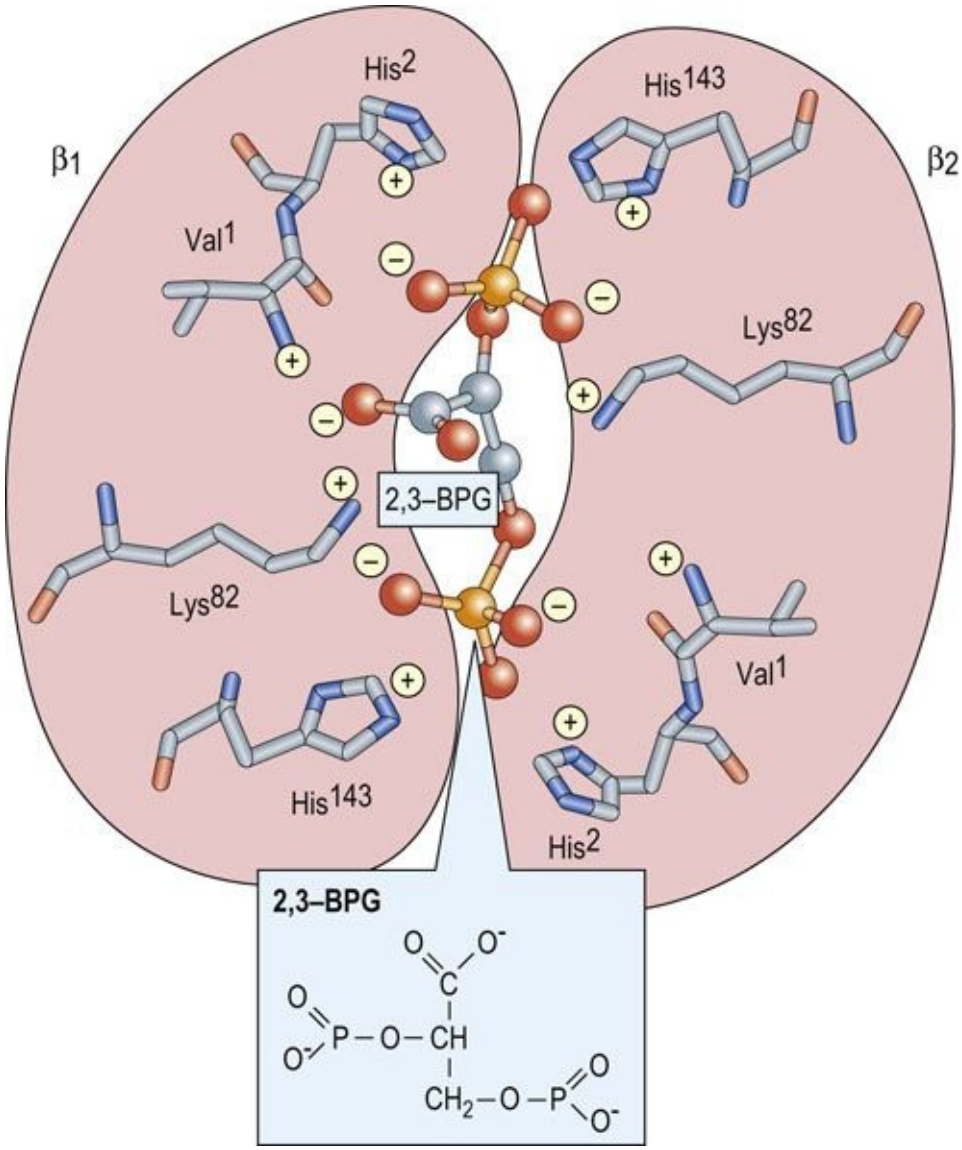


FIG. 5.8 2,3-Bisphosphoglycerate binds preferentially to deoxygenated hemoglobin. On the surface of the deoxygenated Hb tetramer where the two β-globins (purple) interact, there is a cleft formed by the N-terminal amino acid residue ($Val^{1(\beta)}$) and the side chains of $His^{2(\beta)}$, $Lys^{82(\beta)}$, and $His^{143(\beta)}$ (stick models). This site consists of eight cationic groups, sufficient to bind with high affinity one molecule of 2,3-BPG (ball-and-stick model; phosphorus, orange), a molecule with five anionic groups at physiologic pH. This array of positive charges does not exist in oxygenated Hb. In fetal Hb (HbF) $His^{143(\beta)}$ is replaced by a Ser residue.

The importance of 2,3-BPG as an allosteric effector is underscored by

observations that its concentration in the erythrocyte changes in response to various physiologic and pathologic conditions. During chronic hypoxia (decreased pO_2) secondary to pulmonary disease, anemia or shock, the level of 2,3-BPG increases. Such compensatory increases have also been described in cigarette smokers and on adaptation to high altitudes. The net result is a greater stabilization of the deoxygenated, low-affinity T state and a further shift of the saturation curve to the right, thereby facilitating release of more O_2 to tissues. Under most circumstances, the rightward shift has an insignificant effect on the O_2 saturation of Hb in the lungs.

Selected Topics

Interaction of hemoglobin with nitric oxide

Nitric oxide, a potent vasodilator, is stored on Hb as S-nitrosoHb (SNO-Hb)

Nitric oxide (NO) is a gaseous free radical capable of oxidative modification (nitration, nitrosation, nitrosylation) of biological macromolecules. Yet this highly reactive molecule, also known as **endothelium-derived relaxing factor (EDRF)**, is synthesized in endothelial cells and participates in normal vascular physiology, including vasodilation (smooth muscle), hemostasis (platelet), and adhesion molecule expression (endothelial cell). Erythrocytes are the largest intravascular reservoir of bioactive NO, and Hb is indispensable for its formation, storage, and release. SNO-Hb is the product of *S*-nitrosylation of the $Cys^{93\beta}$ side chains of Hb. These Cys thiol groups can accept NO by transfer from intracellular *S*-nitrosoglutathione or from heme-bound NO (nitrosyl-Hb). NO is released by exchange from SNO-Hb to Cys side chains of anion exchanger 1, an erythrocyte membrane protein that can then deliver NO to the plasma. The formation and breakdown of SNO-Hb are sensitive to pO; NO is released from Hb in response to hypoxia or on conversion to the T state, *e.g.* in venous capillaries.

Another remarkable process within the erythrocyte is the allosterically regulated conversion of nitrite (NO_2^-) to NO, a reaction performed by deoxygenated Hb. This intrinsic 'nitrite reductase' activity takes advantage of the moderate NO_2^- concentration in the erythrocyte (up to 0.3 µmol/L). While the chemistry is complex, the reaction is thought to yield a labile intermediate nitrosyl-metHb(ferric) that can readily transfer NO to $Cys^{93(\beta)}$ on oxygenated Hb.

Advanced Concept Box Acute mountain sickness – too high, too fast

Acute mountain sickness (AMS) develops in individuals who ascend rapidly to ambient conditions of hypobaric hypoxia. Symptoms include shortness of breath, rapid heart rate, headache, nausea, anorexia, and sleep disturbance. These can develop at

altitudes of 2000 m (25% incidence) and higher to 4000 m or more (50% incidence). The most severe form is high-altitude cerebral edema (2% incidence), a potentially fatal condition characterized by ataxia and other neuromuscular and neurologic problems.

At 4000 m the barometric pressure is 460 mmHg, leading to an ambient partial pressure of O_2 of 96 mmHg (sea level, 160). Physiologic calculations yield values of a tracheal pO_2 of 86 mmHg (sea level, 149), an alveolar pO_2 of 50 mmHg (sea level, 105), and an arterial pO_2 of 45 mmHg (sea level, 100). At this arterial partial pressure of O_2, Hb saturation is only 81% (see Fig. 5.4). Consequently, the O_2-carrying capacity of arterial blood decreases to 160 mL/L (sea level, 198). Hypoxia can also lead to overperfusion of vascular beds, endothelial leakage, and edema.

Humans adapt to high altitude (acclimatization) by several mechanisms. Hyperventilation is a critical short-term response that serves to decrease alveolar pCO_2 and, in turn, increase alveolar pO_2. Arterial pH is also increased during hyperventilation, leading to a higher affinity of Hb for O_2. A gradual increase in 2,3-BPG typically occurs in response to chronic hypoxia. Another important adaptive mechanism is polycythemia, an increase in erythrocyte concentration that results from erythropoietin stimulation of bone marrow cells. Within one week of acclimatization the Hb concentration can increase by as much as 20% to provide near-normal arterial O_2 content.

Neuroglobin and cytoglobin: minor mammalian hemoglobins

Two other globins have recently been identified in humans

Neuroglobin (Ngb) is expressed primarily in the central nervous system and some endocrine tissues; cytoglobin (Cygb) is ubiquitously expressed, primarily in cells of fibroblast origin. Tissue concentrations of both are <1 mmol/L. The

Ngb polypeptide has 151 amino acid residues (16,933 Da), whereas Cygb contains 190 residues (21,405 Da), with 'extensions' of 20 amino acids at both the *N*- and C-termini (see Fig. 5.1). Both human proteins share only about 25% sequence identity with Mb and Hb. Yet all key elements of the globin fold are present: the three-over-three α-helix sandwich; the proximal and distal His residues; and a hydrophobic, heme-containing pocket.

In contrast to Mb and Hb, Ngb and Cygb contain hexacoordinate hemes for both the Fe^{2+} and Fe^{3+} valency states. The distal HisE7, serving as the sixth ligand, must be displaced to permit binding of O_2. Yet the O_2 affinities of Ngb and Cygb are surprisingly high, with P_{50} values in the range 1–7.5 mmHg and 0.7–1.8 mmHg, respectively, compared to a P_{50}<27 mmHg for Hb. Binding of O_2 to the dimeric Cygb is cooperative (Hill coefficient = 1.2–1.7) but independent of pH. On the other hand, monomeric Ngb exhibits a pH-dependent O_2 affinity. The functions of these minor globins remain elusive. Ngb appears to be comparable to Mb, mediating the delivery of O_2 to retina mitochondria. Cygb is thought to function as an enzyme cofactor, supplying O_2 for the hydroxylation of Pro and Lys side chains in some proteins.

Clinical box A student with hyperventilation numbness, and dizziness

A college student with severe muscle spasms in her arms, numbness in her extremities, some dizziness, and respiratory difficulty was brought to the student health center. The patient had been vigorously exercising in an attempt to relieve the stress of forthcoming examinations when she suddenly began to experience forced, rapid breathing. Suspecting hyperventilation, a health care worker began to reassure the student and helped her recover by getting her to breathe into a paper bag. After 20 minutes the spasms ceased, feeling returned to her fingers, and the lightheadedness resolved.

Comment.

Alveolar hyperventilation is an abnormally rapid, deep, and prolonged breathing pattern that leads to respiratory alkalosis, *i.e.* a profound decrease in pCO_2 and an increase in blood pH that can be

attributed to an increased loss of CO_2 from the body. With decreased [CO_2] and [H^+], two allosteric effectors of O_2 binding and release, the affinity of Hb for O_2 increases sufficiently to reduce the efficiency of delivery of O_2 to peripheral tissues, including the central nervous system. Another characteristic of alkalosis is a decreased level of ionized calcium in plasma, a situation that contributes to muscle spasms and cramps. In general, hyperventilation may be triggered by hypoxemia, pulmonary and cardiac diseases, metabolic disorders, pharmacologic agents, and anxiety. See also Chapter 25.

Hemoglobin variants

Over 95% of the Hb found in adult humans is **HbA**, with the $\alpha_2\beta_2$ globin subunit composition. HbA_2 accounts for 2–3% of the total and has an $\alpha_2\delta_2$ polypeptide composition. **HbA_2** is elevated in β-thalassemia, a disease characterized by a deficiency in β-globin biosynthesis. Functionally, these two adult hemoglobins are indistinguishable. Not surprisingly, mutations of the gene encoding δ-globin are without clinical consequence.

Another minor Hb is **fetal Hb, HbF;** its subunits are α-globin and γ-globin. While it accounts for no more than 1% of adult Hb, HbF predominates in the fetus during the second and third trimesters of gestation and in the neonate. Gene switching on chromosome 11 causes HbF to decrease shortly after birth. The most striking functional difference between HbF and HbA is its decreased sensitivity to 2,3-BPG. Comparison of the primary structures of the β-and γ-polypeptides reveals a replacement of $His^{143\beta}$ by Ser in γ-globin (see Fig. 5.1). Consequently, two of the cationic groups that participate in the binding of the anionic allosteric effector are no longer available (see Fig. 5.8). Predictably, the interaction of 2,3-BPG with HbF is weaker, resulting in an increased affinity for O_2 (P_{50} of 19 mmHg for HbF compared to 27 mmHg for HbA) and a greater stabilization of the oxygenated R state. **The direct benefit of this structural and functional change in the HbF isoform is a more efficient transfer of O_2 from maternal HbA to fetal HbF** (see Fig. 5.4). Separation of these and other Hb variants in the clinical laboratory is performed by electrophoretic and

chromatographic analysis (Fig. 5.9).

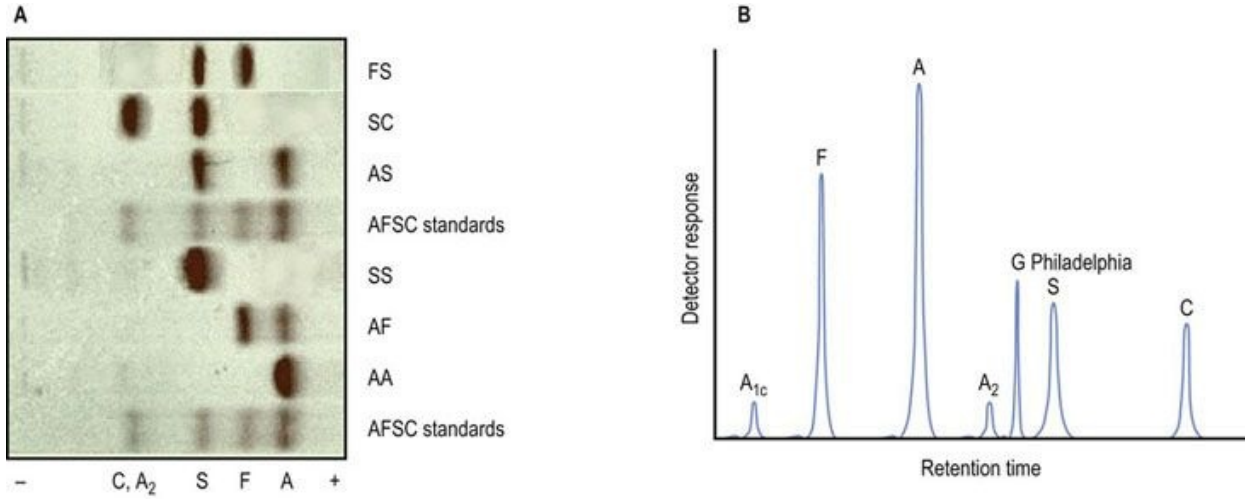


FIG. 5.9 Normal and abnormal hemoglobins can be separated by electrophoretic and chromatographic methods.
(A) This panel shows cellulose acetate electrophoresis (pH 8.4) of blood samples obtained for neonatal screening. This rapid technique will tentatively identify HbS and HbC, two common mutant hemoglobins in the African-American population. Additional tests are required for a definitive diagnosis. FS, newborn with sickle cell disease; SC, double heterozygote child with sickle cell-like disease; AS, child with sickle cell trait; AF, normal neonate. **(B)** This trace illustrates high-pressure liquid chromatography (HPLC) with a cation-exchanger solid phase, a technique capable of separating and quantifying more than 40 hemoglobins. HPLC may also be used to measure HbA_{1c}, a glycated protein that is measured clinically as an index of mean blood glucose concentration in diabetes mellitus. Also shown is the elution profile of Hb G Philadelphia ($Asn^{68(\alpha)} \rightarrow Lys$), a common but benign variant that co-migrates with HbS on electrophoresis.

Sickle cell disease, a common hemoglobinopathy

In sickle cell disease (SCD), distortion of erythrocyte structure (sickling) limits capillary blood flow

Clinically, an individual with SCD presents with intermittent episodes of hemolytic anemia, resulting from chronic lysis of red cells, and painful vasoocclusive crises. Common features also include impaired growth, increased susceptibility to infections, and multiple organ damage. In the African-American

population in the US, SCD affects 90,000–100,000 individuals, a frequency of 0.2%. Heterozygous, mostly asymptomatic, carriers, number 8% in this same population.

SCD is caused by an inherited, single point mutation in the gene encoding β-globin, leading to the expression of the Hb **variant HbS**. Indeed, HbS has been studied biochemically, biophysically, and genetically for over 50 years, making SCD the paradigm of a molecular disease. The mutation is $Glu^{6(\beta)} \rightarrow Val$: a surface-localized charged amino acid is replaced by a hydrophobic residue. Valine on the mutant β-globin subunit fits into a complementary pocket (sometimes called a 'sticky patch') formed on the β-globin subunit of a deoxygenated Hb molecule, a pocket that becomes exposed only upon the release of bound O_2 in tissue capillaries.

HbA remains a true solute at rather high concentrations, largely as a result of a polar exterior surface that is compatible and nonreactive with nearby Hb molecules. In contrast, HbS, when deoxygenated, is less soluble. It forms long, filamentous polymers that readily precipitate, distorting erythrocyte morphology to the characteristic sickle shape. In the homozygous individual with SCD (HbS/HbS), the complex process of nucleation and polymerization occurs rapidly, producing about 10% of circulating erythrocytes that are sickled. In the heterozygous individual (HbA/HbS, sickle cell trait), the kinetics of sickling are decreased by at least a factor of 1000, thereby accounting for the asymptomatic nature of this genotype. In dilute solution, HbS has interactions with O_2 (P_{50} value, Hill coefficient) that are similar to those for HbA. However, the Bohr effect on concentrated HbS is more pronounced, leading to greater release of O_2 in the capillaries and increased propensity for sickling.

Clinical box Acquired methemoglobinemia

In a rural region of the state, a 4-month-old infant was seen at the local emergency room for episodes of seizures, breathing difficulty, and vomiting. The infant's skin and mucous membranes were bluish, indicating cyanosis. Analysis of arterial blood revealed a chocolate brown color, a normal pO_2, an O_2 saturation of 60%, and a metHb (ferric-heme) level of 35%.

The tentative cause of the acute toxic methemoglobinemia was found to be well water contaminated by a nitrate/nitrite

concentration of 34 mg/L. The infant was treated successfully by intravenous administration of methylene blue (1–2 mg/kg) that serves to accelerate indirectly the enzymatic reduction of metHb to normal (ferrous) Hb by NADPH metHb reductase, which is normally a minor pathway for conversion of metHb to Hb.

Comment.

MetHb is formed when the ferrous iron of heme is oxidized to ferric iron; it is produced spontaneously at a low rate and more rapidly in the presence of certain drugs, nitrites, and aniline dyes. In genetic forms of methemoglobinemia, mutation of either the proximal or distal His to Tyr makes the heme iron more susceptible to oxidation (see Table 5.1). The extent of oxidation in Hb tetramers can range from one heme group to all four. Erythrocytes contain an NADH-cytochrome b_5 reductase, or NADH diaphorase, that is responsible for the majority of metHb reduction. Infants are particularly vulnerable to methemoglobinemia because their level of NADH-cytochrome b_5 reductase is half that of adults. Moreover, their higher level of HbF is more sensitive to oxidants compared to HbA.

Sickled erythrocytes exhibit less deformability. They no longer move freely through the microvasculature and often block blood flow, especially in the spleen and joints. Moreover, these cells lose water, become fragile, and have a considerably shorter life span, leading to **hemolysis** and **anemia**. Except during extreme physical exertion, the heterozygous individual appears normal. For reasons that remain to be elucidated, heterozygosity is associated with an increased resistance to malaria, specifically growth of the infectious agent *Plasmodium falciparum* in the erythrocyte. This observation represents an example of a selective advantage that the HbA/HbS heterozygote exhibits over either the HbA/HbA normal or the HbS/HbS homozygote and probably offers an explanation for the persistence of HbS in the gene pool.

Clinical Box Analgesic treatment

of sickle cell vasoocclusive crises

Acute vasoocclusive crises are the most common problem reported by individuals with SCD; they are also the most frequent reason for emergency room treatment and hospital admission. Episodes of vasoocclusive pain are unpredictable and are often excruciating and incapacitating. The origin of this progressive pain involves altered rheologic and hematologic properties of erythrocytes attributable to HbS polymerization and aggregation. Microvascular dysfunction is precipitated by an inflammatory response, indicated by elevation of plasma acute-phase proteins. Ultimately, impaired vasomotor responses in arterioles and adhesive interactions between sickled erythrocytes and endothelial cells in postcapillary venules restrict blood flow to tissues throughout the body.

Epidemiologic data indicate that 5% of patients with SCD can expect to experience 3–10 episodes of severe pain annually. Typically, the pain crisis resolves within 5–7 days, but a severe crisis may cause pain that persists for weeks. To provide relief to the patient, nonnarcotic, narcotic, and adjuvant analgesics are used alone or in combination.

The severity and duration of the pain dictate the most appropriate analgesic regimen. Parenterally administered opioids (morphine, hydromorphone, meperidine) are frequently used for treatment of severe pain in vasoocclusive crises. Several recent studies suggest additional options for the patient and physician: continuous intravenous infusion of a nonsteroidal antiinflammatory drug (ketorolac) and continuous epidural administration of local anesthetic (lidocaine) and opioid analgesic (fentanyl) effectively decreased pain that was unresponsive to conventional measures. In addition to analgesia, oxygen therapy and fluid management are also initiated.

Other hemoglobinopathies

More than 1000 mutations in the genes encoding the α-and β-globin polypeptides have been documented

As with most mutational events, most lead to few, if any, clinical problems. There are, however, several hundred mutations that give rise to abnormal Hb and pathologic phenotypes. Hb mutants or hemoglobinopathies are usually named after the location (hospital, city or geographical region) in which the abnormal protein was first identified. They are classified according to the type of structural change and altered function and the resulting clinical characteristics (see Tables 5.1, 5.2). While many of these mutants have predictable phenotypes, others are surprisingly pleiotropic in their impact on multiple properties of the Hb molecule. With few exceptions, Hb variants are inherited as autosomal recessive traits. Occasionally, double heterozygotes are identified, *e.g.* HbSC (Fig. 5.9).

Table 5.2

Complete blood count (CBC)

Parameter	Patient (male)	Reference value (SI units)*
White blood cell count, WBC	6.82 × 10^9/L	4.0–11.0 × 10^9/L
Red cell count, RBC	4.78 × 10^{12}/L	4.0–5.2 × 10^{12}/L (F); 4.5–5.9 × 10^{12}/L (M)
Hemoglobin, Hb	6.1 mmol/L	7.4–9.9 mmol/L (F); 8.4–10.9 mmol/L (M)
Hematocrit, HCT	33.4%	41–46% (F); 37–49% (M)
Mean corpuscular volume, MCV	71.9 fL	80–96 fL
Mean corpuscular hemoglobin, MCH	21.3 pg/cell	26–34 pg/cell
Mean corpuscular hemoglobin concentration, MCHC	296 g/L	320–360 g/L
Red cell distribution width, RDW	17.7%	11.5–14.5%
Platelet count, PLT	274 × 10^9/L	150–350 × 10^9/L
Mean platelet volume, MPV	8.6 fL	6.4–11.0 fL

*F, female; M, male; fL, 10^{-15} L; pg, 10^{-12} g. To convert mmol Hb/L to g Hb/dL, multiply by 0.01611.

Automated laboratory evaluation of blood provides invaluable information for the diagnosis and monitoring of health problems. The complete blood count, performed on a sample of whole blood, includes counts of red cells (erythrocytes), white cells (leukocytes), and platelets and quantitative indices of the red cells (MCV, MCH, MCHC, and RDW). The results describe the hematopoietic status of the bone marrow and the presence of anemia and its possible cause. Data presented are characteristic of an individual with iron deficiency anemia: low HGB, low MCV (microcytosis), and low MCH (hypochromia). See also reference values in the Appendix 1.

Clinical test box Separation of hemoglobin variants and mutants; diagnosis of hemoglobinopathies

The mobility of a protein during electrophoresis or chromatography is determined by its charge and interaction with

the matrix. Three commonly used techniques provide sufficient resolution to separate Hb variants differing in a single charge from HbA: electrophoresis, isoelectric focusing, and ion exchange chromatography. Electrophoretic and chromatographic separations of Hb are illustrated in Figure 5.9.

The volume of hemolysate required (<100 µL) makes these techniques suitable for neonatal and adult blood samples. Quantification is performed by scanning densitometry or absorption spectrometry. Indications of abnormalities in screening tests are followed by complete blood count (Table 5.2), additional protein analysis, and DNA analysis to identify specific mutations to the globin genes.

Clinical test box Complete blood count

A complete blood count (CBC) provides information on blood cell populations and their characteristics. Data are obtained from whole blood samples by automated hematology analysis. Some instruments also provide leukocyte differentials, reticulocyte count, and red cell morphology. A typical printout of the results for one individual and the reference range is shown in Table 5.2.

Summary

■ This chapter describes two important proteins that reversibly interact with O_2: myoglobin (Mb), a tissue oxygen storage protein, and hemoglobin (Hb), a blood oxygen transport protein. Both use an ancient heme-containing polypeptide domain motif to sequester O_2 and increase its solubility.

■ As a tetramer of globins, Hb is one of the best characterized examples of cooperativity in ligand interactions.

■ With its wide variety of effector molecules, Hb is also a prototype of an allosteric proteins and enzymes.

■ 2,3-Bisphosphoglycerate is an important allosteric effector of Hb, decreasing the oxygen affinity of hemoglobin; this is an important adaptation to high altitude and in pulmonary disease. Protons, through the Bohr effect, and CO_2 also promote the release of oxygen from hemoglobin in peripheral tissue. Conformational changes in both the tertiary and quaternary structures characterize the transition between deoxygenated and oxygenated states.

■ Mutations to globin genes lead to a spectrum of structural and functional variants, some of which are pathogenic, such as HbS, which causes sickle cell disease.

Active learning

1. Discuss why some genetic mutations to α-globin or β-globin result in a pathologic phenotype while the majority remain silent or benign. Describe the mutations that are the most difficult to detect.
2. Speculate on the mechanisms by which an adult with sickle cell disease would benefit from a fetal hemoglobin (HbF) level of 20%.
3. Many hemoglobin-based oxygen carriers (HBOCs) have a decreased sensitivity to pH and an increased susceptibility to oxidation. Discuss the consequences of a reduced Bohr effect to oxygen delivery to the periphery, tissue acid–base balance, and CO_2 transport to the lungs.
4. Summarize the observations of experimental animals in which the gene encoding myoglobin has been ablated (‘knocked out’).

Further reading

Allen, BW, Stamler, JS, Piantadosi, CA. Hemoglobin, nitric oxide and molecular mechanisms of hypoxic vasodilation. *Trends Mol Med*. 2009; 15:452–460.

Bauer, I, Pannen, BH. Bench-to-bedside review: Carbon monoxide – from mitochondrial poisoning to therapeutic use. *Crit Care*. 2009; 13:220.

Jahr, JS, Akha, AS, Holtby, RJ. Crosslinked, polymerized, and PEG-conjugated hemoglobin-based oxygen carriers: clinical safety and efficacy of recent and current products. *Curr Drug Discov Technol*. 2012; 9:158–165.

Kutlar, F. Diagnostic approach to hemoglobinopathies. *Hemoglobin*. 2007; 31:243–250.

Malowany, JI, Butany, J. Pathology of sickle cell disease. *Semin Diagn Pathol*. 2012; 29:49–55.

Skold, A, Cosco, DL, Klein, R. Methemoglobinemia: pathogenesis, diagnosis, and management. *South Med J*. 2011; 104:757–761.

Thein, SL. Milestones in the history of hemoglobin research. *Hemoglobin*. 2011; 35:450–462.

Vogt, M, Hoppeler, H. Is hypoxia training good for muscles and exercise performance? *Prog Cardiovasc Dis*. 2010; 52:525–533.

Winslow, RM. Oxygen: the poison is in the dose. *Transfusion*. 2012.

Websites

The Red Cell and Anemia (detailed five-part presentation by pathologist E Uthman): Blood cells and the CBC. http://web2.airmail.net/uthman/blood_cells.html.

Anemia: Pathophysiologic Consequences, Classification, and Clinical Investigation. http://web2.airmail.net/uthman/anemia/anemia.html.

Nutritional Anemias and Anemia of Chronic Disease. http://web2.airmail.net/uthman/nutritional_anemia/nutritional_anemia.html.

Hemolytic Anemias. http://web2.airmail.net/uthman/hemolytic_anemia/hemolytic_anemia.html.

Hemoglobinopathies and Thalassemias. http://web2.airmail.net/uthman/hemoglobinopathy/hemoglobinopathy.html.

Sickle Cell Information Center (comprehensive site for both patients and professionals). www.scinfo.org/.

Animations:

Interactive examination (Jmol format) of hemoglobin and heme structures, binding of oxygen, and impact of sickle cell disease mutation. Molecular Visualization Resources, University of Massachusetts. www.umass.edu/molvis/tutorials/hemoglobin.

Teaching Cases, American Society of Hematology. teachingcases.hematology.org.

CHAPTER 6

Catalytic Proteins – Enzymes

Junichi Fujii

Learning objectives

After reading this chapter you should be able to:

■ Describe the characteristics of enzymatic reactions from the viewpoint of free energy, equilibrium and kinetics.

■ Discuss the structure and composition of enzymes, including the role of cofactors, and conditions that affect enzymatic reactions.

■ Describe enzyme kinetics based on the Michaelis–Menten equation and the significance of the Michaelis constant (K_m).

■ Describe the elements of enzyme structure that explain their substrate specificity and catalytic activity.

■ Describe regulatory mechanisms affecting enzymatic reactions, including regulation by allosteric effectors and covalent modification.

■ Differentiate among the major types of enzyme inhibition from the viewpoint of enzyme kinetics.

■ Discuss the therapeutic use of enzyme inhibitors and the diagnostic utility of clinical enzyme assays.

Introduction

Almost all biological functions are supported by chemical reactions catalyzed by biological catalysts, called enzymes

Efficient metabolism is controlled by orderly, sequential, and branching metabolic pathways. Enzymes accelerate chemical reactions under physiologic conditions. However, an enzyme cannot alter the equilibrium for a reaction; it can only accelerate the reaction rate, by decreasing the activation energy of the reaction (Fig. 6.1). Regulation of enzymatic activities allows metabolism to adapt to rapidly changing conditions. **Nearly all enzymes are proteins,** although some ribonucleic acid molecules, termed ribozymes, also have catalytic activity (Chapter 32). Based on analysis of the human genome, it is estimated that about a quarter of human genes encode for enzymes that catalyze metabolic reactions.

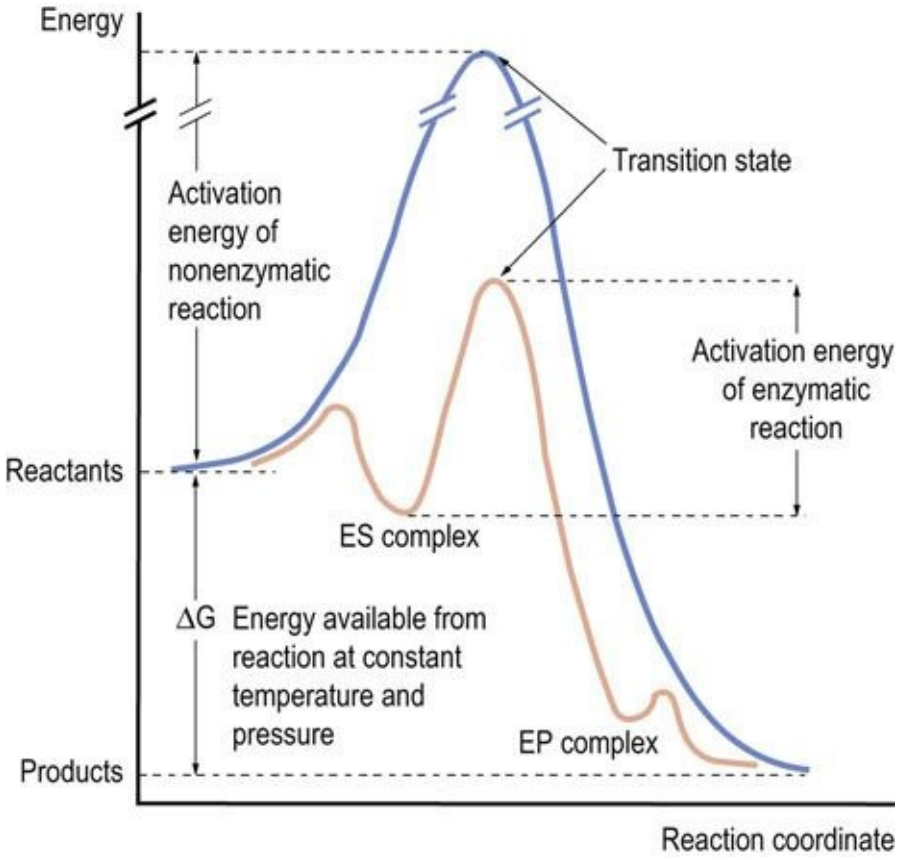


FIG. 6.1 Reaction profile for enzymatic and nonenzymatic reactions.
The basic principles of an enzyme-catalyzed reaction are the same as any chemical reaction. When a chemical reaction proceeds, the substrate must gain activation energy to reach a point called the transition state of the reaction, at which the energy level is maximum. Since the **transition state** of the enzyme-catalyzed reaction has a lower energy than that of the uncatalyzed reaction, the reaction can proceed faster. ES complex, enzyme–substrate complex; EP complex, enzyme–product complex.

Enzymatic Reactions

Factors affecting enzymatic reactions

Effect of temperature

Enzymes have an optimum temperature at which they function most efficiently

In the case of an inorganic catalyst, the reaction rate increases with the temperature of the system, and high temperature may be used to accelerate a reaction. In contrast, enzymes normally function as catalysts at constant (ambient or body) temperature. In in vitro assays, however, enzyme activity increases with temperature, but then declines at higher temperature. This happens because enzymes, like all proteins, denature at high temperature and lose activity.

Effect of pH

Every enzyme has a pH optimum because ionizable amino acids, such as histidine, glutamate, and cysteine, participate in the catalytic reactions

Cytosolic enzymes have pH optima in the pH 7–8 range. Pepsin, which is secreted by gastric cells and functions in gastric juice, has a pH optimum of 1.5–2.0; trypsin and chymotrypsin have alkaline pH optima, consistent with their digestive activity in alkaline pancreatic juice; lysosomal enzymes typically have acidic pH optima. The pH sensitivity of enzymes results from the effect of pH on the ionic charge of amino acid side chains of enzymes. Various solutes, including substrates, products, metal ions and regulatory molecules, also affect the rate of enzymatic reactions.

Definition of enzyme activity

One international unit (IU) of enzyme catalyzes conversion

of 1 µmol of substrate to product per minute

For the purposes of standardization, the activity of an enzyme is measured under defined conditions (temperature, pH, buffer, substrate and coenzyme concentration). The rate or velocity (v) of an enzymatic reaction under these conditions is defined as the rate of conversion of substrate to product per unit of time. A unit of enzyme is a measure of the amount of enzyme. The commonly used international unit (IU) is the amount of enzyme that catalyzes conversion of one micromole of substrate to product per min (1 IU = 1 µmol/min). The **katal** is an international unit for the amount of enzyme that catalyzes conversion of 1 mole of substrate into 1 mole of product per second (1 kat = 1 mol/s). Because the katal is generally a very small number, the much larger international unit is more commonly used as the standard unit of activity.

The specific activity of an enzyme is a measure of the number of IU/mg protein

The specific activity of an enzyme, a measure of activity per amount of protein, is expressed as µmol/min/mg of protein or IU/mg of protein. The specific activity of enzymes varies greatly among tissues, depending on the metabolic function of the tissue. The enzymes for cholesterol synthesis, for example, have a higher specific activity (IU/mg tissue) in liver than in muscle, consistent with the role of liver in biosynthesis of cholesterol. The specific activity of an enzyme is useful for estimating its purity – the higher the specific activity of an enzyme, the higher its purity or homogeneity.

Reaction and substrate specificity

Most enzymes are highly specific for both the type of reaction catalyzed and the nature of the substrate(s)

The reaction that the enzyme catalyzes, is determined chemically by the amino acid residues in the catalytic center of the enzyme. In general, the active site of the enzyme is composed of the substrate binding site and the catalytic site. Substrate specificity is determined by the size, structure, charges, polarity, and hydrophobicity of the substrate binding site. This is because the substrate must bind in the active site as the first step in the reaction, setting the stage for catalysis. Highly specific enzymes such as catalase and urease, which degrade H_2O_2 and urea, respectively, catalyze only one specific chemical reaction, but some enzymes have broader substrate specificity. The serine proteases are a

typical example of such a group of enzymes. These are a family of closely related enzymes, such as the pancreatic enzymes, chymotrypsin, trypsin, and elastase, which contain a reactive serine residue in the catalytic site. They catalyze the hydrolysis of peptide bonds on the carboxyl side of a limited range of amino acids in protein. Although they have similar structures and catalytic mechanisms, their substrate specificities are quite different because of structural features of the substrate binding site (Fig. 6.2).

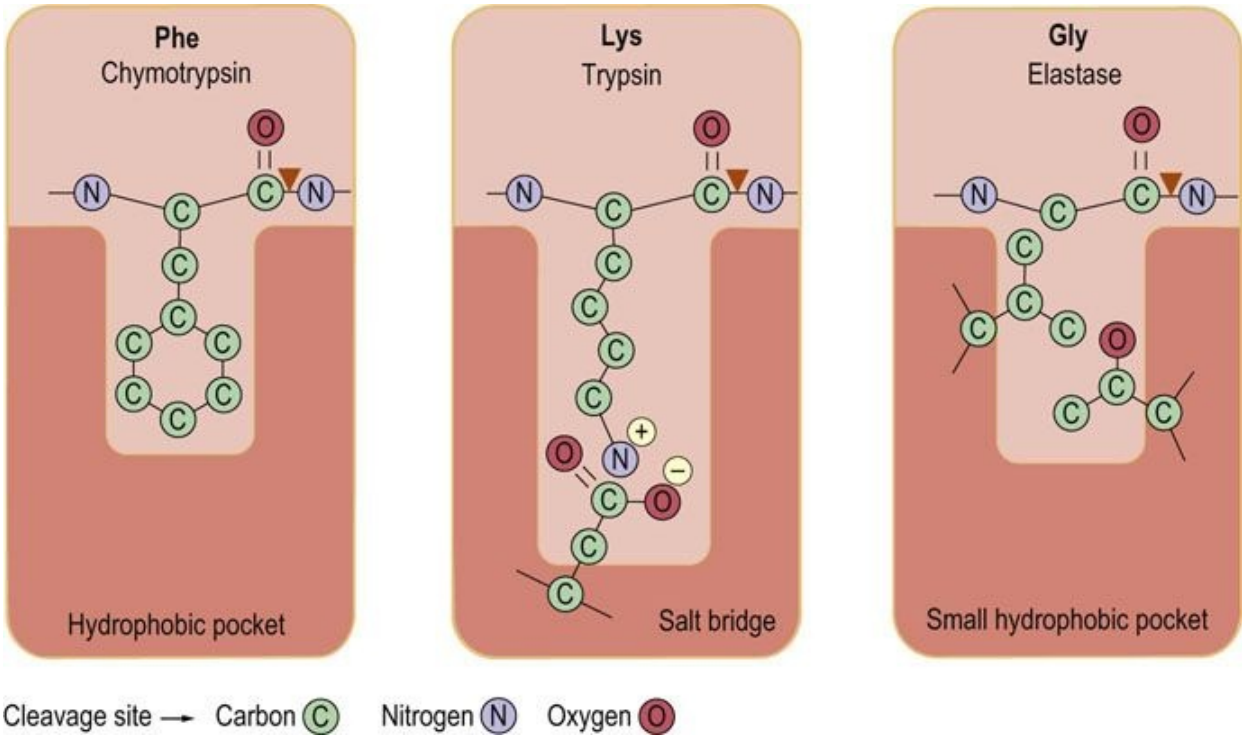


FIG. 6.2 Characteristics of the substrate binding sites in the serine proteases chymotrypsin, trypsin, and elastase.
In chymotrypsin a hydrophobic pocket binds aromatic amino acid residues such as phenylalanine (Phe). In trypsin, the negative charge of the aspartate residue in the substrate binding site promotes cleavage to the carboxyl side of positively charged lysine (Lys) and arginine (Arg) residues. In elastase, side chains of valine and threonine block the substrate binding site and permit binding of amino acids with small or no side chains, such as glycine (Gly). ▼ site of hydrolysis by enzyme.

All enzymes are assigned a four-digit enzyme classification (EC) number to organize the different enzymes that catalyze the many thousands of reactions. The first digit indicates membership of one of the six major classes of enzymes shown in Table 6.1. The next two digits indicate substrate subclasses and sub-

subclasses; the fourth digit indicates the serial number of the specific enzyme. Isozymes are enzymes that catalyze the same reaction, but differ in their primary structure and/or subunit composition. The activities of some tissue-specific enzymes and isozymes are measured in serum for diagnostic purposes (Fig. 6.3 and Table 6.2).

Table 6.1

Enzyme classification

Class	Reaction	Enzymes
1. Oxidoreductases	$A_{red} + B_{ox} \rightarrow A_{ox} + B_{red}$	Dehydrogenases, peroxidases
2. Transferases	A-B + C → A + B-C	Hexokinase, transaminases
3. Hydrolases	A-B + H_2O → A-H + B-OH	Alkaline phosphatase, trypsin
4. Lyases (synthases)	X-A-B-Y → A = B + XY	fumarase, dehydratases
5. Isomerases	A ⇌ isoA	Triose phosphate isomerase, phosphogluco-mutase
6. Ligases (synthetases)	A + B + ATP → A-B + ADP + Pi	Pyruvate carboxylase, DNA ligases

Table 6.2

Some enzymes used for clinical diagnosis

Enzyme	Tissue source(s)	Diagnostic use
AST	Heart, skeletal muscle, liver, brain	Liver disease
ALT	Liver	Liver disease, *e.g.* hepatitis
Amylase	Pancreas, salivary gland	Acute pancreatitis, biliary obstruction
CK	Skeletal muscle, heart, brain	Muscular dystrophy, myocardial infarction
GGT	Liver	Hepatitis, cirrhosis
LDH	Heart, liver, erythrocytes	Lymphoma, hepatitis
Lipase	Pancreas	Acute pancreatitis, biliary obstruction
Alkaline phosphatase	Osteoblast	Bone disease, bone tumors
Acid phosphatase (PSA)	Prostate	Prostate cancer

AST, aspartate aminotransferase, formerly known as serum glutamate oxaloacetate transaminase (SGOT); ALT, alanine aminotransferase, formerly known as serum glutamate pyruvate transaminase (SGPT) CK, creatine phosphokinase; GGT, y-glutamyl transpeptidase; LDH, lactate dehydrogenase; PSA, prostate-specific antigen (kallikrein 3).

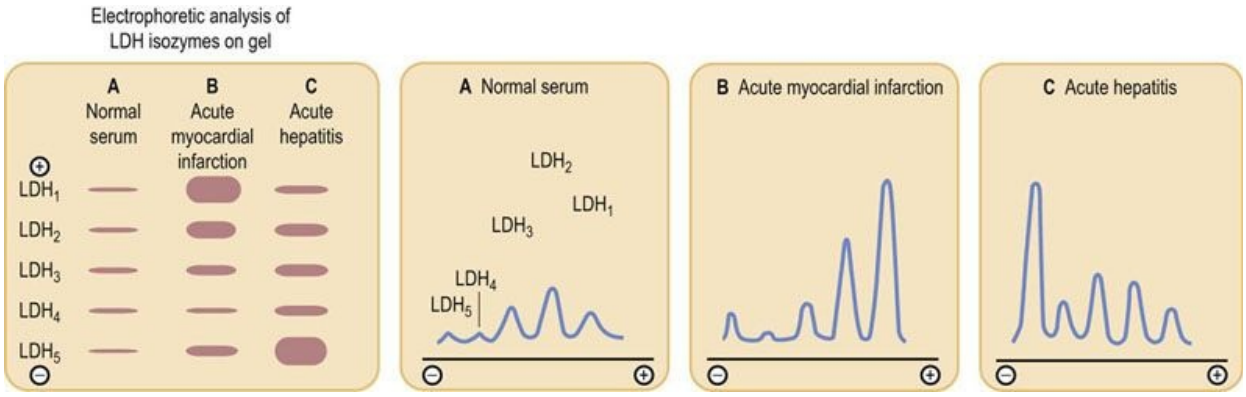


FIG. 6.3 Densitometric patterns of the lactate dehydrogenase (LDH) isozymes in serum of patients diagnosed with myocardial infarction or acute hepatitis.
Isozymes, differing slightly in charge, are separated by electrophoresis on cellulose acetate, visualized using a chromogenic substrate, and quantified by densitometry. Total serum LDH activity is also increased in these patients. Since hemolysis releases LDH from red blood cells and affects isozyme distribution and differential diagnosis, blood samples should be treated with care. The LDH measurements for the diagnosis of myocardial infarction have now been superseded by assay of plasma troponin and other biomarkers.

Roles of coenzymes

Helper molecules, referred to as coenzymes, play an essential part in many enzyme-catalyzed reactions

Enzymes with covalently or noncovalently bound coenzymes are referred to as **holoenzymes**. A holoenzyme without a coenzyme is termed an **apoenzyme**. Coenzymes are divided into two categories. Soluble coenzymes bind reversibly to the protein moiety of the enzyme. They are often modified during the enzymatic reaction, then dissociate from the enzyme and are recycled by another enzyme; oxidoreductases, discussed in Chapter 9, have coenzymes that may be oxidized by one enzyme, then reduced and recycled by another. Coenzymes, such as coenzyme A, assist in the transport of intermediates from one enzyme to another during a sequence of reactions. Most coenzymes are vitamin derivatives. Derivatives of the B vitamins, niacin and riboflavin, act as coenzymes for oxidoreductase reactions. The structure and function of coenzymes will be described in later chapters. **Prosthetic groups** are tightly bound, often covalently linked, to an enzyme and remain associated with the enzyme during the entire catalytic cycle. Some enzymes require inorganic (metal) ions,

frequently termed **cofactors**, for their activity, *e.g.* blood-clotting enzymes that require Ca^{2+}, and oxidoreductases, which use iron, copper, and manganese.

Clinical test box Tissue specificity of lactate dehydrogenase isozymes

A 56-year-old female was admitted to an intensive care unit. The patient had suffered from a slight fever for 1 week, and had some chest pain, and difficulty breathing for the past 24 h. No abnormality was found on chest X-ray or by electrocardiography. However, a blood test showed white blood cells 12,100/mm^3 (normal: 4000–9000/mm^3), red blood cells 240 × 10^4/mm^3 (normal: 380–500 × 10^4/mm^3), hemoglobin 8.6 g/dL (normal: 11.8–16.0 g/dL), lactate dehydrogenase (LDH) 1400 IU/L (normal: 200–400 IU/L). Levels of other enzymes were normal. Based on the blood tests, the LDH isozyme profile and other data, the patient was eventually diagnosed with malignant lymphoma.

Comment.

LDH is a tetrameric enzyme, and oxidoreductases composed of two different 35 kDa subunits. The heart contains mainly the H type, and skeletal muscle and the liver the M type subunit, which are encoded by different genes. Five types of tetrameric isozymes can be formed from these subunits: H_4 (LDH_1), H_3M_1 (LDH_2), H_2M_2 (LDH_3), H_1M_3 (LDH_4), and M_4 (LDH_5). Since isozyme distributions differ among tissues, it is possible to diagnose tissue damage by assaying total LDH activity and then by isozyme profiling (Fig. 6.3).

For hematological reference values, see Table 5.2 and appendix 1.

Advanced concept box Proportion of enzyme genes in whole human

genome

Original data (Venter et al., Science 291:1335, 2001) are quoted here, and so classification does not exactly match nomenclature in Table 6.2. About a quarter of genes encode enzymes. Names of enzyme groups with number and proportion (percentage in parenthesis) in a total of 26,383 human genes were as follows: transferase, 610 (2.0); synthase and synthetase, 313 (1.0); oxidoreductase, 656 (2.1); lyase, 117 (0.4); ligase, 56 (0.2); isomerase, 163 (0.5); hydrolase, 1227 (4.0); kinase, 868 (2.8); nucleic acid enzyme, 2308 (7.5).

Clinical test box Isozymes

Isozyme profiles are often performed in the clinical laboratory for diagnostic purposes (see Fig. 6.3). The definition of isozymes is often operational, i.e. based on simple and reproducible substrate-specific assay methods that sometimes do not require precise knowledge of enzyme structure.

The term isozyme is commonly used to refer to: (1) genetic variants of an enzyme; (2) genetically independent proteins with little homology; (3) heteropolymers of two or more noncovalently bound polypeptide chains; (4) unrelated enzymes that catalyze similar reactions, *e.g.* enzymes conjugated with different prosthetic groups or requiring different coenzymes or cofactors; (5) different forms of a single polypeptide chain, *e.g.* varying in carbohydrate composition, deamination of amino acids, or proteolytic modification.

Enzyme Kinetics

The Michaelis–Menten equation: a simple model of an enzymatic reaction

Enzyme reactions are multistep in nature and comprise several partial reactions

In 1913, long before the structure of proteins was known, Leonor Michaelis and Maud Leonora Menten developed a simple model for the kinetics of enzyme-catalyzed reactions (Fig. 6.4). The Michaelis–Menten model assumes that the substrate S binds to the enzyme E, forming an essential intermediate, the enzyme-substrate complex (ES), which then undergoes reaction on the enzyme surface and decomposes to E + P (product). The model assumes that E, S, and ES are all in rapid equilibrium with one another, so that a steady-state concentration of ES is rapidly achieved, and that decomposition of the ES complex to E + P is the **rate-limiting step** in catalysis. The rate of this reaction is directly dependent on the activation energy of the enzyme-catalyzed reaction (Fig. 6.1).

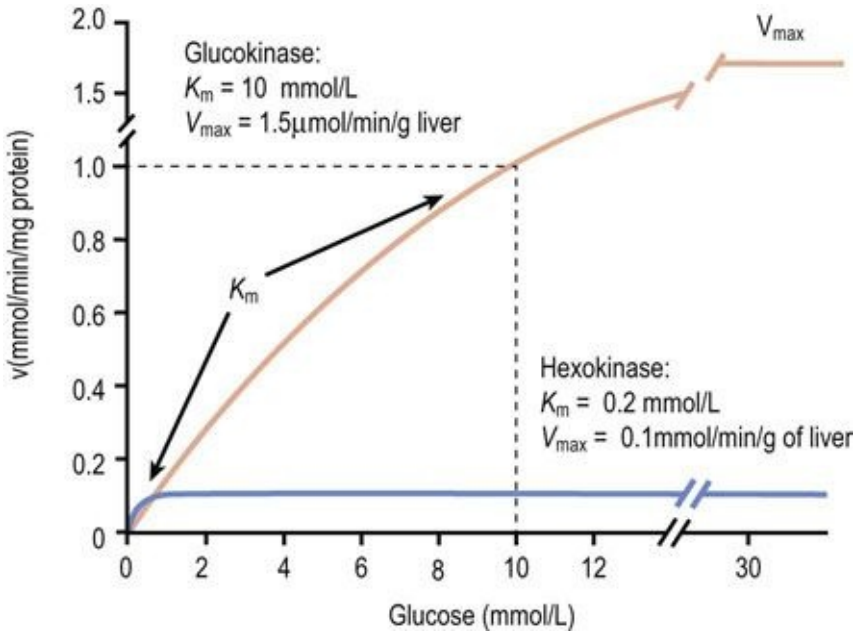


FIG. 6.4 Properties of glucokinase and hexokinase.

Glucokinase and hexokinase catalyze the same reaction, phosphorylation of glucose to glucose 6-phosphate (Glc-6-P). They exhibit different kinetic properties and have different tissue distributions and physiologic functions.

The catalytic constant, k_{cat}, also known as the turnover number, is a rate constant which describes how quickly an enzyme can catalyze a reaction. **The k_{cat} is defined as the number of substrate molecules converted to product per *enzyme molecule* per unit time.** The proportion of ES, in relation to the total number of enzyme molecules $[E]_t$, *i.e.* the ratio $[ES]/[E]_t$, limits the velocity of an enzyme (v) so that:

$$v = k_{cat}[ES]$$

Since E, S, and ES are all in chemical equilibrium, the enzyme achieves maximal velocity, V_{max}, at very high (saturating) substrate concentrations [S], when $[ES] \approx [E]_t$ (total enzyme). Thus:

For the dissociation of the ES complex, the law of mass action yields:

$$K_d = \frac{[E][S]}{[ES]}$$

Given that:

$$[E]_t = [E] + [ES]$$

it can be shown that:

$$\frac{[ES]}{[E]_t} = \frac{[S]}{K_m + [S]}$$

$$\text{where } K_m = K_d$$

Consequently, the velocity of the enzymatic reaction, *v*, is given by:

$$v = \frac{k_{cat}[E]_t[S]}{k_m + [S]}$$

Since k_{cat} $[E]_t$ corresponds to the maximum velocity, V_{max}, that is attained at high (saturating) substrate concentrations, we obtain the Michaelis–Menten equation:

$$v = V_{max} \cdot \frac{[S]}{K_m + [S]}$$

Analysis of the above equations indicates that the Michaelis constant, K_m, is expressed in units of concentration and corresponds to the substrate concentration at which v is 50% of the maximum velocity, i.e. [ES] = ½ $[E]_t$ and v = V_{max}/2 (Fig. 6.4)

K_m is a useful constant for estimating the affinity of an enzyme for its substrate. Enzymes with a high K_m require high substrate concentration for efficient activity, while those with low K_m operate efficiently on trace levels of substrate.The Michaelis–Menten model is based on the assumptions that:

- E, S, and ES are in rapid equilibrium.
- There are no forms of the enzyme present other than E and ES.
- The conversion of ES into E + P is a rate-limiting, irreversible step. While all enzyme-catalyzed reactions are theoretically reversible, initial velocities are normally measured, *i.e.* when product concentration, and therefore the rate of the reverse reaction, is negligible.

Importantly, similar types of kinetic models have been developed for describing the kinetics of multisubstrate, multiproduct enzymes.

Advanced concept box glucokinase and hexokinase

Hexokinase catalyzes the first step in glucose metabolism in all

cells, namely the phosphorylation reaction of glucose by adenosine triphosphate (ATP) to form glucose 6-phosphate (Glc-6-P):

$$\text{glucose} + \text{ATP} = \text{glucose 6-phosphate} + \text{ADP}$$

This enzyme has a low K_m for glucose (0.2 mmol/L) and is inhibited allosterically by its product, Glc-6-P. Since normal glucose levels in blood are about 5 mmol/L and intracellular levels are 0.2–2 mmol/L, hexokinase efficiently catalyzes this reaction (50–90% of V_{max}) under normal conditions, e.g. in muscle.

Hepatocytes, which store glucose as glycogen, and pancreatic β-cells, which regulate glucose consumption in tissues and its storage in liver by secreting insulin, contain an isozyme called glucokinase.

Glucokinase catalyzes the same reaction as hexokinase, but has a higher K_m for glucose (10 mmol/L) and is not inhibited by the product, Glc-6-P. Since glucokinase has a much higher K_m than hexokinase, glucokinase phosphorylates glucose with increasing efficiency as blood glucose levels increase following a meal (see Fig. 6.4). One of the physiologic roles of glucokinase in the liver is to provide Glc-6-P for the synthesis of glycogen, a storage form of glucose. In the pancreatic β-cell, glucokinase functions as the glucose sensor, determining the threshold for insulin secretion. Mice lacking glucokinase in the pancreatic β-cell die within 3 days of birth of profound hyperglycemia, because of failure to secrete insulin.

Use of the Lineweaver–Burk and Eadie–Hofstee plots

Alternative graphical analyses permit more accurate determination of the Km and Vmax of an enzyme

In a plot of reaction rate versus substrate concentration, the rate of the reaction approaches the maximum velocity (V_{max}) asymptotically (Fig. 6.5A), so that it is difficult to obtain accurate values for V_{max} and, as a result, K_m (substrate concentration required for half-maximal activity), by simple extrapolation. To solve this problem, several linear transformations of the Michaelis–Menten equation have been developed.

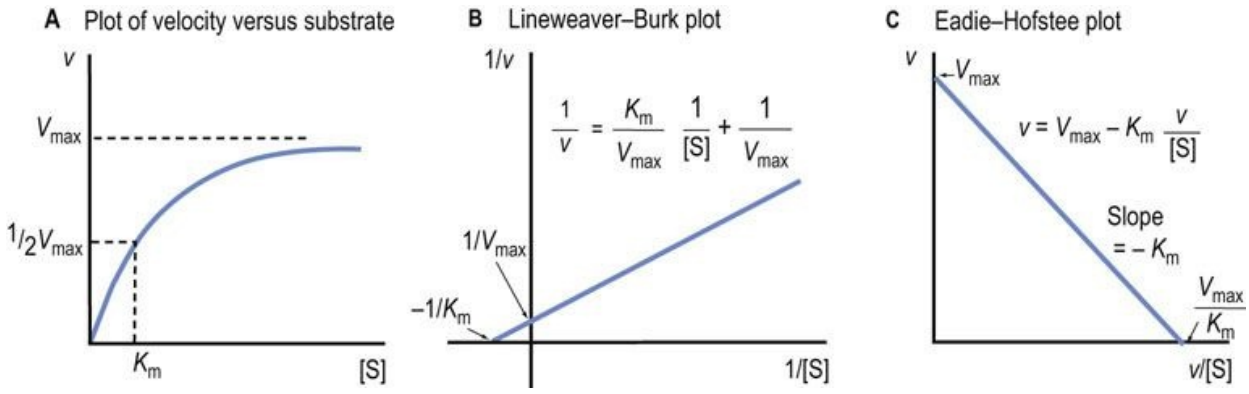


FIG. 6.5 Enzyme kinetics plot.
Kinetic representations of the properties of enzymes. (A) Michaelis–Menten plot of velocity (*v*) versus substrate concentration ([S]). (B) Lineweaver–Burk plot. (C) Eadie–Hofstee plot.

Lineweaver–Burk plot

The Lineweaver–Burk, or double reciprocal, plot is obtained by taking the reciprocal of the Michaelis–Menten equation (Fig. 6.5B). By rearranging the equation, we obtain:

$$\frac{1}{v} = \frac{1}{V_{max}} + \frac{k_m}{V_{max}} \times \frac{1}{[S]}$$

This equation yields a straight line (y = mx + b), with $y = 1/v$, $x = 1/[S]$, m = slope, b = y intercept. Therefore, a graph of $1/v$ versus $1/[S]$ (Fig. 6.5B) has a slope of K_m/V_{max}, a $1/v$ intercept of $1/V_{max}$ and a $1/[S]$ intercept of $-1/K_m$. Although the Lineweaver–Burk plot is widely used for kinetic analysis of enzyme reactions, because reciprocals of the data are calculated, a small experimental error, especially at low substrate concentration, can result in a large error in the graphically determined values of K_m and V_{max}. An additional disadvantage is that important data obtained at high substrate concentrations are concentrated into a narrow region near the $1/v$ axis.

Eadie–Hofstee plot

A second, widely used linear form of the Michaelis–Menten equation is the Eadie–Hofstee plot (Fig. 6.5C), described by the equation:

$$v = V_{max} - K_m \times \frac{v}{[S]}$$

In this case, a plot of v versus $v/[S]$ has a y axis (v-intercept) of V_{max}, an x axis ($v/[S]$) intercept of V_{max}/K_m, and a slope of $-K_m$. The Eadie–Hofstee plot does not compress the data at high substrate concentrations.

Clinical test box Measurement of enzyme activity in clinical samples

In clinical laboratories, enzyme activity is measured in the presence of saturating substrate(s) and coenzyme concentrations. Initial rates are recorded to minimize errors resulting from the reverse reaction. Under these conditions $v \approx V_{max}$, and enzyme activity is directly proportional to enzyme concentration. The amount of enzyme (enzyme activity) is commonly expressed in IU/mL of plasma, serum, cerebrospinal fluid, rather than per mg protein. For interlaboratory comparisons, the conditions for the enzyme assay must be standardized, e.g. by specifying the

substrate and coenzyme concentrations used, the buffer, buffer concentration, ionic species and ionic strength, pH and temperature.

Most clinical samples are collected under fasting conditions; this assures consistency in measurement of analytes whose concentration may vary diurnally or others, such as glucose or lipids, which vary in response to food intake. Lipemic samples are cloudy and may yield unreliable data by spectrophotometric or fluorometric methods. To avoid such problems, clinical samples must be delipidated, commonly by extraction with organic solvent.

Mechanism of Enzyme Action

Enzymatic reactions involve functional groups on enzymes, coenzymes, substrates and products

Enzymes vary significantly in their mechanism of action. In some cases catalysis is carried out on the substrate, noncovalently, reversibly bound to the enzyme. In other cases, a covalent intermediate is formed on, and then released from, the enzyme and, in others, all the action takes place on a coenzyme which forms a covalent bond with substrate. The mechanisms of action of many enzymes are discussed in later chapters in this text.

The serine proteases, introduced in Figure 6.2, are representative of enzymes that form a covalent intermediate with their substrates. These enzymes cleave peptide bonds in proteins and, as in all enzymatic reactions, functional groups on amino acid side chains participate in the enzyme-catalyzed reaction. In the serine protease family, an active site serine residue catalyzes cleavage of the peptide bond. The functional group on serine, a primary alcohol, is not among the more reactive functional groups in organic chemistry. To enhance its activity in serine proteases, this serine residue is part of a ‘catalytic triad’, in the case of chymotrypsin: Asp^{102}, His^{57} and Ser^{195} (Fig. 6.6). Concerted hydrogen bonding interactions between these amino acids increase the nucleophilicity of the serine residue, so that it can attack the carbonyl carbon atom of the peptide bond in the substrate. Chymotrypsin is specific for cleavage on the carboxyl side of peptide bonds containing aromatic amino acids, such as phenylalanine. The mechanism of the enzymatic reaction is outlined in Figure 6.7, showing the formation and cleavage of an enzyme-bound intermediate.

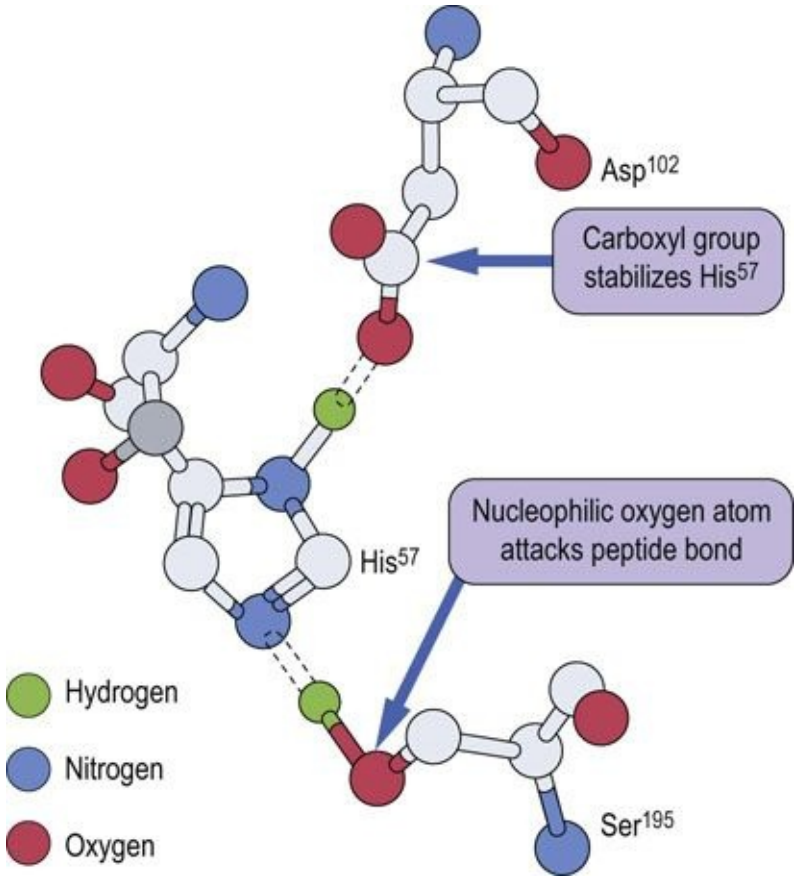


FIG. 6.6 A schematic model of a catalytic triad of serine protease.

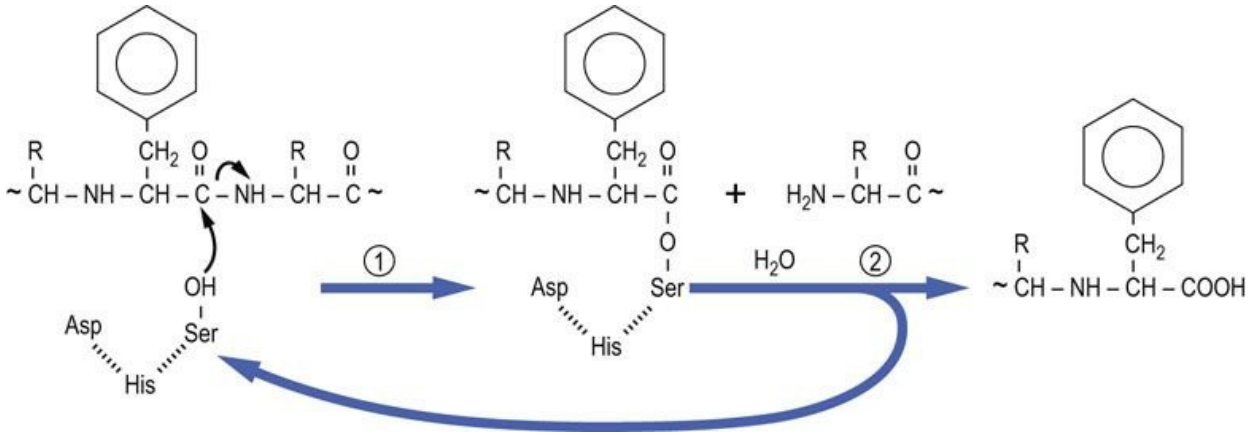


FIG. 6.7 Mechanism of action of chymotrypsin.
The active site serine residue attacks the carbonyl group of the peptide bond on the carboxyl side of a phenylalanine residue. The carboxy-terminal peptide is released and the amino-terminal peptide remains an enzyme-bound intermediate – the amino-terminal

peptide linked covalently through its carboxy-terminal phenylalanine esterified to the active site serine residue. The ester bond is hydrolyzed in the second step of the reaction to release the amino-terminal peptide and regenerate active enzyme.

Trypsin and elastase, two other digestive enzymes with different amino acid specificities (Fig. 6.2), are similar to chymotrypsin in many respects. About 40% of the amino acid sequences of these three enzymes are identical, and their three-dimensional structures are very similar. All three enzymes contain the aspartate-histidine-serine catalytic triad, and are inactivated by reaction of fluorophosphates with the active serine residue. The nerve gas, **diisopropylfluorophosphate**, forms a sterically hindered, very slowly hydrolyzed serine-diisopropylphospate ester and inhibits serine proteases.

Enzyme Inhibition

Among numerous substances affecting metabolic processes, enzyme inhibitors are particularly important. Many drugs, either naturally occurring or synthetic, act as enzyme inhibitors. Metabolites of these compounds may also inhibit enzyme activity. Most enzyme inhibitors act reversibly, but there are also irreversible inhibitors that permanently modify the target enzyme. Using the Lineweaver–Burk plots, it is possible to distinguish three forms of reversible inhibition: competitive, uncompetitive, and noncompetitive inhibition.

Competitive inhibitors cause an apparent increase in K_m, without changing V_{max}

An enzyme can be inhibited competitively by substances that are similar in chemical structure to the substrate. These compounds bind in the active site and compete with substrate for the active site of the enzyme; they cause an apparent increase in K_m, but no change in V_{max} (Fig. 6.8). The inhibition is not the result of an effect on enzyme activity, but on substrate access to the active site. The reaction scheme for competitive inhibition is:

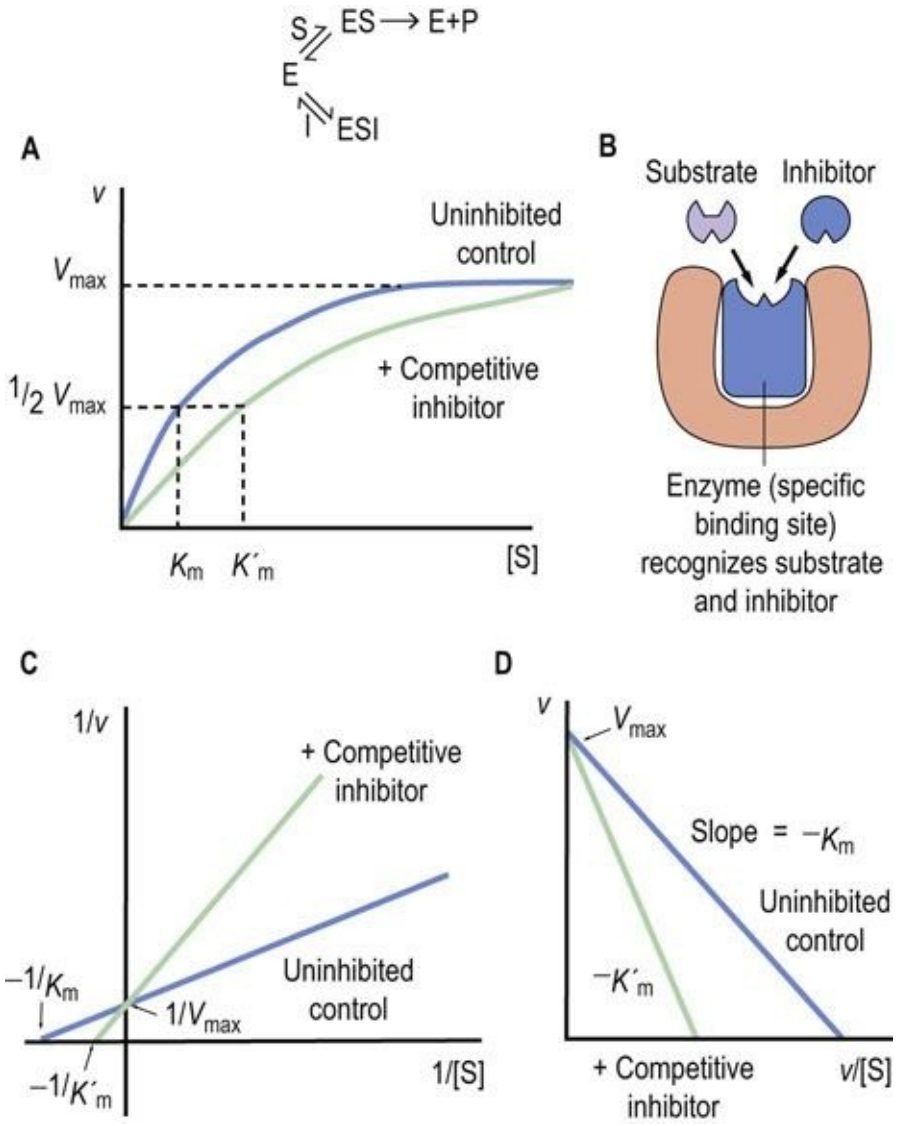


FIG. 6.8 Competitive enzyme inhibition.
(A) Plot of velocity versus substrate concentration. (B) Mechanism of competitive inhibition. (C) Lineweaver–Burk plot in the presence of a competitive inhibitor. (D) Eadie–Hofstee plot in the presence of a competitive inhibitor. K'_m is the apparent K_m in presence of inhibitor.

$$\begin{matrix} E + S \rightleftharpoons ES \rightarrow E + P \\ + \\ I \rightleftharpoons EI \end{matrix}$$

The inhibition constant (K_i) is the dissociation constant of the enzyme–

inhibitor complex (EI) and the lower the K_i, the more efficient the inhibition of enzyme activity. Regardless of the K_i, however, **the rate of the enzyme-catalyzed reaction in the presence of a competitive inhibitor can be increased by increasing the substrate concentration,** since substrate, at higher concentration, competes more effectively with the inhibitor.

Clinical box Treatment with an inhibitor of angiotensin-converting enzyme (ACE)

A 50-year-old man was admitted to hospital suffering from general fatigue, a stiff shoulder, and headache. The patient was 1.8 m tall and weighed 84 kg. His blood pressure was 196/98 mmHg (normal below 140/90 mmHg; optimal below 120/80 mmHg) and his pulse was 74. He was diagnosed as hypertensive. The patient was given captopril, an angiotensin-converting enzyme (ACE) inhibitor. After 5 days' treatment, his blood pressure returned to near-normal levels.

Comment.

Renin in the kidney converts angiotensinogen into angiotensin I, which is then proteolytically cleaved to angiotensin II by ACE. Angiotensin II increases renal fluid and electrolyte retention, contributing to hypertension. Inhibition of ACE activity is therefore an important target for hypertension treatment. Captopril inhibits ACE competitively, decreasing blood pressure. (See also Chapter 23.)

Clinical box Methanol poisoning can be treated by ethanol administration

A 46-year-old male presented to the emergency room 7 h after

consuming a large quantity of bootleg alcohol. He could not see clearly and complained of abdominal and back pain. Laboratory results indicated severe metabolic acidosis, a serum osmolality of 465 mmol/kg (reference range 285–295 mmol/kg), and serum methanol level of 4.93 g/L (156 mmol/L!). By aggressive treatment, including an ethanol drip, bicarbonate, and hemodialysis, he survived and regained his eyesight.

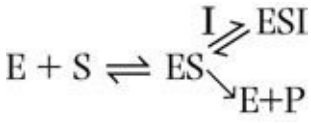
Comment.

Methanol poisoning is uncommon but extremely hazardous. Ethylene glycol poisoning is more common and exhibits similar clinical characteristics. The most important initial symptom of methanol poisoning is visual disturbance. Laboratory evidence of methanol poisoning includes severe metabolic acidosis and increased plasma solute (methanol) concentration. Methanol is slowly metabolized to formaldehyde, which is then rapidly metabolized to formate by alcohol dehydrogenase. Formate accumulates during methanol intoxication and is responsible for the metabolic acidosis in the early stage of intoxication. In later stages lactate may also accumulate as a result of formate inhibition of respiration. Ethanol is metabolized by alcohol dehydrogenase, which binds ethanol with much higher affinity than either methanol or ethylene glycol. Ethanol is therefore a useful agent to inhibit competitively the metabolism of methanol and ethylene glycol to toxic metabolites. The unmetabolized methanol and ethylene glycol are gradually excreted in urine. Early treatment with ethanol, together with bicarbonate to combat acidosis and hemodialysis to remove methanol and its toxic metabolites, yields a good prognosis.

Noncompetitive inhibitors cause an apparent decrease in V_{max}

An uncompetitive inhibitor binds only to the enzyme–substrate complex and not to the free enzyme. The equation below shows the reaction scheme for

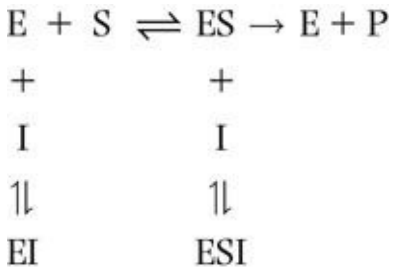
uncompetitive inhibition. In this case, the K_i is the dissociation constant for the enzyme–substrate–inhibitor complex (ESI).



The inhibitor causes a decrease in V_{max} because a fraction of the enzyme–substrate complex is diverted by the inhibitor to the inactive ESI complex. Binding of the inhibitor and the increase in the ESI complex may also affect the dissociation of substrate, causing an apparent decrease in K_m, *i.e.* an apparent increase in substrate affinity.

Noncompetitive inhibitors may bind to sites outside the active site and alter both the K_m and V_{max} of the enzyme

An uncompetitive inhibitor can bind either to the free enzyme or to the enzyme–substrate complex, *e.g.* at an allosteric site. Uncompetitive inhibitors exhibit more complex effects and may alter both the K_m and V_{max} of an enzymatic reaction. The equation below shows the reaction scheme observed for noncompetitive inhibition.



Many drugs and poisons irreversibly inhibit enzymes

Prostaglandins are key inflammatory mediators. Their synthesis is initiated by cyclooxygenase-mediated oxidation and cyclization of arachidonate under inflammatory conditions (Chapter 40). Compounds that suppress

cyclooxygenase have anti-inflammatory activity. **Aspirin** (acetylsalicylic acid) inhibits cyclooxygenase activity by acetylating Ser^{530}, which blocks access of arachidonate to the active site of the enzyme. Other **nonsteroidal anti-inflammatory drugs (NSAIDs)**, such as indomethacin, inhibit cyclooxygenase activity by reversibly blocking the arachidonate binding site.

Disulfiram (**Antabuse®**) is a drug used for the treatment of alcoholism. Alcohol is metabolized in two steps to acetic acid. The first enzyme, alcohol dehydrogenase, yields acetaldehyde, which is then converted into acetic acid by aldehyde dehydrogenase. The latter enzyme has an active site cysteine residue that is irreversibly modified by disulfiram, resulting in accumulation of alcohol and acetaldehyde in the blood. People who take disulfiram become sick because of the accumulation of acetaldehyde in blood and tissues, leading to alcohol avoidance.

Alkylating agents, such as iodoacetamide (ICH_2CONH_2), irreversibly inhibit the catalytic activity of some enzymes by modifying essential cysteine residues. **Heavy metals,** such as mercury and lead salts, also inhibit enzymes with active site sulfhydryl residues. The mercury adducts are often reversible by thiol compounds. Eggs or egg white are sometimes administered as an antidote for accidental ingestion of heavy metals The egg white protein, ovalbumin, is rich in sulfhydryl groups; it traps the free metal ions and prevents their absorption from the gastrointestinal tract.

In many cases, irreversible inhibitors are used to identify active-site residues involved in enzyme catalysis, and to gain insight into the mechanism of enzyme action. By sequencing or mass spectrometric analysis of the modified peptide, it is possible to identify the specific amino acid residue modified by the inhibitor and involved in catalysis.

Advanced concept box Enzyme inhibition: transition-state inhibition and suicide substrate

Enzymes catalyze reactions by inducing the transition state of the reaction. It should therefore be possible to construct molecules that bind very tightly to the enzyme by mimicking the transition state of the substrate. Transition states themselves cannot be isolated, because they are not a stable arrangement of atoms, and some bonds are only partially formed or broken. But for some enzymes,

analogues can be synthesized that are stable, but still have some of the structural features of the transition state.

Penicillin (Fig. 6.9) is a good example of a transition state analogue. It inhibits the transpeptidase that crosslinks bacterial cell-wall peptidoglycan strands, the last step in cell-wall synthesis in bacteria. It has a strained 4-membered lactam ring that mimics the transition state of the normal substrate. When penicillin binds to the active site of the enzyme, its lactam ring opens, forming a covalent bond with a serine residue at the active site. Penicillin is a potent irreversible inhibitor of bacterial cell-wall synthesis, making the bacterium osmotically fragile and unable to survive in the body.

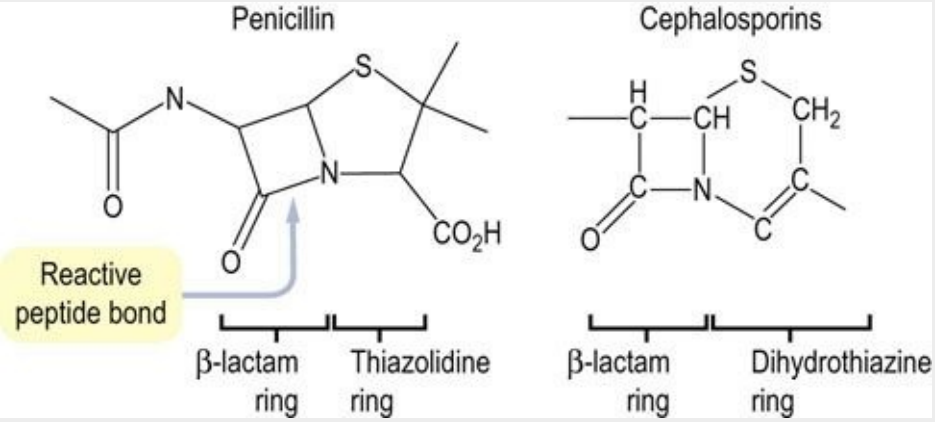


FIG. 6.9 Structure of penicillin showing the reactive peptide bond in the β-lactam ring and core structure of cephalosporins.
Penicillins contain a β-lactam ring fused to a thiazolidine ring. Cephalosporins are another class of compounds containing the β-lactam ring fused to a six-membered dihydrothiazine ring. Because of their effectiveness and lack of toxicity, β-lactam compounds are widely used antibiotics. Bacteria with β-lactamase, which breaks the β-lactam ring, are resistant to these antibiotics.

Regulation of Enzyme Activity

In multistep metabolic pathways, the slowest step limits the overall rate of the reaction

It is most efficient to regulate a metabolic pathway by controlling key enzymes that are involved in the 'rate-limiting' step. Generally, five independent mechanisms are involved in the regulation of enzyme activity.

■ The expression of the enzyme protein from the corresponding gene changes in response to the cell's changing environment or metabolic demands.

■ Enzymes may be irreversibly activated or inactivated by proteolytic enzymes.

■ Enzymes may be reversibly activated or inactivated by covalent modification, such as phosphorylation.

■ Allosteric regulation modulates the activity of key enzymes through reversible binding of small molecules at sites distinct from the active site in a process that is relatively rapid and, hence, the first response of cells to changing conditions.

■ The degradation of enzymes by intracellular proteases in the lysosome or by proteasomes in the cytosol also determines the lifetimes of the enzymes and consequently enzyme activity over a much longer period of time.

Despite extensive studies it is still impossible to deduce precisely the activity of specific enzymes in the body; many factors including substrate concentrations, pH and allosteric effectors change with time under physiologic conditions. Advances in technology now permit the measurement of thousands of metabolites quantitatively and simultaneously in tissues – a developing field known as metabolomics (Chapter 36). Using the tools of basic enzymology and metabolomics, it may eventually be possible to measure the endogenous flux through complex metabolic pathways in vivo.

Proteolytic activation of digestive enzymes

Some enzymes are stored in subcellular organelles or compartments in an inactive precursor form

Several digestive enzymes are stored as inactive zymogens or proenzymes in secretory vesicles in the pancreas. The zymogens are secreted in pancreatic juice following a meal and are activated in the gastrointestinal tract. Trypsinogen is converted into trypsin by the action of intestinal enteropeptidase.

Enteropeptidase, located on the inner surface of the duodenum, hydrolyzes an *N*-terminal peptide from the inactive trypsinogen. Rearrangement of the tertiary structure yields the proteolytically active form of trypsin. The active trypsin then digests other zymogens, such as procarboxypeptidase, proelastase and chymotrypsinogen, as well as other trypsinogen molecules (Chapter 10). Similar proteolytic cascades are observed in blood clotting and fibrinolysis (dissolution of clots) (Chapter 7).

Allosteric regulation of rate-limiting enzymes in metabolic pathways

Allosteric enzymes display sigmoidal, rather than hyperbolic, plots of reaction rate vs. substrate concentration

The substrate saturation curve for an 'isosteric' (single shape) enzyme is hyperbolic (see Fig. 6.5A). In contrast, allosteric enzymes show sigmoidal plots of reaction velocity versus substrate concentration [S] (Fig. 6.10). An allosteric (other site) effector molecule binds to the enzyme at a site that is distinct and physically separate from the substrate binding site, and affects substrate binding (K_m) and/or k_{cat}. In some cases, the substrate may exert allosteric effects; this is referred to as a **homotropic** effect. If the allosteric effector is different from the substrate, it is referred to as a **heterotropic** effect. Homotropic effects are observed when the reaction of one substrate molecule with a multimeric enzyme affects the binding of a second substrate molecule at a different active site on the enzyme. The interaction between subunits makes the binding of substrate cooperative and results in a sigmoidal curve in the plot of v versus [S]. This effect is essentially identical with that described for the binding of O_2 to hemoglobin (Chapter 5), except that in the case of enzymes, substrate binding leads to an enzyme-catalyzed reaction.

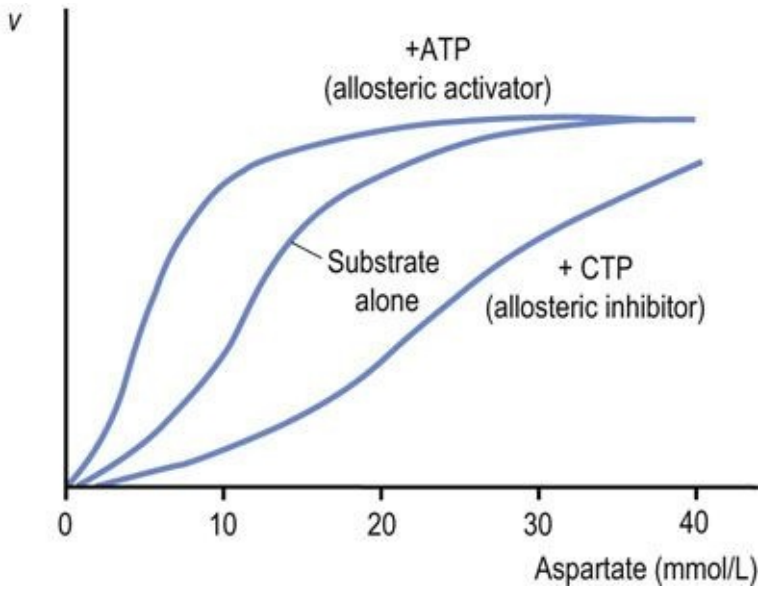


FIG. 6.10 Allosteric regulation of aspartate transcarbamoylase (ATCase). Plot of velocity (*v*) versus substrate concentration in the presence of an allosteric activator or allosteric inhibitor ATCase is an example of an allosteric enzyme. Aspartate (substrate) homotropically regulates ATCase activity, providing sigmoidal kinetics. CTP, an end product, heterotropically inhibits, but ATP, a precursor, heterotropically activates ATCase.

Clinical box Hemophilia is caused by a defect in zymogen activation

A child was admitted to hospital with muscle bleeding affecting the femoral nerve. Laboratory findings indicated a blood-clotting disorder, hemophilia A, resulting from deficiency of factor VIII. Factor VIII was administered to the patient to restore blood-clotting activity.

Comment.

Formation of a blood clot results from a cascade of zymogen-activation reactions. Over a dozen different proteins, known as blood-clotting factors, are involved. In the final step, the blood clot is formed by conversion of a soluble protein, fibrinogen (factor I), into an insoluble, fibrous product, fibrin, which forms the matrix of the clot. This last step is catalyzed by the serine protease, thrombin (factor IIa). Hemophilia is a disorder of blood clotting caused by a defect in one of the sequence of clotting factors. Hemophilia A, the

major (85%) form of hemophilia, is caused by a defect of clotting Factor VIII (see Chapter 7).

Advanced concept box Nucleoside analogues as ANTIVIRAL agents

Nucleoside analogues such as acyclovir and ganciclovir have been used for treatment of herpes simplex virus (HSV), varicella-zoster (VZV), and cytomegalovirus (CMV). They are pro-drugs that are activated by phosphorylation and terminate viral DNA synthesis by inhibiting the viral DNA polymerase reaction. The thymidine kinase (TK), more properly a nucleoside kinase, of the viruses phosphorylate these compounds to their monophosphate form. Cellular kinases next add phosphates to form the active triphosphate compounds, which are competitive inhibitors of the viral DNA polymerase during DNA replication (Chapter 32).

While viral TK has low substrate specificity and efficiently phosphorylates nucleoside analogues, cellular nucleoside kinases have high substrate specificity and barely phosphorylate the nucleoside analogues. Thus, virus-infected cells are prone to be arrested at specific cell cycle stage, G_2-M checkpoint (Chapter 42), but uninfected cells are resistant to the nucleoside analogues.

Positive and negative cooperativity

Positive cooperativity indicates that the reaction of a substrate with one active site makes it easier for another substrate to react at another active site. **Negative cooperativity** means that the reaction of a substrate with one active site makes it more difficult for a substrate to react at the other active site. Since the affinity of the enzyme changes with substrate concentration, it cannot be described by simple Michaelis–Menten kinetics. Instead, it is characterized by the substrate concentration giving a half-maximal rate, $[S]_{0.5}$, and the **Hill coefficient** (H)

(Chapter 5). The H-values are larger than 1 for enzymes with positive cooperativity and less than 1 for those with negative cooperativity. For most allosteric enzymes, intracellular substrate concentrations are poised near the $[S]_{0.5}$, so that the enzyme's activity responds to slight changes in substrate concentration.

The model most often invoked to rationalize allosteric behavior was established by Monod, Wyman and Changeaux, the so-called **concerted model** (Fig. 6.11). As with O_2 binding to Hb, in the absence of substrate, the enzyme has a low affinity for substrate and is in the T state (tense state). The other conformation of the enzyme is the R state (relaxed state). Binding of allosteric effector molecules shifts the fraction of enzyme from one state to the other. In this model, all the active sites in the R state are the same and all have higher substrate affinity than in the T state. Because the transition between the T and R states occurs at the same time for all subunits, this is called the concerted (two-state) model. An alternative model, the so-called **sequential (multistate) model**, was proposed by Koshland, Nèmethy and Filmer. It postulates that each subunit changes independently to a different conformation and that different subunits may have different affinities for substrate. It is now recognized that both models are applicable to different enzymes.

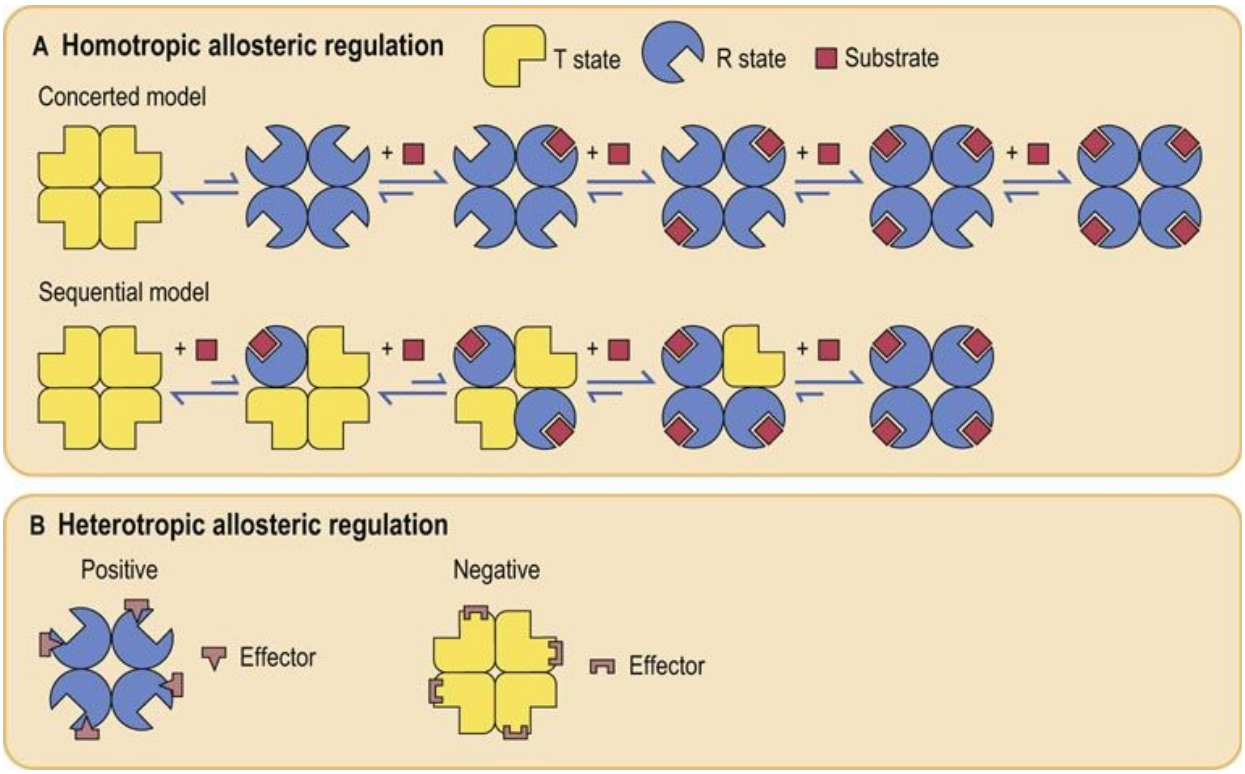


FIG. 6.11 Schematic representation of allosteric regulation.
(A) In homotropic regulation, the substrate acts as an allosteric effector. Two models are presented. In the concerted model, all of the subunits convert from the T (tense; low affinity for substrate) – into the R (relaxed; high affinity for substrate) – state at the same time; in the sequential model, they change one by one, with each substrate binding reaction. **(B)** In heterotropic regulation, the effector is distinct from the substrate, and binds at a structurally different site on the enzyme. Positive and negative effectors stabilize the enzyme in R and T state, respectively.

Clinical box Insecticide poisoning

A 55-year-old man was spraying an insecticide containing organic fluorophosphates in a rice field. He suddenly developed a frontal headache, eye pain, and tightness in his chest, typical signs of over-exposure to toxic organic fluorophosphates. He was taken to hospital and treated with an intravenous injection of 2 mg of atropine sulfate, and gradually recovered.

Comment.

Organic fluorophosphates form covalent phosphoryl-enzyme complexes with both serine proteases and esterases, such as acetylcholinesterase. irreversibly inhibiting the enzymes.

Acetylcholinesterase terminates the action of acetylcholine during neuromuscular activity (Chapter 41) by hydrolyzing the acetylcholine to acetate and choline. Inhibition of this enzyme prolongs the action of acetylcholine, leading to constant neuromuscular stimulation. Atropine competitively blocks acetylcholine binding and muscle stimulation at the neuromuscular junction.

Enzymatic Measurement of Blood Glucose

The glucose oxidase/peroxidase assay

In clinical laboratories, most compounds are measured by automated enzymatic methods

The most common assay procedure uses a mixture of glucose oxidase and peroxidase (Fig. 6.12). Glucose oxidase is highly specific for glucose, but oxidizes only the β-anomer of the sugar, which represents *64% of glucose in solution. The assay mixture is therefore supplemented with mutarotase, which rapidly catalyzes the interconversion of the anomers, enhancing assay sensitivity by* 50%. The H_2O_2 produced in the oxidase reaction is then used in a peroxidase reaction to oxidize a chromogen to yield a colored chromophore. The color yield is directly proportional to the glucose content of the sample. There are fluorometric versions of this assay for high sensitivity, and one commercial analyzer uses an oxygen electrode to measure the rate of decrease in oxygen concentration in the sample, which is also directly proportional to the glucose concentration.

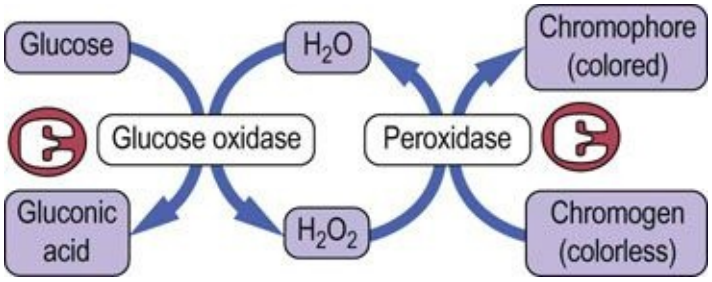


FIG. 6.12 The glucose oxidase/peroxidase assay for blood glucose.
The color produced in this assay is directly proportional to blood glucose concentration.

Reagent strips and glucometers

People with diabetes normally monitor their blood glucose several times a day using reagent strips or glucose meters

The glucose reagent strips are impregnated with a glucose oxidase/peroxidase (GOP) reagent. In the manual version of this assay, the extent of color change on a dipstick is related to glucose concentration – typically on a 1–4 scale. Modern glucometers use a small drop of blood (~1 µL) and amperometric electrodes to measure the current produced by the redox reaction catalyzed by glucose dehydrogenase (GDH), which oxidizes glucose to gluconic acid, but reduces a coenzyme rather than oxygen. These assays are commonly used where rapid or frequent measurements of blood glucose are required. When the GOP and GDH assays were compared at high altitude on a trek up Mount Kilimanjaro, the GOP assay, which depends on ambient oxygen, had a greater error. Both methods were less accurate at the low temperatures at high altitude.

Kinetic assays

Kinetic assays are more rapid than endpoint assays

In the assay described in Figure 6.12 and plotted for several glucose concentrations in Figure 6.13A, the reaction is allowed to proceed to its endpoint, *i.e.* until all the glucose has been oxidized; then the color change is measured. The color yield is then plotted against a standard to determine blood glucose concentration (Fig. 6.13B). High-throughput kinetic analyzers estimate the glucose concentration in a sample by measuring the initial rate of the reaction. Analysis of the kinetic plots in Figure 6.13A, for example, indicates that both the endpoint and the initial rate of the glucose oxidase assay are dependent on glucose concentration. Thus, the analyzer can measure the change in absorbance (or some other parameter) during the early stages of the reaction and compare this rate to that of a standard solution to estimate the glucose concentration (Fig. 6.13C). These assays are performed on flow-injection or centrifugal analyzers to insure rapid mixing of reagents and sample.

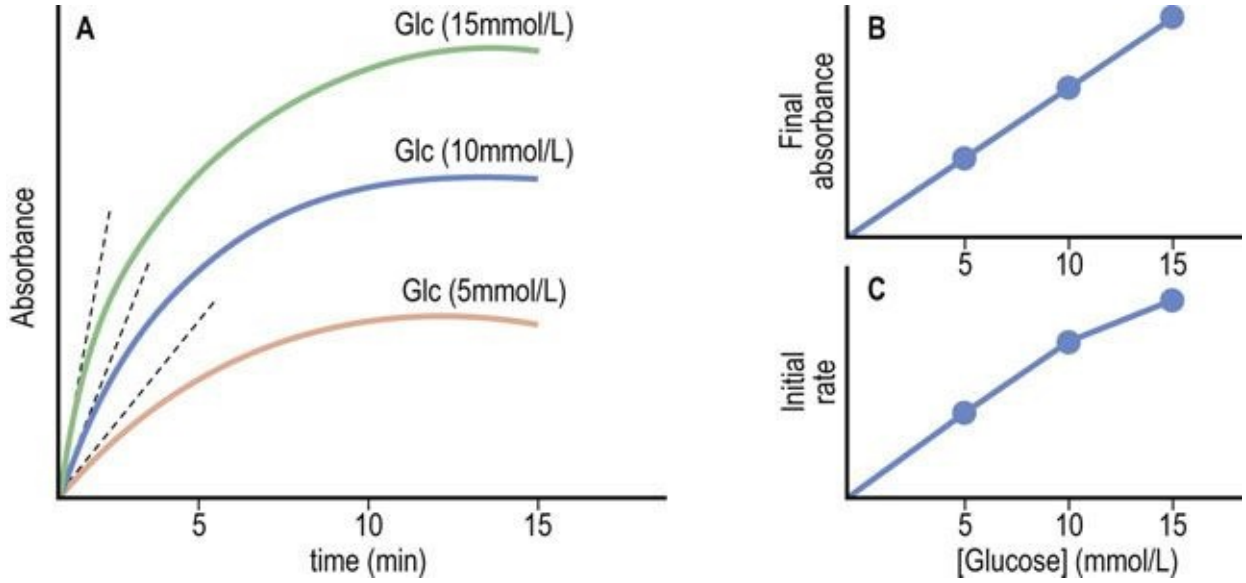


FIG. 6.13 Glucose oxidase/peroxidase assays – endpoint versus kinetic assays. **(A)** Graphical analysis of an endpoint assay. **(B)** The final (endpoint) absorbances are plotted as a function of glucose concentration, yielding a straight line. **(C)** Initial rates of reactions are estimated by multiple measurements early in the assay (dotted lines in frame A), and plotted versus glucose concentration. Nonlinear plots, when obtained, are analyzed by computer.

Kinetic analyzers are inherently faster than endpoint assays because they estimate glucose concentration before the assay reaches its endpoint. These assays work because glucose oxidase and glucose dehydrogenase have a high K_m for glucose. At the concentrations of glucose found in blood, the rate of the oxidase reaction is proportional to glucose concentration, *i.e.* in the first-order region of the Michaelis–Menten equation where the substrate concentration is less than the K_m (see Fig. 6.4).

Summary

■ Most metabolism is catalyzed by biological catalysts called enzymes. Their catalytic activities are apparent at body temperature, and they are strictly regulated by several mechanisms.
■ Both covalent and noncovalent modifications are involved in this regulation and allow for efficient metabolic control. Enzyme activity can be inhibited (or activated) by synthetic compounds (drugs), exogenous compounds (toxins), and endogenous compounds (allosteric effectors).
■ Kinetic analyses of enzymatic reaction are beneficial for evaluating the biological role of enzymes and for elucidating their reaction mechanisms.
■ Assays of enzymes in blood are useful for diagnosis and monitoring of many clinical conditions.

Active learning

1. In a multistep sequence of enzymatic reactions, where is the most effective site for controlling the flux of substrate through the pathway? What effect will an inhibitor of a rate-limiting enzyme have on the concentration of substrates in a multistep pathway?
2. Most drugs are designed to inhibit specific enzymes in biological systems. The drug Prozac has had a profound effect on the medical treatment of depression. Review the history of development of Prozac, illustrating the importance of specificity in the mechanism of drug action.
3. Discuss some examples of reversible and irreversible enzyme inhibitors used in medical practice.
4. Knock-out mice are mice that lack a specific gene. Discuss the impact of KO mice on the direction of drug development in the pharmaceutical industry.

Further reading

Androutsopoulos, VP, Hernandez, AF, Liesivuori, J, *et al.* A mechanistic overview of health associated effects of low levels of organochlorine and organophosphorous pesticides. *Toxicology*. 2012.

Christenson, RH, Duh, SH. Methodological and analytic considerations for blood biomarkers. *Prog Cardiovasc Dis*. 2012; 55:25–33.

Claudino, WM, Goncalves, PH, di Leo, A, *et al.* Metabolomics in cancer: A bench-to-bedside intersection. *Crit Rev Oncol Hematol*. 2012.

DeFeo, EM, Wu, CL, McDougal, WS, *et al.* A decade in prostate cancer: from NMR to metabolomics. *Nat Rev Urol*. 2011; 8(6):301–311.

Lippi, G, Valentino, M, Cervellin, G. Laboratory diagnosis of acute pancreatitis: in search of the Holy Grail. *Crit Rev Clin Lab Sci*. 2012; 49:18–31.

Lopez-Otin, C, Matrisian, LM. Emerging roles of proteases in tumour suppression. *Nat Rev Cancer*. 2007; 7:800–808.

Niemela, O, Alatalo, P. Biomarkers of alcohol consumption and related liver disease. *Scand J Clin Lab Invest*. 2010; 70:305–312.

Öberg, D, Östenson, C-G. Performance of glucose dehydrogenase-and glucose oxidase-based blood glucose meters at high altitude and low temperature. *Diabetes Care*. 2005; 28:1261.

Spratt, BG. The 2011 Garrod Lecture: From penicillin-binding proteins to molecular endocrinology. *J Antimicrob Chemother*. 2012; 67:1578–1588.

Teicher, BA, Linehan, WM, Helman, LJ. Targeting cancer metabolism. *Clin Cancer Res*. 2012; 18:5537–5545.

Zhang, GF, Sadhukhan, S, Tochtrop, GP, *et al.* Metabolomics, pathway regulation, and pathway discovery. *J Biol Chem*. 2011; 286:23631–23635.

Websites

Clinical Enzymology. www.labtestsonline.org/.

International Federation of Clinical Chemistry and Laboratory Medicin. www.ifcc.org/

Chymotrypsin, mechanism of action. www.chembio.uoguelph.ca/educmat/chm258/lecture12.pdf.

Serine proteases. www.med.unibs.it/~marchesi/pps97/course/section12/serprot1.html.

Enzyme classification. www.chem.qmul.ac.uk/iubmb/enzyme/rules.html.

The Comprehensive Enzyme Information Syste. www.brenda-enzymes.info/

A Database of Enzyme Catalytic Mechanism. http://mbs.cbrc.jp/EzCatDB/

The Enzyme Databas. www.enzyme-database.org/

Enzyme nomenclature. www.chem.qmul.ac.uk/iubmb/enzyme/.

Enzyme nomenclature database. http://enzyme.expasy.org/.

Enzyme structures database. www.ebi.ac.uk/thornton-srv/databases/enzymes/.

IntEnz (Integrated relational Enzyme database). www.ebi.ac.uk/intenz/.

A resource portal for enzymes. www.enzymedirectory.com/.

CHAPTER 7

Hemostasis and Thrombosis

Gordon D.O. Lowe and Catherine N. Bagot

Learning objectives

After reading this chapter you should be able to:

- Outline the sequential mechanisms involved in normal hemostasis.
- Summarize the processes through which the vessel wall regulates hemostasis and thrombosis.
- Describe the role of platelets in hemostasis and thrombosis.
- Outline pathways through which antiplatelet drugs act.
- Describe the pathways of blood coagulation, and how these are tested in the clinical hemostasis laboratory to identify coagulation disorders.
- Describe the physiologic inhibitors of blood coagulation.
- Outline pathways through which anticoagulant drugs act.
- Describe the main components of the fibrinolytic system.
- Describe how thrombolytic (fibrinolytic) drugs act.

Introduction

It is essential that blood should not leak excessively from blood vessels when they are injured by the traumas of daily life

Circulation of the blood within the cardiovascular system is essential for transportation of gases, nutrients, minerals, metabolic products, and hormones between different organs. It is also essential that blood should not leak excessively from blood vessels when they are injured by the traumas of daily life. Animal evolution has therefore resulted in the development of an efficient but complex series of hemodynamic, cellular, and biochemical mechanisms that limit such blood loss by forming platelet–fibrin plugs at sites of vessel injury (**hemostasis**). Genetic disorders that result in loss of individual protein functions, and therefore in excessive bleeding (e.g. hemophilia), have played an important part in the identification of many of the biochemical mechanisms in hemostasis.

It is essential also that these hemostatic mechanisms are appropriately controlled by inhibitory mechanisms, otherwise an exaggerated platelet–fibrin plug may produce local occlusion of a major blood vessel (artery or vein) at its site of origin (**thrombosis**), or may break off and block a blood vessel downstream (**embolism**).

Arterial thrombosis is the major cause of heart attacks, stroke, and non-traumatic limb amputations in developed countries (atherothrombosis is discussed in Chapter 18). Venous thrombosis and embolism are also major causes of death and disability. Clinical use of antithrombotic drugs (antiplatelet, anticoagulant, and thrombolytic agents) is now widespread in developed countries, and requires an understanding of how they interfere with hemostatic mechanisms to exert their antithrombotic effects.

Hemostasis

Hemostasis means ‘the arrest of bleeding’

After tissue injury that ruptures smaller vessels (including everyday trauma, injections, surgical incisions, and tooth extractions), a series of interactions between the vessel wall and the circulating blood normally occurs, resulting in cessation of blood loss from injured vessels within a few minutes (hemostasis). Hemostasis results from effective sealing of the ruptured vessels by a hemostatic plug composed of blood platelets and fibrin. Fibrin is derived from circulating fibrinogen, whereas platelets are small cell fragments that circulate in the blood and have an important role in the initiation of hemostasis.

Hemostasis requires the coordinated function of blood vessels, platelets, coagulation factors and the fibrinolytic system

Figure 7.1 provides an overview of hemostatic mechanisms and illustrates some of the interactions between blood vessels, platelets, and the coagulation system in hemostasis; each of these components of hemostasis also interacts with the fibrinolytic system. The initial response of small blood vessels to injury is arteriolar vasoconstriction, which temporarily reduces local blood flow. Flow reduction transiently reduces blood loss, and may also promote formation of the platelet–fibrin plug. Activation of blood platelets is followed by their adhesion to the vessel wall at the site of injury, and their subsequent aggregation to each other, building up an occlusive platelet mass that forms the initial (primary) hemostatic plug. This platelet plug is friable and, unless subsequently stabilized by fibrin, will be washed away by local blood pressure when vasoconstriction reverses.

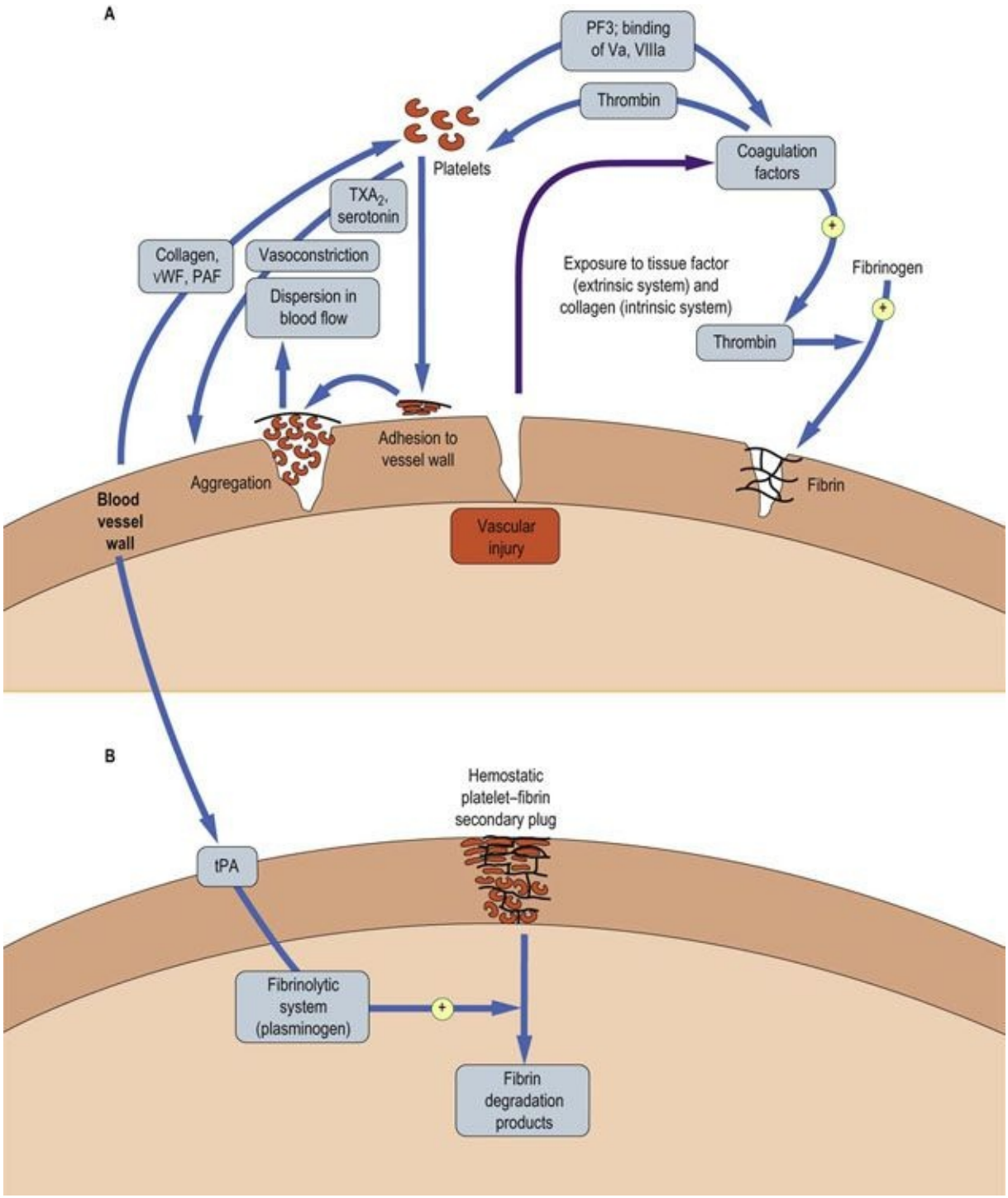


FIG. 7.1 Overview of hemostatic mechanisms.
(A) Vascular injury sets in motion a series of events that culminate in formation of a primary plug of platelets. This can be dispersed by blood flowing through the vessel unless the plug is stabilized. **(B)** The primary plug is stabilized by a network of fibrin (formed from crosslinked fibrinogen). The secondary plug is stable and is degraded only when the fibrinolytic system has been activated. PAF, platelet-activating factor; PS, phosphatidylserine; tPA, tissue-type plasminogen activator; TXA_2, thromboxane A_2; Va,

activated coagulation factor V; VIIIa, activated coagulation factor VIII; vWF, von Willebrand factor. Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012.

Vascular injury also activates coagulation factors, which interact sequentially to form thrombin, which converts circulating soluble plasma fibrinogen to insoluble, crosslinked fibrin. This forms the subsequent (secondary) hemostatic plug, which is relatively resistant to dispersal by blood flow or fibrinolysis. There are two pathways of the activation of coagulation factors: the extrinsic pathway, which is initiated by the exposure of the flowing blood to tissue factor, released from subendothelial tissue, and the intrinsic pathway, which has an important amplification role in generating thrombin and fibrin.

The lysis of fibrin is as important to health as its formation

Hemostasis is a continuous process throughout life, and would result in excessive fibrin formation and vascular occlusion if unchecked. Evolution has therefore produced a fibrinolytic system; this is activated by local fibrin formation, resulting in local generation of plasmin, an enzyme which digests fibrin plugs (in parallel with tissue repair processes), thus maintaining vascular patency. Digestion of fibrin results in generation of circulating fibrin degradation products (FDPs). These are detectable in plasma of healthy individuals at low concentration, which illustrates that fibrin formation and lysis are continuing processes in health.

Excessive bleeding may result from defects in each of the components of hemostasis, which may be caused by disease (congenital or acquired) or by antithrombotic drugs (Table 7.1). The vascular, platelet, coagulation and fibrinolytic components of hemostasis will now be discussed in turn.

Table 7.1

Congenital and acquired causes of excessive bleeding

	Congenital	Acquired
Vessel wall	Disorders of collagen synthesis (Ehlers–Danlos syndrome)	Vitamin C deficiency (scurvy) Corticosteroid excess
Platelets	Disorders of adhesion vWF deficiency (von Willebrand disease) Platelet GPIb-IX deficiency (Bernard–Soulier syndrome) Disorders of aggregation Platelet GPIIb-IIIa deficiency (Glanzmann's thrombasthenia) Disorders of storage granules (i.e.storage pool disorders affecting alpha granules, dense granules or both) Disorders of platelet secretion and signal transduction (e.g. defects in platelet-agonist interactions, abnormalities in arachidonic acid pathway)	Antiplatelet drugs (e.g. aspirin, dipyridamole, clopidogrel) Defective formation of platelets Excessive destruction of platelets
Coagulation	Coagulation factor deficiencies (hemophilias): factor VIII factor IX factor XI fibrinogen, etc.	Vitamin K deficiency (factors II, VII, IX, X) Parenteral anticoagulants e.g. unfractionated heparin (UFH), low-molecular-weight heparin (LMWH) Oral anticoagulants (vitamin K antagonists, e.g. warfarin, direct thrombin inhibitors, e.g. dabigatran, direct Xa inhibitors, e.g. rivaroxaban), liver disease Disseminated intravascular coagulation (DIC)
Fibrinolysis	Antiplasmin deficiency PAI-1 deficiency	Fibrinolytic drugs (e.g. tPA, urokinase, streptokinase)

GPIb-IX, GPIIb-IIIa, glycoprotein receptors Ib-IX and IIb-IIIa; PAI-1, plasminogen activator inhibitor type 1; tPA tissue plasminogen activator.

The vessel wall

Vascular injury has a key role in initiating local formation of the platelet–fibrin plug and in its subsequent removal by the fibrinolytic system

All blood vessels are lined by a flat sheet of endothelial cells, which have important roles in the interchange of chemicals, cells, and microbes between the blood and the body tissues. Endothelial cells in the smallest blood vessels (capillaries) are supported by a thin layer of connective tissue, rich in collagen fibers, called the intima. In veins, a thin layer (the media) of contractile smooth muscle cells allows some venoconstriction: for example, superficial veins under the skin constrict in response to surface cooling. In arteries and arterioles, a well-developed muscle layer allows powerful vasoconstriction, including the vasoconstriction after local injury that forms part of the hemostatic response. Larger vessels also have a supportive connective tissue outer layer (the adventitia).

Normal endothelium has an antithrombotic surface

Intact normal endothelium does not initiate or support platelet adhesion or blood coagulation. Its surface is antithrombotic. This thrombo-resistance is partly due to endothelial production of two potent vasodilators and inhibitors of platelet function: prostacyclin (prostaglandin I_2, PGI_2) and nitric oxide, otherwise known as endothelium-derived relaxing factor (EDRF).

Advanced concept box
Prostacyclin and nitric oxide: biochemical mediators of vasoconstriction and vasodilatation

The diameters of arteries and arterioles throughout the body continuously alter to regulate blood flow according to local and general metabolic and cardiovascular requirements. Control mechanisms include neurogenic (sympathetic/adrenergic; Chapter 41.1) and myogenic pathways, and local biochemical mediators,

including prostacyclin (PGI_2) and nitric oxide.

Prostacyclin is the major arachidonic acid metabolite formed by vascular cells. It is a potent vasodilator, and also a potent inhibitor of platelet aggregation. It has a short half-life in plasma (3 min).

Nitric oxide is also a potent vasodilator formed by vascular endothelial cells, also with a short half-life. It was initially termed endothelium-derived relaxing factor (EDRF). In common with that of prostacyclin, its generation by endothelial cells is enhanced by many compounds, and also by blood flow and shear stress (the tangential force applied to the cells by the flow of blood). In the normal circulation, nitric oxide appears to have a key role in flow-mediated vasodilatation. It is synthesized by two distinct forms of endothelial nitric oxide synthase (eNOS): constitutive and inducible. Constitutive eNOS rapidly provides relatively small amounts of nitric oxide for short periods, related to vascular flow regulation. The beneficial effects of nitrate drugs in hypertension and angina may partly reflect their effects on this pathway (Chapter 18). Inducible eNOS is stimulated by cytokines in inflammatory reactions, and releases large amounts of nitric oxide for long periods. Its suppression by glucocorticoids may partly account for their antiinflammatory effects.

Both prostacyclin and nitric oxide appear to exert their vasodilator actions by diffusing locally from endothelial cells to vascular smooth muscle cells, where they stimulate guanylate cyclase, resulting in increased formation of cyclic guanosine 3′5′-monophosphate (cGMP) and relaxation of vascular smooth muscle via alteration of the intracellular calcium concentration (Chapter 40).

Endothelial damage exposes blood to the tissue factor and to collagen

The vasoconstriction that occurs after vascular injury is partly mediated by two platelet activation products: serotonin (5-hydroxytryptamine), and thromboxane A_2 (TXA_2), a product of platelet prostaglandin metabolism. The endothelial cell

damage also exposes flowing blood to subendothelial tissue factor, which activates the extrinsic pathway of blood coagulation (Fig. 7.1). In addition, after a vascular injury that disrupts the endothelial cell lining, flowing blood is exposed to subendothelial collagen, which activates the intrinsic pathway of blood coagulation.

Advanced concept box
Thromboxane A_2 and aspirin

It has already been noted that prostacyclin, PGI_2, the major arachidonic acid metabolite formed by vascular cells, is a potent vasodilator and inhibitor of platelet aggregation. In contrast, the major arachidonic acid metabolite formed by platelets is **thromboxane A_2 (TXA_2)**, which is a potent vasoconstrictor and stimulates platelet aggregation. In common with prostacyclin, TXA_2 has a short half-life. In the late 1970s, Salvador Moncada and John Vane contrasted the effects of PGI_2 and TXA_2 on blood vessels and platelets, and hypothesized that a balance between these two compounds was important in the regulation of hemostasis and thrombosis.

Congenital deficiencies of cyclooxygenase or thromboxane synthase (the enzymes involved in TXA_2 synthesis) result in a mild bleeding tendency. Ingestion of even low doses of **acetylsalicylic acid (aspirin)** irreversibly acetylates cyclooxygenase and suppresses TXA_2 synthesis and platelet aggregation for several days, resulting in an antithrombotic effect and a mild bleeding tendency. Bleeding is especially likely from the stomach, as a result of the formation of stomach ulcers secondary to the inhibition of cytoprotective gastric mucosal prostaglandins by aspirin. Although in persons at high risk of arterial thrombosis, *e.g.* previous myocardial infarction, this bleeding tendency is outweighed by a reduction in risk of thrombosis, aspirin is contraindicated in individuals with a history of bleeding disorders, or existing stomach or duodenal ulcers.

Exposure of flowing blood to collagen as a result of endothelial damage also stimulates platelet activation

Platelets bind to collagen via von Willebrand factor (vWF), which is released from the endothelial cells. vWF in turn binds both to collagen fibers and to platelets (via a platelet membrane glycoprotein receptor, GPIb-IX). Platelet-activating factor (PAF) from the vessel wall may also activate platelets in hemostasis (Fig. 7.1; see also box on p. 27).

Collagen has a key role in the structure and hemostatic function of small blood vessels

Because collagen has a key role in the structure and hemostatic function of small blood vessels, vascular causes of excessive bleeding include congenital or acquired deficiencies of collagen synthesis (Table 7.1). Congenital disorders include the rare **Ehlers–Danlos syndrome**. Acquired disorders include the relatively common vitamin C deficiency, scurvy (Chapter 11), and excessive exogenous or endogenous corticosteroids.

Platelets and platelet-related bleeding disorders

Blood platelets form the initial hemostatic plug in small vessels, and the initial thrombus in arteries and veins

Platelets are circulating, anuclear microcells of mean diameter 2–3 μm. They are fragments of bone marrow megakaryocytes, and circulate for approximately 10 days in the blood. The concentration of platelets in normal blood is 150–400 × 10^9/L

Advanced concept box Platelet activation exposes glycoprotein receptors

Platelets can be activated by several chemical agents, including adenosine diphosphate (ADP, released by platelets, erythrocytes, and endothelial cells), epinephrine, collagen, thrombin, and PAF; by infection *e.g.* HIV, *Helicobacter pylori*; and by high physical shear stresses. Most of the chemical agents appear to act by binding to specific receptors on the platelet surface membrane. After receptor stimulation, several pathways of platelet activation can be initiated, resulting in several phenomena:

- **Change in platelet shape** from a disk to a sphere with extended pseudopodia, which facilitates aggregation and coagulant activity.
- **Release of several compounds involved in hemostasis** from intracellular granules, *e.g.* ADP, serotonin, fibronectin and vWF.
- **Aggregation**, via exposure of GPIb-IX membrane receptor and linking by vWF (under high shear conditions), and via exposure of another membrane glycoprotein receptor, GPIIb-IIIa, and linking by fibrinogen (under low shear conditions).
- **Adhesion to the vessel wall** via exposure of the GPIb-IX membrane receptor, through which vWF binds platelets to subendothelial collagen.

Finally, stimulation of the platelet membrane receptor triggers

the activation of platelet membrane phospholipases, which hydrolyze membrane phospholipids, releasing arachidonic acid. Arachidonic acid is metabolized by cyclooxygenase and thromboxane synthase to TXA_2, a potent but labile (half-life 30 seconds) mediator of platelet activation and vasoconstriction.

Congenital defects in platelet adhesion/aggregation can cause lifelong excessive bleeding

A simple screening test – measurement of the skin bleeding time (reference range, 2–9 minutes) – can frequently detect congenital defects of platelet adhesion/aggregation, in which the time is characteristically prolonged. The most common such defect is von Willebrand disease (Table 7.1), a group of both autosomal dominant and autosomal recessive disorders that result in either quantitative or qualitative defects of vWF multimers. These multimers are composed of subunits (molecular weight 220–240 kDa) that are released from storage granules known as the Weibel–Palade bodies in endothelial cells and alpha granules in platelets. Not only does vWF have an important role in platelet hemostatic function but it also transports coagulation factor VIII (antihemophilic factor) in the circulation and delivers it to sites of vascular injury. Hence, plasma concentrations of factor VIII may also be low in **von Willebrand disease**. Treatment of this disease is to increase the low plasma vWF activity, usually by means of either desmopressin (a synthetic analogue of vasopressin (Chapter 24), which releases vWF from endothelial cells into plasma), or administering VWF concentrates derived from human plasma.

Less common congenital platelet-related bleeding disorders include **GPIb-IX deficiency** (Bernard–Soulier syndrome), **GPIIb-IIIa deficiency** (Glanzmann's thrombasthenia), and **fibrinogen deficiency** (because fibrinogen bridges GPIIb-IIIa receptors of adjacent platelets).

Acquired disorders may be caused by defective formation, and excessive destruction or consumption of platelets

Acquired disorders of platelets include a low platelet count (**thrombocytopenia**), which may be the result of either defective formation of

platelets by bone marrow megakaryocytes, *e.g.* in myelodysplasia or acute myeloid leukaemia, excessive destruction of platelets, *e.g.* by antiplatelet antibodies, and excessive consumption of platelets, *e.g.* in disseminated intravascular coagulation or by sequestration in an enlarged spleen.

Antiplatelet drugs are used in the prevention or treatment of arterial thrombosis

Antiplatelet drugs are used in the prevention or treatment of arterial thrombosis; their sites of action are illustrated in Figure 7.2. Aspirin inhibits cyclooxygenase and hence reduces the formation of TXA_2. Because it also has the effect of reducing the formation of PGI_2, which itself has antiplatelet activity, agents acting more specifically as thromboxane synthase inhibitors, *e.g.* picotamide, or thromboxane receptor antagonists, such as ifetroban, have also been investigated as potential antiplatelet agents. However, these do not appear to be more effective than aspirin. Dipyridamole acts by both reducing the availability of ADP and inhibiting thromboxane synthase, and ticlopidine and clopidogrel inhibit the ADP receptor (Fig. 7.2). These drugs have antithrombotic effects similar to those of aspirin, but cause less gastric bleeding because they do not interfere with synthesis of prostaglandins in the stomach. GPIIb-IIIa antagonists, *e.g.* tirofiban or abciximab, can also be used in acute coronary thrombosis. Each of these antiplatelet drugs adds to the antithrombotic efficacy of aspirin but also increases the risk of bleeding when used in combination.

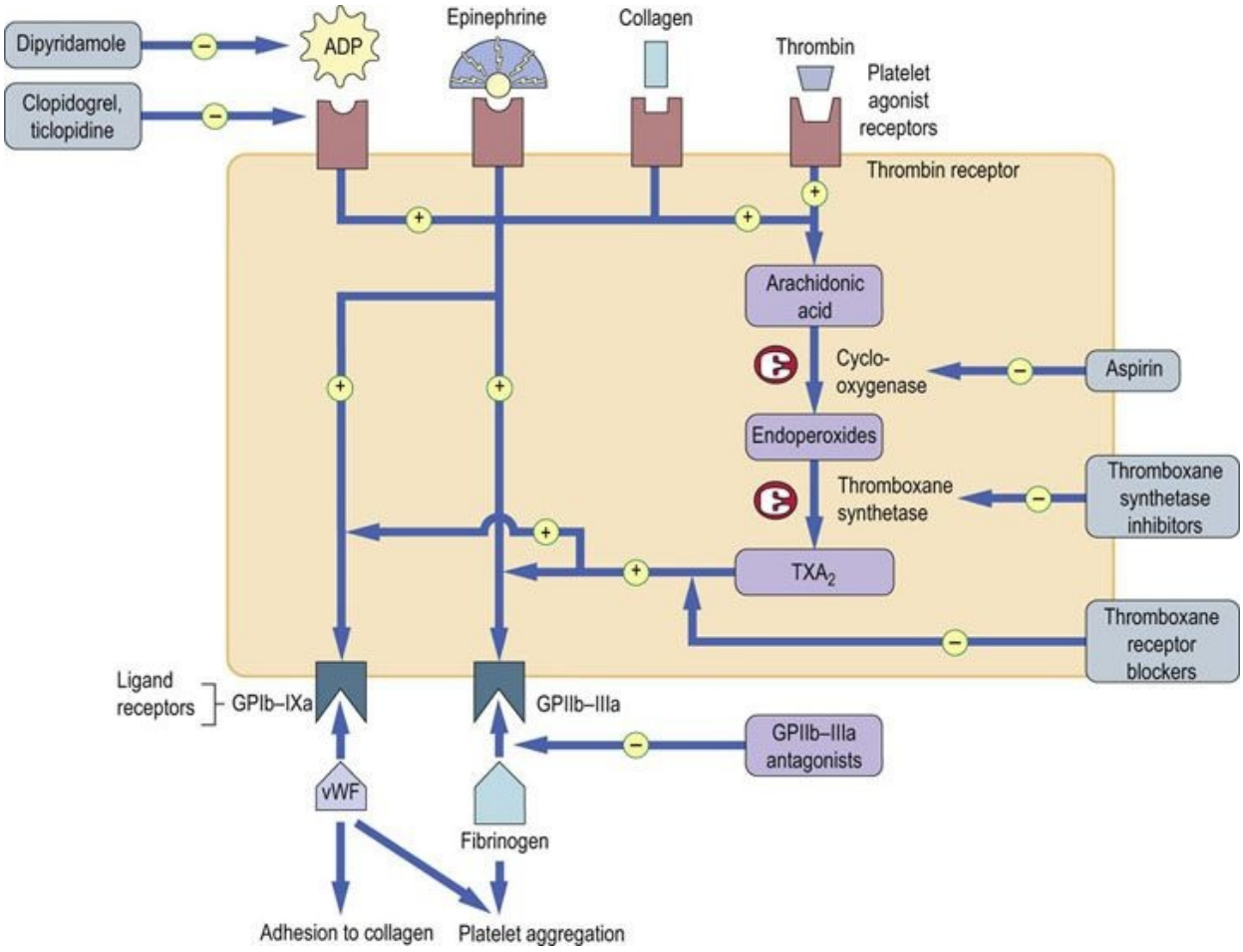


FIG. 7.2 Pathways of platelet activation and mechanisms of action of antiplatelet drugs. Stimulation of platelet agonist receptors results in exposure of platelet ligand receptors, partly through the platelet prostaglandin (cyclooxygenase) pathway. Ligand receptors bind vWF and fibrinogen in platelet adhesion/aggregation. vWF, von Willebrand factor; TXA_2, thromboxane A_2.

Advanced concept box Platelet membrane receptors: their ligands vWF and fibrinogen

Platelets have a key role in hemostasis and thrombosis, through adhesion to the vessel wall and subsequent aggregation to form a platelet-rich hemostatic plug or thrombus. These processes involve exposure of specific membrane glycoprotein receptors after platelet activation by several compounds.

Platelet receptor GPIb-IX plays a key part in the adhesion of

platelets to subendothelium. It binds vWF, which also interacts with specific subendothelial receptors, including those on subendothelial collagen. Congenital deficiencies of GPIb-IX (Bernard–Soulier syndrome) or, more commonly, of vWF, result in a bleeding tendency.

Another receptor, **GPIIb-IIIa**, has a key role in platelet aggregation. After platelet activation, hundreds of thousands of GPIIb-IIIa receptors can be exposed in a single platelet. These receptors interact with fibrinogen or vWF, which bind platelets together, forming a hemostatic or thrombotic plug. Congenital deficiency of GPIIb-IIIa (the rare Glanzmann's thrombasthenia) causes a severe bleeding disorder; in contrast, deficiencies of either fibrinogen or vWF cause a milder bleeding disorder, because these two ligands can substitute for each other. GPIIb-IIIa inhibitors (e.g. tirofiban, abciximab) have been developed for patients undergoing angioplasty for coronary artery disease to prevent further coronary events.

Coagulation

Blood coagulation factors interact to form the secondary, fibrin-rich, hemostatic plug in small vessels, and the secondary fibrin thrombus in arteries and veins

Plasma coagulation factors are identified by Roman numerals: they are listed in Table 7.2, together with some of their properties. Tissue factor was formerly known as factor III, calcium ion as factor IV; factor VI does not exist.

Table 7.2

Coagulation factors and their properties

Factor	Synonyms	Molecular weight (Da)	Plasma concentration (mg/dL)
I	Fibrinogen	340,000	200–400
II	Prothrombin	70,000	10
III	Tissue factor (thromboplastin)	44,000	0
IV	*Calcium ion	40	9–10
V	Proaccelerin, labile factor	330,000	1
VII	Serum prothrombin conversion accelerator (SPCA), stable factor	48,000	0.05
VIII	Antihemophilic factor (AHF)	220,000	0.01
(vWF)		(250,000)n	1
IX	Christmas factor	55,000	0.3
X	Stuart–Prower factor	59,000	1
XI	Plasma thromboplastin antecedent (PTA)	160,000	0.5
XII	Hageman factor	80,000	3
XIII	Fibrin-stabilizing factor (FSF)	32,000	1–2
Prekallikrein	Fletcher factor	85,000	5
High-molecular-weight kininogen (HMWK)	Fitzgerald, Flaujeac or Williams factor, contact activation cofactor	120,000	6

n indicates number of subunits.

*To convert calcium ion to mmol/L divide by 4.

The coagulation cascade

Figure 7.3 illustrates the currently accepted scheme of blood coagulation. Since the early 1960s, this has been accepted as a ‘waterfall’ or ‘cascade’ sequence of interactive proenzyme to enzyme conversions, each enzyme activating the next proenzyme in the sequence(s). Activated factor enzymes are designated by the letter ‘a’: for example, factor XIa. Although the process of blood coagulation is complex and nonlinear, traditionally, the scheme has been divided into three parts:

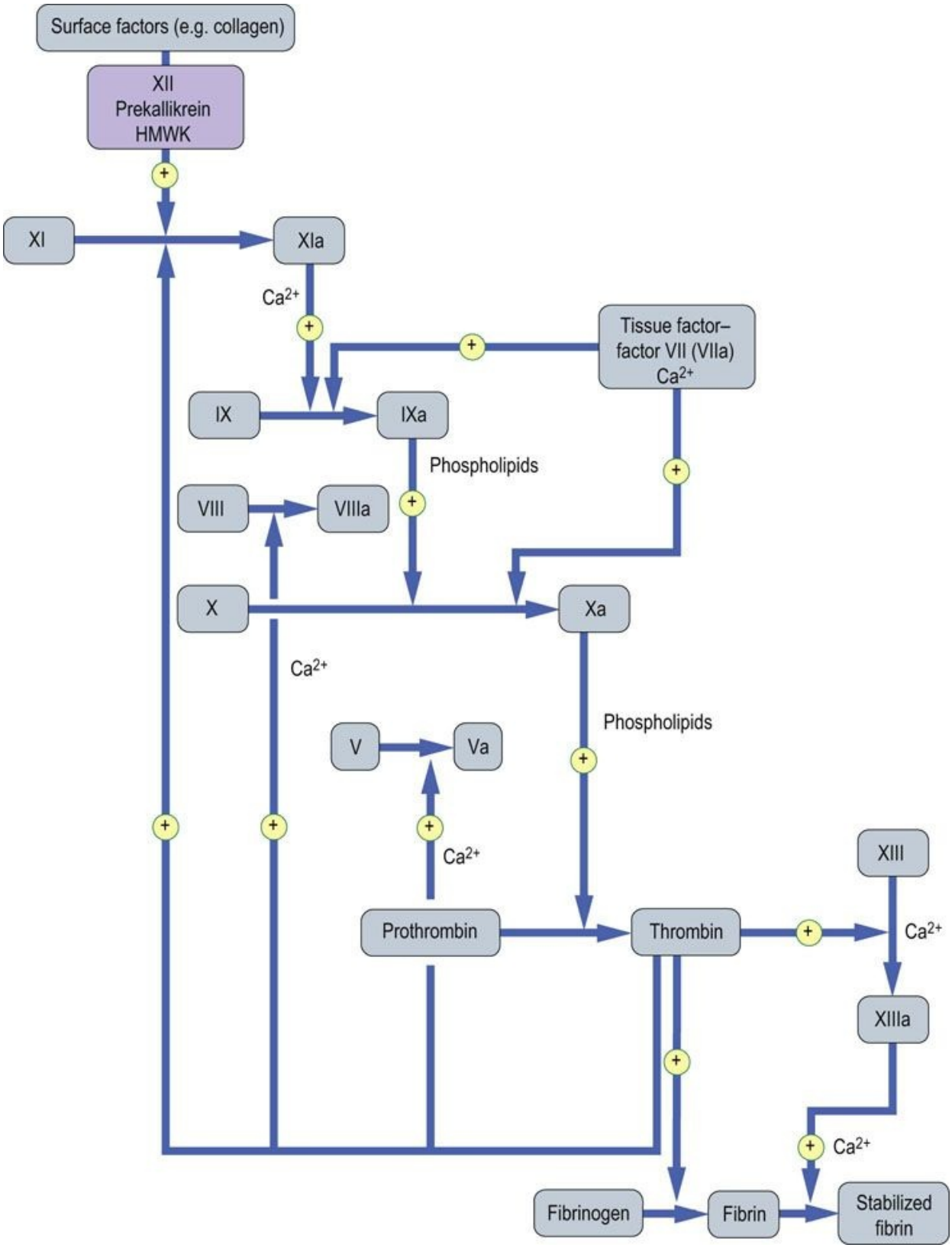


FIG. 7.3 Blood coagulation: activation of coagulation factors.
After the initiation of blood coagulation, the coagulation factor proenzymes are activated

sequentially: activated factor enzymes are designated by the letter ‘a’. The purple box indicates contact factors that have no apparent function in in vivo hemostasis. Phospholipids are supplied in vivo by platelets. The intrinsic pathway: blue arrows. The extrinsic pathway: red arrows. The common pathway: green arrows. HMWK, high-molecular-weight kininogen. Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012.

- the intrinsic pathway
- the extrinsic pathway, and
- the final common pathway.

The status of the intrinsic, extrinsic and final common pathway is assessed by specific laboratory tests

The three components of the coagulation system are distinguished on the basis of the nature of the initiating factor and its corresponding test in the clinical hemostasis laboratory; hence, three tests of coagulation are performed in clinical laboratories on citrated, platelet-poor plasma:

- activated partial thromboplastin time (APTT), testing the intrinsic pathway
- prothrombin time (PT) – the extrinsic pathway, and
- thrombin clotting time (TCT) testing the final common pathway.

Platelet-poor plasma is used in these tests because the platelet count influences clotting time results. To obtain platelet-poor plasma, blood is collected in tubes containing citrate anticoagulant to sequester calcium ions reversibly, and the blood is centrifuged at 2000 *g* for 15 minutes. The coagulation time tests are initiated by adding calcium and appropriate initiating agents.

However, these tests have their limitations in describing the in vivo phenotype of a patient's blood to coagulate effectively. The so-called global assays of coagulation have therefore been developed, and they are thought to better reflect an individual's ability to clot. These include thromboelastography and thrombin generation.

Clinical test boxGlobal coagulation assays

Thromboelastography (TEG) and **rotational thromboelastometry (ROTEM)** assess the ability of whole blood to clot in response to a mechanical stimulus, allowing an assessment of all aspects of hemostasis: platelet function, fibrin

crosslinking and fibrinolysis.

Thrombin generation assay is a global coagulation assay thought better able to assess an individual's ability to coagulate than standard coagulation assays. Tests such as the PT and APTT described above measure only 5% of the total thrombin generated: *i.e.* at the time of first clot generation (Fig. 7.4).

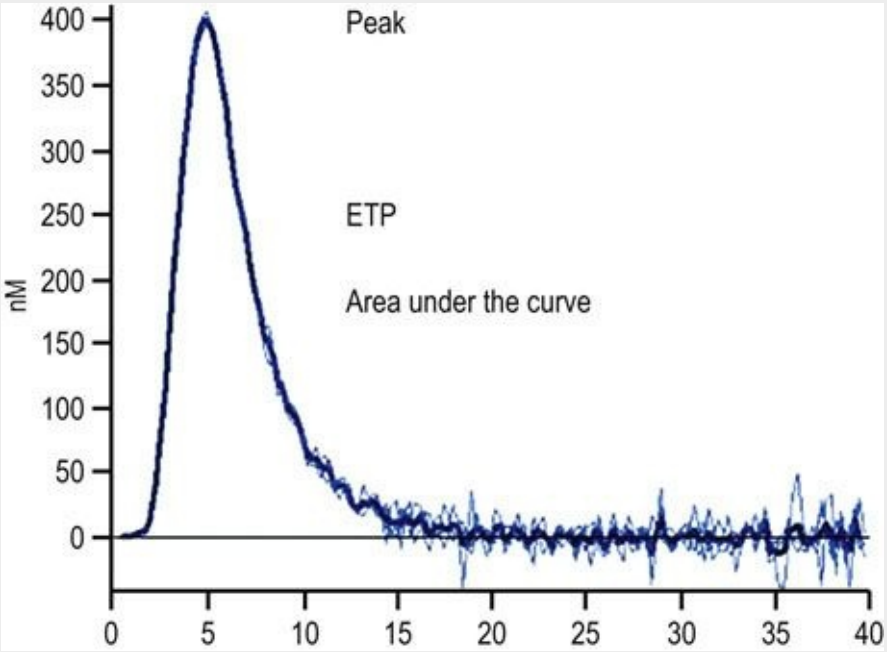


FIG. 7.4 Thrombin generation curve.
The curve is a measurement of thrombin concentration over time with the parameters lag time, time to the first development of thrombin, peak thrombin, the maximum amount of thrombin generated, time to peak, and endogenous thrombin potential (ETP) or the area under the curve.

Thrombin is central to the coagulation cascade in that it converts fibrinogen to fibrin and also has numerous positive and negative feedback roles. The measurement of thrombin generation enables quantification over time of all thrombin generated in a plasma sample by its ability to ‘cut’ either a chromophore or fluorochrome and measure the resultant chromogenic or fluorescent activity.

Despite promising results from ROTEM and thrombin generation, both assays are limited by numerous preanalytical and analytical variables, which make comparisons between laboratories

difficult. There is still no reliable standardization of these assays with robust internal or external quality control. For this reason, both remain research tools.

Congenital deficiencies of coagulation factors (I–XIII) result in excessive bleeding

Congenital deficiencies of coagulation factors (I–XIII) result in excessive bleeding, which illustrates their physiologic importance in hemostasis. The exception is factor XII deficiency, which does not increase the bleeding tendency, despite prolonging blood clotting times in vitro; the same is true for its cofactors, prekallikrein or high-molecular-weight kininogen (HMWK). A possible explanation for this is given below.

Activated partial thromboplastin time (APTT) assesses the intrinsic pathway

The term ‘intrinsic’ implies that no extrinsic factor such as tissue factor or thrombin is added to the blood, besides a contact with nonendothelial ‘surface’. The clinical test of this pathway is the activated partial thromboplastin time (APTT), also known as the kaolin–cephalin clotting time (KCCT) because kaolin (microparticulated clay) is added as a standard ‘surface’, and cephalin (brain phospholipid extract) as a substitute for platelet phospholipid. The reference range of the APTT is about 30–40 seconds; prolongations are observed in deficiencies of factors XII (or its cofactors, prekallikrein or HMWK), XI, IX (or its cofactor, factor VIII), X (or its cofactor, factor V), or prothrombin (factor II) (see Table 7.1 and 7.2).

The test is used to exclude the common congenital hemophilias (deficiencies of factors VIII, IX or XI), and to monitor unfractionated heparin treatment. Hemophilias caused by factor VIII or IX deficiency occur in approximately 1 in 5000 and 1 in 30,000 males, respectively; inheritance is X-linked recessive, transmitted by carrier females. Treatment is usually with recombinant factor VIII or IX concentrates.

Clinical boxA boy with extensive

bruising: classic Hemophilia (Congenital Factor VIII Deficiency)

A 3-year-old boy was admitted from the emergency room of his local hospital because of extensive bruising after a fall down a few stairs. A routine coagulation screen test showed a greatly prolonged APTT of more than 150 seconds (normal range 30–40 seconds). Assay of coagulation factor VIII showed a very low level; the vWF level was normal. His mother recollected a family history of excessive bleeding, which had affected her brother and father.

Comment.

Because of this typical history of an X-linked recessive bleeding disorder, a low coagulation factor VIII level, and a normal vWF level, the diagnosis of congenital factor VIII deficiency was made. The family were referred to the local hemophilia center and counseled about the risks of further affected sons and carrier daughters. The child was treated with intravenous recombinant factor VIII concentrate for the present bleed, and prophylactically in the future to prevent further bleeding episodes.

Prothrombin time assesses the extrinsic pathway

The term ‘extrinsic’ refers to the effect of tissue factor, which (after combining with coagulation factor VII) greatly accelerates coagulation, by activating both factor IX and factor X (Fig. 7.3). Tissue factor is a polypeptide that is expressed in all cells other than endothelial cells. The clinical test of this pathway is the prothrombin time (PT), in which tissue factor is added to plasma. The reference range is approximately 10–15 seconds; prolongations are observed in deficiencies of factors VII, X, V, or II. In clinical practice, the test is used to diagnose both the rare congenital defects of these factors and, much more commonly, acquired bleeding disorders, resulting from:

■ **Vitamin K deficiency**, *e.g.* in malabsorption or obstructive jaundice (Chapter

11), which reduces hepatic synthesis of factors II, VII, IX and X. Treatment is by oral or intravenous administration of vitamin K.

■ **Administration of Oral vitamin K antagonists**, *e.g.* warfarin, which reduce hepatic synthesis of these factors. Excessive bleeding in patients taking warfarin can be treated by stopping the drug, giving vitamin K or replacing factors II, VII, IX and X with prothrombin complex concentrates containing only the relevant factors, *e.g.* Beriplex or fresh frozen plasma.

■ **Liver disease**, which reduces hepatic synthesis of these factors. For example, the prothrombin time is a prognostic marker of liver failure after acetaminophen (paracetamol) overdose (Chapter 30). Treatment is by replacing factors II, VII, IX and X with fresh frozen plasma.

Clinical test boxMonitoring oral anticoagulant therapy

Oral anticoagulant therapy with **vitamin K antagonists**, *e.g.* **warfarin**, is given long term to patients at risk of thrombosis within the chambers of the heart, *e.g.* patients with atrial fibrillation or heart valve prostheses, which may embolize to the brain, causing a stroke.

Monitoring of an internationally standardized prothrombin time, *i.e.* the **International Normalized Ratio (INR)**, every few weeks is essential to minimize the risk not only of thromboembolism but also of excessive bleeding. Up to 1% of the adult population in developed countries now receive long-term anticoagulants; hence traditional monitoring by doctors and nurses (taking blood samples, sending them to the laboratory, getting results and giving dosage instructions to patients) has created an unsustainable workload.

In recent years, near-patient or point-of-care INR testing has become available for warfarin monitoring with a ‘finger prick’ capillary sample being drawn into a portable INR analyzer. With this technique some patients can self-test and occasionally self-monitor, similar to blood glucose self-monitoring by persons with diabetes. Computerized algorithms for dosing warfarin have also been developed to assist health care workers to alter dosing accurately. New oral anticoagulants, *e.g.* dabigatran and

rivaroxaban, for patients with atrial fibrillation have recently been introduced, which require no monitoring of anticoagulation levels.

Thrombin clotting time assesses the final common pathway

The term ‘final common pathway’ refers to the conversion of prothrombin to thrombin via Xa, with Va acting as a cofactor

This in turn allows the conversion of fibrinogen to fibrin. This final stage of fibrin production in the common pathway is tested clinically by the thrombin clotting time (TCT), in which exogenous thrombin is added to plasma. The reference range of values is approximately 10–15 seconds; prolongations are observed in fibrinogen deficiency and in the presence of inhibitors, *e.g.* heparin, fibrin degradation products. Fibrinogen deficiency may be congenital or due to acquired consumption of fibrinogen in disseminated intravascular coagulation (DIC), or may occur after administration of fibrinolytic drugs (see below). Treatment is with cryoprecipitate or fibrinogen concentrates.

Several assays assess platelet function

Apart from assessing platelet number, size and morphology through the ‘complete blood count’ analysis and blood film review, platelet function can also be assessed in other ways.

One method of platelet function assessment is the **Platelet Function Analyzer** (PFA-100, Siemens). Whole blood is passed over a cartridge containing an aperture coated with a combination of two platelet agonists: *e.g.* collagen/epinephrine, or collagen/ADP. The time to aperture closure as a result of platelet aggregation is measured. It cannot define specific disorders but an abnormal result is suggestive of a platelet disorder and can be used as a screening test.

Light transmission aggregometry (LTA) is considered the gold standard for investigating specific disorders of platelet function. Platelet-rich plasma is exposed to various platelet agonists, *e.g.* collagen, ADP and epinephrine, and light transmission is monitored to produce standard curves. The pattern of curves

obtained with a combination of agonists can help determine which platelet function defect is present.

The production and release of platelet nucleotides, *i.e.* ATP and ADP, can be measured to assess nucleotide production and nucleotide release from granules. Flow cytometry analysis of various **platelet receptors** can also be performed.

Such an array of tests should mean that platelet function disorders should be easy to diagnose. However, preanalytical and analytical variables mean that results can often be unreliable and the results are often difficult to interpret.

Thrombin

Thrombin converts circulating fibrinogen to fibrin and activates factor XIII, which crosslinks the fibrin, forming a clot

It is currently believed that activation of blood coagulation is usually initiated by vascular injury, causing exposure of flowing blood to tissue factor, which results in activation of factors VII and IX. Subsequently, activation of factors X and II (prothrombin) occurs preferentially at sites of vascular injury, alongside activated platelets: the latter provide procoagulant activity as a result of exposure of negatively charged platelet surface membrane phospholipids, such as phosphatidylserine, and high affinity binding sites for several activated coagulation factors, allowing the formation of the prothrombinase complex (Va, Xa and II) and tenase complex (VIIIa, IXa and Xa), which both greatly enhance the production of thrombin. As a result of these biochemical interactions, thrombin and fibrin formation are efficiently localized at sites of vascular injury.

Thrombin has a central role in hemostasis

Not only does thrombin convert circulating fibrinogen to fibrin at sites of vascular injury, producing the secondary, fibrin-rich hemostatic plug, It also activates factor XIII, a transglutaminase, which crosslinks fibrin, rendering it resistant to dispersion by local blood pressure or by fibrinolysis (Figs 7.1 and 7.3). Furthermore, thrombin stimulates its own generation in a positive feedback cycle in three ways:

- **It catalyzes activation of factor XI:** this may explain why congenital deficiencies of factor XII, prekallikrein or HMWK are not associated with

excessive bleeding (Fig. 7.3).

- **It catalyzes activation of factors VIII and V.**
- **It activates platelets** (Fig. 7.2).

Thrombin inhibitors have been developed as anticoagulant drugs

Now that the central role of thrombin in hemostasis and thrombosis has been recognized, a number of direct thrombin inhibitors (DTIs) have been developed as anticoagulant drugs. Dabigatran, an oral direct thrombin inhibitor (DTI), has been demonstrated in large randomized controlled clinical trials to be as effective as warfarin in the treatment of acute venous thrombosis and in the prevention of stroke in patients with atrial fibrillation. The major advantage of this drug for these indications is that dose monitoring is not required. Recently it has obtained a licence in the UK for stroke prevention in atrial fibrillation. However, to date there is no effective antidote and for this reason, it is being introduced into clinical practice with caution.

Argatroban is another oral DTI which is an effective alterative to heparin when the latter is contraindicated following an episode of heparin-induced thrombocytopenia (HIT). It has recently been granted a licence in the UK for this indication. Bivalirudin, a derivative of hirudin, originally obtained from the medicinal leech, *Hirudo medicinalis*, is a parenteral DTI which has been shown to be effective for the treatment of acute coronary syndromes. It is also an alternative to heparin for patients with acute coronary syndromes who have HIT.

This central role of thrombin is also the rationale for the intensive research that is now being performed to refine the thrombin generation assay and apply it to both bleeding and thrombotic clinical pathologies.

Coagulation inhibitors are essential to prevent excessive thrombin formation and thrombosis

Three systems of naturally occurring coagulation inhibitors have been identified (Fig. 7.5 and Table 7.3).

Table 7.3

Properties of coagulation inhibitors

Inhibitor (synonym)	Molecular weight	Plasma concentration (mg/dL)
Antithrombin (antithrombin III)	65,000	18–30
Protein C	56,000	0.4
Protein S	69,000	2.5
Tissue factor pathway inhibitor, TFPI	32,000	0.1
(lipoprotein-associated coagulation inhibitor, LACI)		

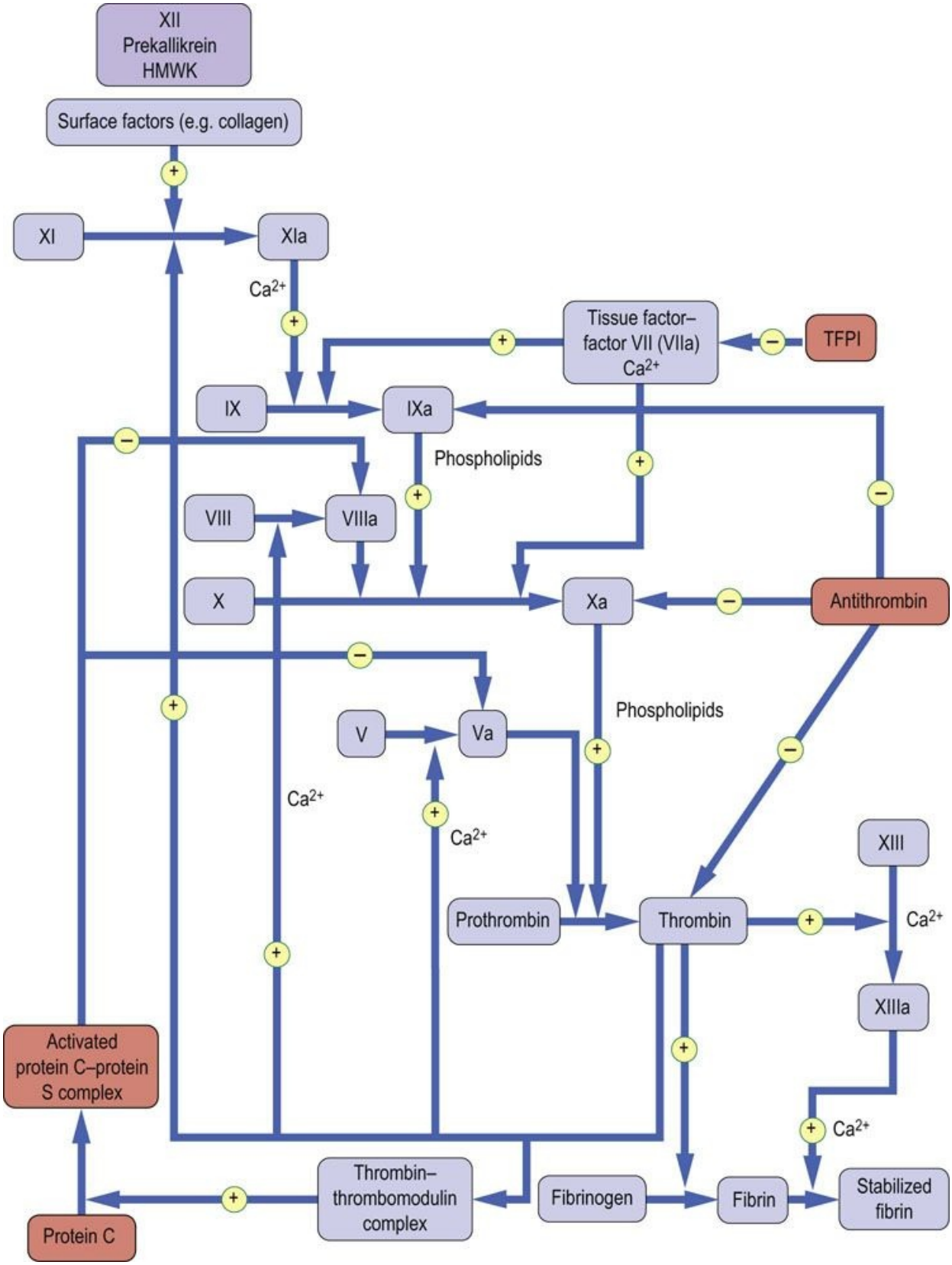


FIG. 7.5 Sites of action of blood coagulation inhibitors.

Antithrombin, protein C and protein S, and tissue factor pathway inhibitor (TFPI).

■ **Antithrombin:** this is a protein synthesized in the liver. Its activity is catalyzed by the antithrombotic drug heparin (unfractionated and low-molecular-weight heparins), and by heparin-like endogenous glycosaminoglycans (GAGs) that are present on the surface of vascular endothelial cells. It inactivates not only thrombin, but also factors IXa and Xa (Fig. 7.5). As a result, congenital antithrombin deficiency results in an increased risk of venous thromboembolism.
■ **Heparins** are referred to as indirect Xa inhibitors due to their augmentation of antithrombin activity. Heparins are used both in the treatment and prevention of acute venous thrombosis, usually in the form of a low-molecular-weight heparin (LMWH), *e.g.* enoxaparin or dalteparin. They are usually replaced by oral anticoagulants such as warfarin for longer-term anticoagulation. LMWHs also have a role in the management of acute arterial thrombosis as it pertains to acute coronary syndromes, *e.g.* enoxaparin or fondaparinux.
■ **Direct Xa inhibitors** have now been developed as anticoagulants. Rivaroxaban has been shown to be an effective alternative to warfarin for the treatment and prevention of venous thrombosis and for the prevention of stroke in patients with atrial fibrillation. Apixaban is licenced for the prevention of venous thrombosis in patients undergoing total knee or hip arthroplasties. Trials have also shown it to be effective in preventing thromboembolic stroke in patients with atrial fibrillation. Both medications are in an oral form and no monitoring is required of their anticoagulation effect. However, there is currently no effective antidote.
■ **Protein C and its cofactor, protein S:** these are vitamin K-dependent proteins, synthesized in the liver. When thrombin is generated, it binds to thrombomodulin (molecular weight 74 kDa), which is present on the surface of vascular endothelial cells. The thrombin–thrombomodulin complex activates protein C, which forms a complex with its cofactor, protein S. This complex selectively degrades factors Va and VIIIa by limited proteolysis (Fig. 7.5). Hence, this pathway forms a negative feedback upon thrombin generation. Congenital deficiencies of protein C or protein S result in an increased risk of venous thromboembolism. A further cause of increased risk of venous thromboembolism is a mutation in coagulation factor V (factor V Leiden), which confers resistance to its inactivation by activated protein C. This mutation is common, occurring in approximately 5% of the population in Western countries.
■ **Tissue factor pathway inhibitor (TFPI):** this protein is synthesized in

endothelium and the liver; it circulates bound to lipoproteins. It inhibits the tissue factor–VIIa complex (Fig. 7.5). However, deficiency of TFPI does not appear to increase the risk of thrombosis.

Clinical boxA 40-year-old man with pain and leg swelling: antithrombin Deficiency

A 40-year-old man was admitted from the emergency room of his local hospital because of acute pain and swelling of his left leg 10 days after recent major surgery. Ultrasound imaging of the leg confirmed occlusion of the left femoral vein by thrombus.

Comment.

He was prescribed anticoagulant therapy with low-molecular-weight heparin at standard doses. The patient volunteered a strong family history of ‘clots on the legs’ at a young age. He was commenced on warfarin and LMWH, the latter being discontinued when his INR was >2. He was followed up at the anticoagulation and specialist thrombophilia clinic for long-term management.

Fibrinolysis

Fibrinolytic system acts to limit excessive formation of fibrin through plasmin-mediated fibrinolysis

The coagulation system acts to form fibrin; the fibrinolytic system acts to limit excessive formation of fibrin (both intra-and extravascular) through plasmin-mediated fibrinolysis. Circulating plasminogen binds to fibrin via lysine-binding sites; it is converted to active plasmin by plasminogen activators. Tissue-type plasminogen activator (tPA) is synthesized by endothelial cells; it normally circulates in plasma in low basal concentrations (5 ng/mL), but is released into plasma by stimuli that include venous occlusion, exercise, and epinephrine. Together with plasminogen, it binds strongly to fibrin, which stimulates its activity (the K_m for plasminogen decreases from 65 to 0.15 µmol/L in the presence of fibrin), thereby localizing plasmin activity to fibrin deposits.

Clinical test boxMeasurement of fibrin D-dimer in the diagnosis of suspected deep vein thrombosis

Fibrin D-dimer (a degradation product of crosslinked fibrin and a marker of fibrin turnover) is normally present in blood at concentrations of <250 µg/L. In **deep vein thrombosis of the leg (DVT)**, deposition of a large mass of crosslinked fibrin within the leg veins, followed by partial lysis by the body's fibrinolytic system, increases fibrin turnover and blood D-dimer levels are elevated. Many patients attend accident and emergency departments with a swollen and/or painful leg, which may be due to a DVT.

Rapid immunoassays for blood D-dimer can be performed in the emergency department and are now widely used as an adjunct to clinical diagnosis. About one-third of patients with clinically suspected DVT have normal D-dimer levels, which in combination with a low clinical score usually excludes the diagnosis and may allow early discharge of such patients without the need for further

investigation or treatment. In patients with raised D-dimer levels, heparin treatment is started and imaging of the leg performed (usually by ultrasound) to confirm the presence and extent of a DVT.

Clinical boxAntithrombotic treatment in acute coronary syndrome

Occlusion of a coronary artery by thrombus causes the features of acute coronary syndrome which include electrocardiographic and biochemical changes. **Myocardial infarction** refers to permanent death of that part of the heart muscle which is supplied by the artery. In acute coronary syndromes, including myocardial infarction, the patient typically experiences **severe chest pain**.

Aspirin and heparin are usually given in acute myocardial infarction and other acute coronary syndromes, to inhibit the platelet and fibrin components of the developing coronary artery thrombus. Some patients may require the addition of clopidogrel and/or GPIIb-IIIa inhibitors.

Many patients with evolving acute myocardial infarction are candidates for **thrombolytic treatment** with a plasminogen activator drug, given intravenously. Prompt thrombolysis dissolves the coronary artery thrombus, reduces the size of the infarct, and reduces the risk of complications, including death and heart failure. However, in recent years direct removal of the thrombus **(percutaneous coronary intervention (PCI))** is given instead of thrombolytic therapy as it appears to give favorable outcomes to thrombolysis and does not increase the risk of bleeding, for example into the brain. Patients undergoing PCI should in addition receive a GPIIb-IIIa inhibitor.

Plasmin inhibitors prevent excessive fibrinolytic activity

Excessive tPA activity in plasma is normally prevented by an excess of its major inhibitor, plasminogen activator inhibitor type 1 (PAI-1), which is synthesized by both endothelial cells and hepatocytes. Urinary-type plasminogen activator (uPA) circulates in plasma both as an active single-chain precursor form (scuPA, pro-urokinase) and as a more active two-chain form (tcuPA, urokinase). One activator of scuPA is surface-activated coagulation factor XII, which therefore links the coagulation and fibrinolytic systems. The major components of the fibrinolytic system are illustrated in Table 7.4 and Figure 7.6. Excessive formation of plasmin is normally prevented by:

Table 7.4

The components of fibrinolytic system

Component (synonym)	Molecular weight (Da)	Plasma concentration (mg/dL)
Plasminogen	92,000	0.2
Tissue-type plasminogen activator, tPA	65,000	5 (basal)
Urinary-type plasminogen activator type 1, uPA	54,000	20
Plasminogen activator inhibitor type 1, PAI-1	48,000	200
Antiplasmin (α_2-antiplasmin)	70,000	700

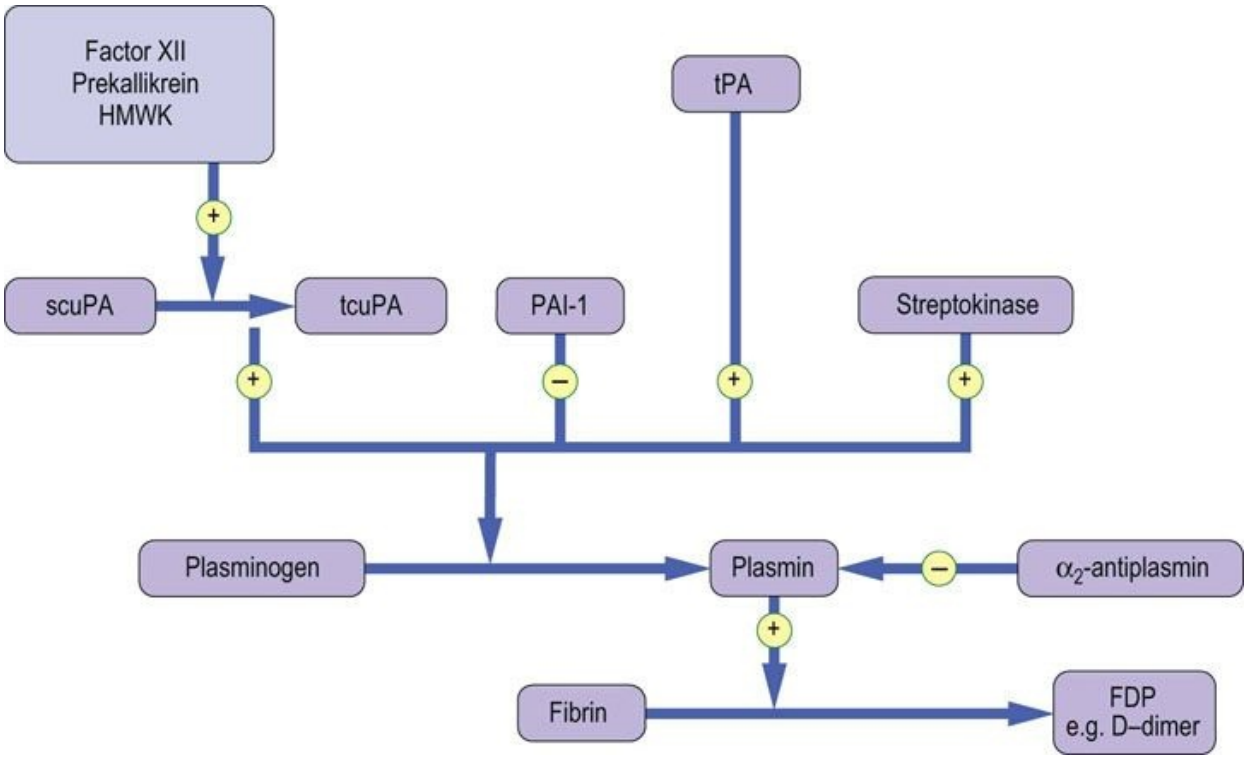


FIG. 7.6 The fibrinolytic system.
Plasminogen can be activated to plasmin by uPA (urokinase), tPA or streptokinase. uPA and tPA are inhibited by plasminogen activator inhibitor type 1 (PAI-1). Plasmin is inhibited by antiplasmin. Plasmin degrades fibrin to fibrin degradation products (FDPs). HMWK, high-molecular-weight kininogen; scuPA, pro-urokinase; tcuPA, two-chain urokinase; tPA, tissue-type plasminogen activator. Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012.

- binding of 50% of plasminogen to histidine-rich glycoprotein (HRG), and
- rapid inactivation of free plasmin by its major inhibitor, α_2-antiplasmin.

The physiologic importance of PAI-1 and α_2-antiplasmin is illustrated by the increased bleeding tendency that is associated with the rare cases of their congenital deficiencies (Table 7.1); the excessive plasma plasmin activity that results from these deficiencies has the effect of lysing hemostatic plugs.

Summary

■ Hemostasis constitutes a number of processes which guard the body against blood loss.
■ Injury to the blood vessel wall sets in motion complex phenomena which involve blood platelets (activation, adhesion, aggregation) and a cascade of coagulation factors, classified into intrinsic, extrinsic and final common pathways.
■ The integrity of these three systems may be tested by simple laboratory tests. Global coagulation assays such as thrombin generation and thromboelastography, currently used in research, may be more effective in assessing an individual's coagulation phenotype.
■ Deficiencies of factors participating in the coagulation cascade, and/or disordered platelet function, result in bleeding disorders.
■ Eventually, blood clots are degraded by the fibrinolytic system. The process of fibrinolysis prevents thrombotic phenomena and there is normally a balance between hemostasis and thrombosis.
■ Aspirin and heparin are used in patients with acute myocardial infarction or other acute coronary syndromes.
■ Aspirin (or other antiplatelet agents) are also used to reduce risk of recurrent myocardial infarction and stroke.
■ Anticoagulant drugs, *e.g.* heparin, warfarin, or rivaroxaban, are used in the treatment of acute venous thrombosis or embolism.
■ Anticoagulant drugs, *e.g.* warfarin, dabigatran and rivaroxaban, are used long term to prevent thromboembolism arising from the heart (atrial fibrillation, heart valve prostheses).

Active learning

1. When a patient presents with excessive bleeding from multiple sites, what laboratory tests are available to identify the likely cause of their hemostatic defect?
2. When a patient presents with a painful swollen leg, possibly due to acute deep venous thrombosis (DVT), what laboratory tests can be performed to help the clinician to:
   ■ Establish or exclude this diagnosis?
   ■ Monitor anticoagulant treatment, after the diagnosis has been

confirmed?
3. When a patient presents with acute coronary artery thrombosis (evolving to myocardial infarction), what antithrombotic drugs should be urgently considered to reduce the risk of complications?

Further reading

Ozaki, Y. State of the art 2011. editor. *J Thromb Haemostas*. 2011; 9(suppl s1):1–395.

Mannucci, PM, Levi, M. Prevention and treatment of major blood loss. *N Engl J Med*. 2007; 356:2301–2311.

Weitz, JI, *et al*. New antithrombotic drugs. *Chest*. 2012; 141(2):e120S–e151S.

Websites and downloads

Anaesthesia, UK. Tests of coagulation. www.frca.co.uk/article.aspx?articleid=100101.

International Society on Thrombosis and Haemostasis. www.isth.org.

Walker HK, Hall WD, Hurst JW, eds. Clinical Methods: The History, Physical, and Laboratory Examinations. ed 3. Butterworths, Boston, 1990. www.ncbi.nlm.nih.gov/books/NBK265

Wright, IS, The nomenclature of blood clotting factors. *Can Med Assoc J* 1962; 86:373–37. www.ncbi.nlm.nih.gov/pmc/articles/PMC1848865/?page=1

CHAPTER 8

Membranes and Transport

Masatomo Maeda

Learning objectives

After reading this chapter you should be able to:

■ Describe differences between passive and active carrier-mediated transport systems.

■ Describe the basic features of membrane channels and pores.

■ Give several specific examples of ion and substrate transport systems, including coupled transport systems.

■ Describe several characteristic diseases resulting from defects in membrane transport.

Introduction

Biomembranes are not rigid or impermeable, but highly mobile and dynamic structures

The plasma membrane is the gatekeeper of the cell. It controls not only the access of inorganic ions, vitamins and nutrients but also the entry of drugs and the exit of waste products. Integral transmembrane proteins have important roles in transporting these molecules through the membrane and often maintain concentration gradients across the membranes. K^+, Na^+ and Ca^{2+} concentrations in the cytoplasm are maintained at $\sim$140, 10, and 10^{-4} mmol/L, respectively, by transport proteins, whereas those outside (in the blood) are $\sim$5, 145, and 1–2 mmol/L, respectively. The driving force for transport of ions and maintenance of ion gradients is directly or indirectly provided by ATP. The transport properties of membranes will be illustrated by several important examples.

Types of transport processes

Simple diffusion through the phospholipid bilayer

Some small, neutral molecules can traverse biomembranes by simple diffusion

Small, nonpolar molecules (such as O_2, CO_2, N_2) and uncharged polar molecules (such as urea, ethanol, and small organic acids) move through membranes by simple diffusion without the aid of membrane proteins (Table 8.1 and Fig. 8.1A). The direction of net movement of these species is always 'downhill', along the concentration gradient, from high to low concentration to establish equilibrium.

Table 8.1

Transport systems of biomembranes

Type		Example	Transport protein	Energy coupling	Specificity	Saturability	Rate (molecules/ transport protein/s)
Passive transport or diffusion	Simple diffusion		–	–	–	–	
	Facilitated diffusion		+	–	+	+	
	Transporter	GLUT-1–5					~10^2
	Channel	H_2O, Na^+, K^+, Ca^{2+}, Cl^-					10^7–10^8
Active transport	Primary	proton pumps	+	+	+	+	10^2–10^4
	Secondary	ABC transporters	+	+	+	+	10^0–10^{2*}
	Symporter	SGLT-1, 2, neutral amino acids					
	Antiporter	Cl^-/HCO_3^-, Na^+/Ca^{2+}, Na^+/H^+					
	Uniporter	Glutamate					

Transport systems are classified according to the role of transport proteins and energy coupling. Typical substrates for various types of transporters and channels are shown in the parentheses.

*The Cl^-/HCO_3^- antiporter seems to be an exception to secondary active transport systems, as its transport rate is high, at 10^5 molecules/transport protein/s.

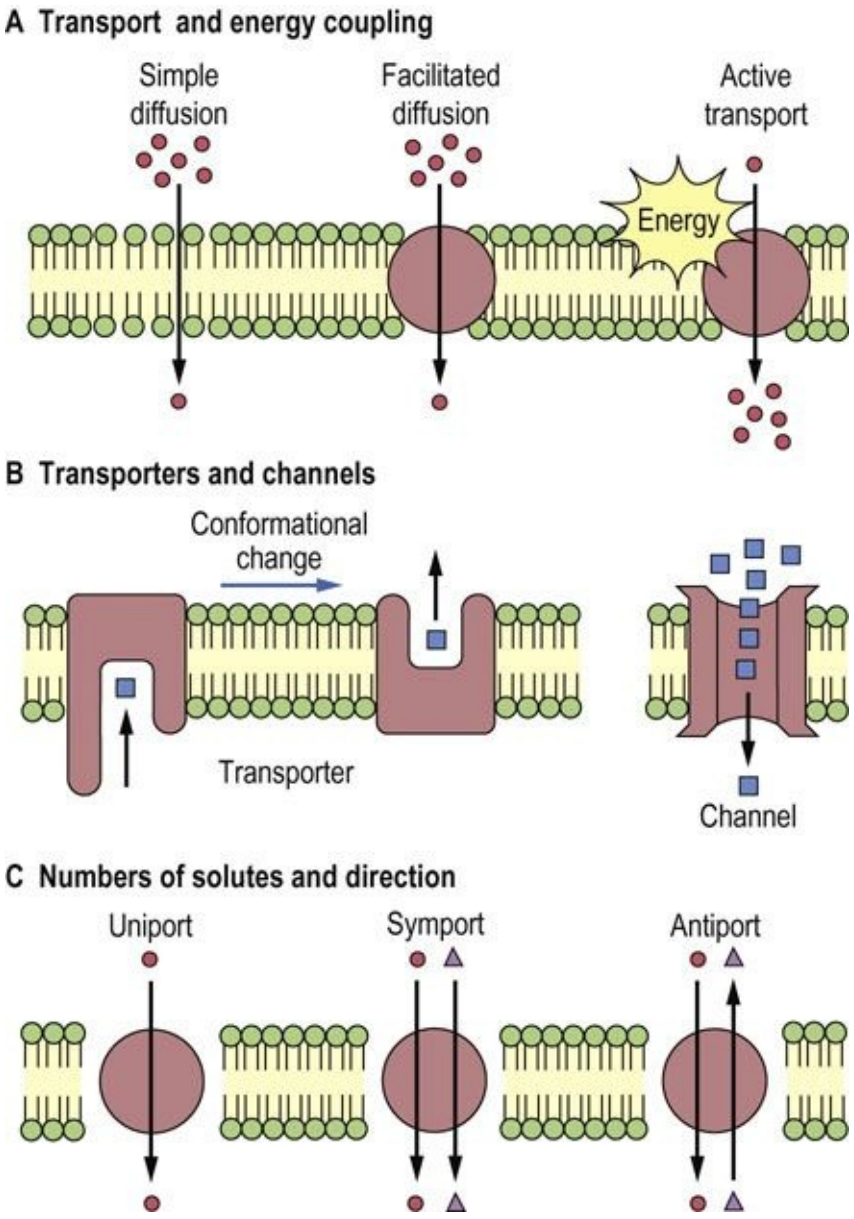


FIG. 8.1 Various models of solute movement across membranes.

The hydrophobicity of the molecules is an important requirement for simple diffusion across the membrane, as the interior of the phospholipid bilayer is hydrophobic. The rate of transport of these molecules is, in fact, closely related to their **partition coefficient** between oil and water.

Although water molecules can be transported by simple diffusion, channel proteins (see below) are believed to control the movement of water across most membranes, especially in the kidney for concentration of the urine. Mutation in a water channel protein gene (aquaporin-2) causes diuresis in patients with

nephrogenic diabetes insipidus, a disease characterized by excessive urination but without the hyper-glycemia characteristic of diabetes mellitus (see Chapter 23).

Transport mediated by membrane proteins

Membrane proteins are required for transport of larger molecules across biomembranes

Transport of larger, polar molecules, such as amino acids or sugars, into a cell requires the involvement of membrane proteins known as transporters, also called porters, permeases, translocases or carrier proteins. The term ‘carrier’ is also applied to ionophores, which move passively across the membrane together with the bound ion (Fig. 8.2). Transporters are as specific as enzymes for their substrates, and work by one of two mechanisms: **facilitated diffusion** or **active transport**. Facilitated diffusion catalyzes the movement of a substrate through a membrane down a concentration gradient and does not require energy. In contrast, active transport is a process in which substrates are transported uphill, against their concentration gradient. Active transport must be coupled to an energy-producing reaction (see Fig. 8.1A).

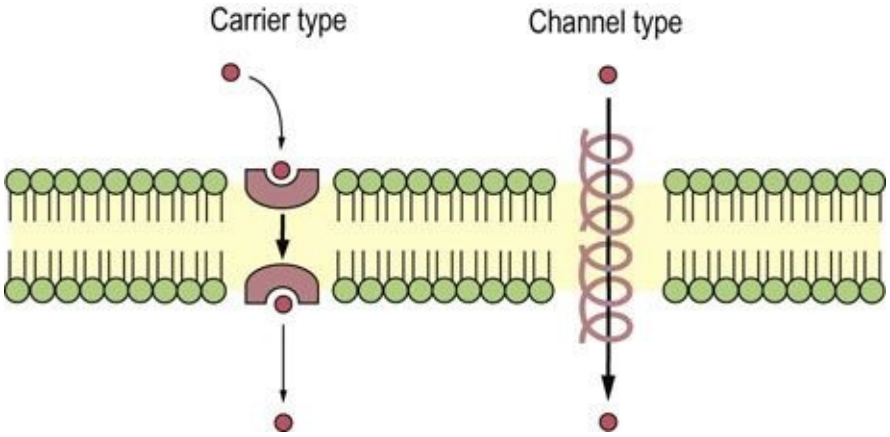


FIG. 8.2 Mobile ion carriers and channel-forming ionophores.
Ionophores permit net movement of ions only down their electrochemical gradients.

Advanced concept box

Antiobiotics and Membrane permeability

Antibiotics act as ionophores and increase the permeability of membranes to specific ions; bactericidal effects of ionophores are attributed to disturbance of the ion transport systems of bacterial membranes. Ionophores permit net movement of ions only down their electrochemical gradients. There are two classes of ionophores: mobile ion carriers (or caged carriers) and channel formers (Fig. 8.2).

Valinomycin is a typical example of a mobile ion carrier. It is a cyclic peptide with a lipophilic exterior and an ionic interior. It dissolves in the membrane and diffuses between the inner and outer surfaces. K^+ binds in the central core of valinomycin, and the complex diffuses across the membrane, releasing the K^+, gradually dissipating the K^+ gradient. The carrier-type ionophores, nigericin and monensin, exchange H^+ for Na^+ and K^+, respectively. Ionomycin and A23187 are Ca^{2+} ionophores.

The β-helical gramicidin A molecule, a linear peptide with 15 amino acid residues, forms a pore. The head-to-head dimer of gramicidin A makes a transmembrane channel that allows movement of monovalent cations (H^+, Na^+, and K^+).

Polyene antibiotics such as amphotericin B and nystatin exert their cytotoxic action by rendering the membrane of the target cell permeable to ions and small molecules. Formation of a sterol–polyene complex is essential for the cytotoxic function of these antibiotics, as they display a selective action against organisms in which the membranes contain sterols. Thus they are active against yeasts, a wide variety of fungi, and other eukaryotic cells, but have no effect on bacteria. Because their affinity toward ergosterol, a fungal membrane component, is higher than that for cholesterol, these antibiotics have been used for the treatment of topical infections of fungal origin.

Saturability and specificity are important characteristics of membrane transport systems

The rate of facilitated diffusion is generally much greater than that of simple diffusion: transport proteins catalyze the transport process. In contrast to simple diffusion, in which the rate of transport is directly proportional to the substrate concentration, facilitated diffusion is a saturable process, having a maximum transport rate, T_{max} (Fig. 8.3). When the concentration of extracellular molecules (transport substrates) becomes very high, the T_{max} is achieved by saturation of the transport proteins with substrate. The kinetics of facilitated diffusion for substrates can be described by the same equations that are used for enzyme catalysis (e.g. Michaelis–Menten and Lineweaver–Burk type equations) (see Chapter 6):

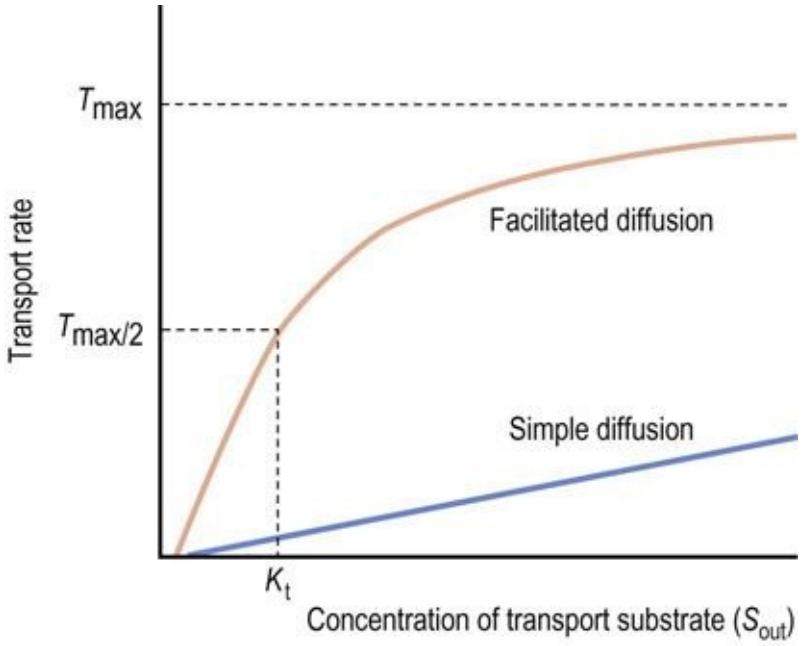


FIG. 8.3 Comparison of the transport kinetics of facilitated diffusion and simple diffusion.
The rate of transport of substrate is plotted against the concentration of substrate in the extracellular medium. In common with enzyme catalysis, transporter-catalyzed uptake has a maximum transport rate, T_{max} (saturable). K_t is the concentration at which the rate of substrate uptake is half-maximal. For simple diffusion, the transport rate is slower and directly proportional to substrate concentration.

$$S_{out} + \text{transporter} \overset{K_t}{\rightleftharpoons} (S \cdot \text{transporter complex}) \rightharpoonup S_{in}$$

where K_t is the dissociation constant of the substrate–transporter complex, and S_{out} is the concentration of transport substrate. Then the transport rate, t, can be calculated as:

$$t = \frac{T_{max}}{1 + \frac{K_t}{S_{out}}}$$

where the K_t is the concentration that gives the half-maximal transport rate. The K_t for a transporter is conceptually the same as the K_m for an enzyme (Chapter 6).

The transport process is usually highly specific: each transporter transports only a single species of molecules or structurally related compounds. The red blood cell GLUT-1 transporter has a high affinity for D-glucose, but 10–20 times lower affinity for the related sugars, D-mannose and D-galactose. The enantiomer L-glucose is not transported; its K_t is more than 1000 times higher than that of the D-form.

Clinical box Cystinosis

An 18-month-old child presented with polyuria, failure to thrive and an episode of severe dehydration. Urine dipstick testing demonstrated glucosuria and proteinuria, with other biochemical analyses showing generalized aminoaciduria and phosphaturia.

Comment.

This is a classical presentation of infantile cystinosis, resulting from accumulation of cystine in lysosomes because of a defect in the lysosomal transport protein, cystinosine. Cystine is poorly soluble, and crystalline precipitates form in cells throughout the body. In vitro experiments with cystine loading have shown that renal proximal tubular cells become ATP-depleted, resulting in impairment of ATP-dependent ion pumps with consequent electrolyte imbalances and metabolite losses.

Treatment with cysteamine increases the transport of cystine

from lysosomes, delaying the decline in renal function. Cysteamine is a weak base; it forms a mixed disulfide with cysteine, which is then secreted through a cationic amino acid transporter. If untreated, renal failure occurs by 6–12 years of age. Unfortunately, there is further accumulation of cystine in the central nervous system, despite therapy, with long-term neurological damage.

Clinical box Hartnup disease

A three-year old had been on holiday to Spain and developed pellagra-like skin changes on his face, neck, forearms and dorsal aspects of his hands and legs. His skin became scaly, rough and hyperpigmented. The child was brought to the GP complaining of headaches and weakness. Urinalysis demonstrated gross hyperaminoaciduria of neutral monoamino-monocarboxylic acids (i.e. alanine, serine, threonine, asparagine, glutamine, valine, leucine, isoleucine, phenylalanine, tyrosine, tryptophan, histidine and citrulline).

Comment.

These amino acids share a common transporter which is expressed only on the luminal border of epithelial cells in the renal tubules and intestinal epithelium. The pellagra-like dermatitis (see Chapter 11) and neurological involvement resemble nutritional niacin deficiency. The reduced tryptophan intake results in reduced nicotinamide production. The disease is treated with oral nicotinamide and application of sun-blocking agents to exposed areas.

Characteristics of glucose transporters (uniporters)

Glucose transporters catalyze downhill transport of glucose into and out of cells

Glucose transporters are essential for facilitated diffusion of glucose into cells. The GLUT family of glucose transporters includes GLUT-1 to GLUT-5 (Table 8.2) and others. They are transmembrane proteins similar in size, all having about 500 amino acid residues and 12 transmembrane helices. GLUT-1, in red blood cells, has a K_m of ~2 mmol/L. The GLUT-1 transporter operates at about 40% of T_{max} under fasting conditions (blood glucose concentration of 5 mmol/L; 90 mg/dL); this level of activity is sufficient to meet the needs of the red cell (Chapter 12). In contrast, pancreatic islet β-cells express GLUT-2, with a K_m of more than 10 mmol/L (180 mg/dL). In response to the intake of food and the resulting increase in blood glucose concentration, GLUT-2 molecules respond by increasing the uptake of glucose into β-cells, stimulating insulin secretion (Chapter 21). Cells in insulin-sensitive tissues such as muscle and adipose have GLUT-4. Insulin stimulates translocation of GLUT-4 from intracellular vesicles to the plasma membrane, facilitating glucose uptake during meals.

Table 8.2

Classification of glucose transporters

Transporter	K_t for D-glucose transport (mmol)	Substrate	Major sites of expression
Facilitated diffusion (uniporter) (passive transport)			
GLUT-1	1–2	Glucose, galactose, mannose	Erythrocyte, blood–tissue barriers
GLUT-2	15–20	Glucose, fructose	Liver, intestine, kidney, pancreatic β-cells, brain
GLUT-3	1.8*	Glucose	Ubiquitous
GLUT-4	5	Glucose	Skeletal and cardiac muscles, adipose tissues
GLUT-5	6–11**	Fructose	Intestine
Na^+-coupled symporter (active transport)			
SGLT-1	0.35	Glucose (2Na^+/1 glucose), galactose	Intestine, kidney
SGLT-2	1.6	Glucose (1Na^+/1 glucose)	Kidney

K_m values are determined from the uptake of 2-deoxy-D-glucose (*), a nonmetabolizable analogue of glucose, and fructose (**).

Clinical box Defective glucose transport across the blood–brain barrier as a cause of seizures, and

developmental delay

A male infant at the age of 3 months suffered from recurrent seizures. His cerebrospinal fluid (CSF) glucose concentrations were low (0.9–1.9 mmol/L; 16–34 mg/dL), and the ratio of CSF to blood glucose ranged from 0.19 to 0.33; the normal value is 0.65.

The potential causes of low CSF glucose concentrations, such as bacterial meningitis, subarachnoid hemorrhage, and hypoglycemia, were not present, and high CSF lactate values would be found in all these conditions except hypoglycemia. In contrast, the CSF lactate concentrations were consistently low (0.3–0.4 mmol/L; 3–4 mg/dL) compared with the normal value (<2.2 mmol/L; <20 mg/dL). These findings suggested a defect in transport of glucose from the blood to the brain.

Comment.

Assuming that the activity of GLUT-1 glucose transporter in the erythrocyte reflects that of the brain microvessels, a transport assay using his erythrocytes was carried out. The T_{max} for uptake of glucose by the patient's erythrocytes was 60% of the mean normal value, suggesting a heterozygous defect. A ketogenic diet (a high-fat, low-protein, low-carbohydrate diet) was started, since the brain can use ketone bodies as oxidizable fuel sources, and the entry of ketone bodies into the brain is not dependent on the glucose transporter system. The patient stopped having seizures within 4 days after beginning the diet.

Transport by channels and pores

Membrane channels or pores are open, less selective conduits for transport of ions, metabolites and even proteins across biomembranes

Channels are often pictured as tunnels across the membrane, in which binding sites for substrates (ions) are accessible from either side of the membrane at the

same time (see Fig. 8.1B). Conformational changes are not required for the translocation of substrates entering from one side of the membrane to exit on the other side. However, voltage changes and ligand binding induce conformational changes in channel structure that have the effect of opening or closing the channels – processes known as **voltage or ligand 'gating'**. Movement of molecules through channels is fast in comparison with the rates achieved by transporters (see Table 8.1).

The terms 'channel' and 'pore' are sometimes used interchangeably. However, 'pore' is used most frequently to describe more open, somewhat nonselective structures that discriminate between substrates, *e.g.* peptides or proteins, on the basis of size. The term 'channel' is usually applied to more specific ion channels.

Three examples of pores important for cellular physiology

The **gap junction** between endothelial, muscle, and neuronal cells is a cluster of small pores, in which two cylinders of six **connexin** subunits in the plasma membranes join each other to form a pore about 1.2–2.0 nm (12–20 Å) in diameter. Molecules smaller than about 1 kDa can pass between cells through these gap junctions. Such cell–cell interchange is important for physiologic communication or coupling, for example in the concerted contraction of uterine muscle during labor and delivery. Mutations of the genes encoding connexin 26 and connexin 32 cause deafness and Charcot–Marie–Tooth disease, respectively.

Nuclear pores have a radius of about 9.0 nm (90 Å) through which larger proteins and nucleic acids enter and leave the nucleus.

A third class of pores is important for protein sorting. Mitochondrial proteins encoded by nuclear genes are transported to this organelle through pores in the outer mitochondrial membrane. Nascent polypeptide chains of secretory proteins and plasma membrane proteins also pass through pores in the endoplasmic reticulum membrane during biosynthesis of the peptide chain.

Active transport

Primary active transport systems use ATP directly to drive transport; secondary active transport uses an electrochemical gradient of Na^+ or H^+ ions, or a membrane potential produced by primary active transport processes

ATP is a high-energy product of metabolism and is often described as the ‘energy currency’ of the cell (Chapter 9). The phosphoanhydride bond of ATP releases free energy when it is hydrolyzed to produce adenosine diphosphate (ADP) and inorganic phosphate. Such energy is used for biosynthesis, cell movement, and uphill transport of molecules against concentration gradients. Primary active transport systems use ATP directly to drive transport; secondary active transport uses an electrochemical gradient of Na^+ or H^+ ions, or a membrane potential produced by primary active transport processes. Sugars and amino acids are generally transported into cells by secondary active transport systems.

Primary active transport systems use ATP to drive ion pumps (ion transporting ATPases or pump ATPases)

The pump ATPases are classified into four groups (Table 8.3). Coupling factor ATPases (F-ATPases) in mitochondrial, chloroplast, and bacterial membranes hydrolyze ATP and transport hydrogen ions (H^+). As discussed in detail in the next chapter, the mitochondrial **F-ATPase** works in the backward direction, synthesizing ATP from ADP and phosphate as protons move down an electrochemical (concentration and charge) gradient generated across the inner mitochondrial membrane during oxidative metabolism. The product, ATP, is released into the mitochondrial matrix, but is needed for biosynthetic reactions in the cytoplasm. ATP is transported to the cytoplasm through an **ATP-ADP translocase** in the mitochondrial inner membrane. This translocase is an example of an **antiport** system (see Fig. 8.1C); it allows one molecule of ADP to enter only if one molecule of ATP exits simultaneously.

Table 8.3

Primary active transporters in eukaryotic cells

Group	Member	Location	Substrate(s)	Functions
F-ATPase (coupling factor)	H^+-ATPase	Mitochondrial inner membrane	H^+	ATP synthesis driven by electrochemical gradient of H^+
V-ATPase (vacuolar)	H^+-ATPase	Cytoplasmic vesicles (lysosome, secretory granules), plasma membranes (ruffled border of osteoclast, kidney epithelial cell)	H^+	Activation of lysosomal enzymes, accumulation of neurotransmitters, turnover of bone, acidification of urine
P-ATPase (phosphorylation)	Na^+/K^+-ATPase	Plasma membranes (ubiquitous, but abundant in kidney and heart)	Na^+ and K^+	Generation of electrochemical gradient of Na^+ and K^+
	H^+/K^+-ATPase	Stomach (parietal cell in gastric gland)	H^+ and K^+	Acidification of stomach lumen
	Ca^{2+}-ATPase	Sarcoplasmic reticulum and endoplasmic reticulum	Ca^{2+}	Ca^{2+} sequestration into sarcoplasmic (endoplasmic) reticulum
	Ca^{2+}-ATPase	Plasma membrane	Ca^{2+}	Ca^{2+} excretion to outside of the cell
	Cu^{2+}-ATPase	Plasma membrane and cytoplasmic vesicles	Cu^{2+}	Cu^{2+} absorption from intestine and excretion from liver
ABC (ATP-binding cassette) transporter	P-glycoprotein	Plasma membrane	Various drugs	Excretion of harmful substances, multidrug resistance for anticancer drugs
	MRP	Plasma membrane	Glutathione conjugate	Detoxification, multidrug resistance
	CFTR*	Plasma membrane	Cl^-	c-AMP-dependent chloride channel, regulation of other channels
	TAP	Endoplasmic reticulum	Peptide	Presentation of peptides for immune response

Various examples of primary active transporters (ATP-powered pump ATPases) are listed, together with their location.

*Some ABC transporters function as channels or channel regulators. MRP, multidrug resistance-associated protein; CFTR, cystic fibrosis transmembrane conductance regulator; TAP, transporter associated with antigen processing.

Cytoplasmic vesicles, such as lysosomes, endosomes, and secretory granules, are acidified by a **V-type (vacuolar) H^+-ATPase** in their membranes. Acidification by this V-ATPase is important for the activity of lysosomal enzymes that have acidic pH optima, and for the accumulation of drugs and neurotransmitters in secretory granules. The V-ATPase also acidifies the extracellular environments of osteoclasts and renal epithelial cells. Defects in the osteoclast plasma membrane V-ATPase result in osteopetrosis (increased bone density), while mutation of the ATPase in collecting ducts of the kidney causes renal tubular acidosis. F-and V-type ATPases are structurally similar, and seem to be derived from a common ancestor. The ATP-binding catalytic subunit and the subunit forming the H^+ pathway are conserved between these ATPases.

P-ATPases form phosphorylated intermediates that drive ion translocation: the 'P' refers to the phosphorylation. These transporters have an active-site aspartate residue that is reversibly phosphorylated by ATP during the transport process. The P-type Na^+/K^+-ATPase in various tissues and the Ca^{2+}-ATPase in the sarcoplasmic reticulum have important roles in maintaining cellular ion gradients. Na^+/K^+-ATPases also create an **electrochemical gradient** of Na^+ that produces the driving force for uptake of nutrients from the intestine (below). The

discharge of this electrochemical gradient is also fundamental to the process of nerve transmission. Mutations of P-ATPase genes cause Brody cardiomyopathy (Ca^{2+}-ATPase), familial hemiplegic migraine type 2 (Na^{+}/K^{+}-ATPase), and Menkes and Wilson's diseases (Cu^{2+}-ATPases).

Clinical box Menkes and wilson's diseases

X-linked **Menkes disease** is a lethal disorder that occurs in 1 in 100,000 newborn infants and is characterized by abnormal and hypopigmented hair, a characteristic facies, cerebral degeneration, connective tissue and vascular defects, and death by the age of 3 years. A copper-transporting P-ATPase that is expressed in all tissues except liver is defective in this disease (Table 8.3). In patients with Menkes disease, copper enters the intestinal cells, but is not transported further, resulting in severe copper deficiency. Subcutaneous administration of a copper histidine complex may be an effective treatment if started early.

The gene for **Wilson's disease** also encodes a copper-transporting P-ATPase and is 60% identical with that of the Menkes gene. It is expressed in liver, kidney, and placenta. Wilson's disease occurs in 1 in 35,000–100,000 newborns and is characterized by failure to incorporate copper into ceruloplasmin in the liver and failure to excrete copper from the liver into bile, resulting in toxic accumulation of copper in the liver and also in the kidney, brain, and cornea. Liver cirrhosis, progressive neurologic damage, or both, occur during childhood to early adulthood. Chelating agents such as penicillamine are used for treatment of patients with this disease. Oral zinc treatment may be useful for decreasing the absorption of dietary copper. Copper is an essential trace metal and an integral component of many enzymes. However, it is toxic in excess, because it binds to proteins and nucleic acids, enhances the generation of free radicals, and catalyzes oxidation of lipids and proteins in membranes.

The **ATP-binding cassette (ABC) transporters** comprise the fourth active transporter family. 'ABC' is the abbreviation for 'ATP-binding cassette', referring to an ATP-binding region in the transporter (see Table 8.3). P-glycoprotein ('P' = permeability) and **MRP (multidrug resistance-associated protein)**, which have a physiologic role in excretion of toxic metabolites and xenobiotics, contribute to resistance of cancer cells to chemotherapy. TAP transporters, a class of ABC transporters associated with **A**ntigen **P**resentation, are required for initiating the immune response against foreign proteins; they mediate antigen peptide transport from the cytosol into endoplasmic reticulum. Some ABC transporters are present in peroxisomal membrane where they appear to be involved in the transport of peroxisomal enzymes necessary for oxidation of very long-chain fatty acids. Defects of ABC transporters are associated with a number of diseases (see box on next page).

Advanced concept box ABC transporter diseases

Human genome data suggests that there are about 50 genes for ABC transporters. An unusually wide range of diseases are caused by defects in ABC transporters, including Tangier disease, Stargardt disease, progressive intrahepatic cholestasis, Dubin–Johnson syndrome, pseudoxanthoma elasticum, familial persistent hyperinsulinemic hypoglycemia of infancy (PHHI), adrenoleukodystrophy, Zellweger syndrome, sitosterolemia and cystic fibrosis.

Cystic fibrosis (CF) is the most common potentially lethal autosomal recessive disease of Caucasian populations, affecting 1 in 2500 newborns. CF is usually manifested as exocrine pancreatic insufficiency, an increase in the concentration of chloride ions (Cl^-) in sweat, male infertility, and airway disease, which is the major cause of morbidity and mortality. The pancreatic and lung pathology results from the increased viscosity of secreted fluids (mucoviscoidosis). CF is caused by mutations in the gene CFTR (cystic fibrosis transmembrane conductance regulator) which encodes a Cl^- channel. ATP binding to CFTR is required for channel opening. The lack of this channel activity in epithelia of CF patients affects both ion and water secretion.

Uniport, symport, and antiport are examples of secondary active transport

Transport processes may be classified into three general types: uniport (monoport), symport (cotransport) and antiport (countertransport) (see Fig. 8.1). Transport substrates move in the same direction during symport, and in opposite directions during antiport. Uniport of charged substrates may be electrophoretically driven by the membrane potential of the cell. The movement of one substrate uphill, against its concentration gradient, can be driven by antiport of another substrate (usually a cation such as Na^+ or H^+) down a gradient. The proteins participating in these transport systems are termed uniporters, symporters, and antiporters, respectively (see Table 8.1). Some examples are presented below.

Advanced concept box
Electrochemical gradients

The permeability of most nonelectrolytes through membranes can be analyzed by assuming that the rate-limiting step is the diffusion within the lipid bilayer. Their permeability across a phospholipid bilayer is experimentally shown to be a function of the partition coefficient between aqueous and organic solvents. The relative rate of simple diffusion of a molecule across the membrane is therefore proportional to the concentration gradient across the bilayer and to the hydrophobicity of the molecule.

For charged molecules and ions, transport across the membrane must be facilitated by a transporter or channel, and is driven by the electrochemical gradient, a combination of the concentration gradient (chemical potential) and the voltage gradient across the membrane (electric potential). These forces may act in the same direction or in opposite directions.

In the case of Na^+ ions, the concentration difference between outside (145 mmol/L) and inside (12 mmol/L) the cell is about a factor of 10, being maintained by the Na^+/K^+-ATPase. **The Na^+/K^+-ATPase** is electrogenic,

pumping out three Na^+ and pumping in two K^+ ions, generating an inside-negative membrane potential. K^+ leaks out through K^+ channels, down its concentration gradient (140 mmol/L to 5 mmol/L), further increasing the electric potential. The concentration gradient of Na^+ ions and the electric potential (inside negative) power the import and export of other molecules with Na^+ against their concentration gradient by symporters and antiporters, respectively.

Examples of transport systems and their coupling

Ca^{2+} transport and mobilization in muscle

Membrane depolarization opens up voltage dependent ion channels at the neuromuscular junction

Striated muscle (skeletal and cardiac) is composed of bundles of muscle cells (Chapter 20). Each cell is packed with bundles of actin and myosin filaments (myofibrils) that produce contraction. During muscle contraction, nerves at the neuromuscular junction stimulate local depolarization of the membrane by opening voltage-dependent Na^+ channels. The depolarization spreads rapidly into invaginations of the plasma membrane called the transverse (T) tubules, which extend around the myofibrils (see Fig. 20.5).

Voltage-dependent Ca^{2+} channels (VDCCs) located in the T tubules of skeletal muscle change their conformation in response to membrane depolarization, and directly activate a Ca^{2+}-release channel in the sarcoplasmic reticulum membrane, a network of flattened tubules that surrounds each myofibril in the muscle cell cytoplasm. The escape of Ca^{2+} from the lumen (interior compartment) of the sarcoplasmic reticulum increases the cytoplasmic concentration of Ca^{2+} (depolarization-induced Ca^{2+} release) about 100-fold, from 10^{-4} mmol/L (0.0007 mg/L) to about 10^{-2} mmol/L (0.07 mg/dL), triggering ATP hydrolysis by myosin, which initiates muscle contraction. A Ca^{2+}-ATPase in the sarcoplasmic reticulum then hydrolyzes ATP to transport Ca^{2+} back out of the cytoplasm into the lumen of the sarcoplasmic reticulum, decreasing the cytoplasmic Ca^{2+} and allowing the muscle to relax (Fig. 8.4, left).

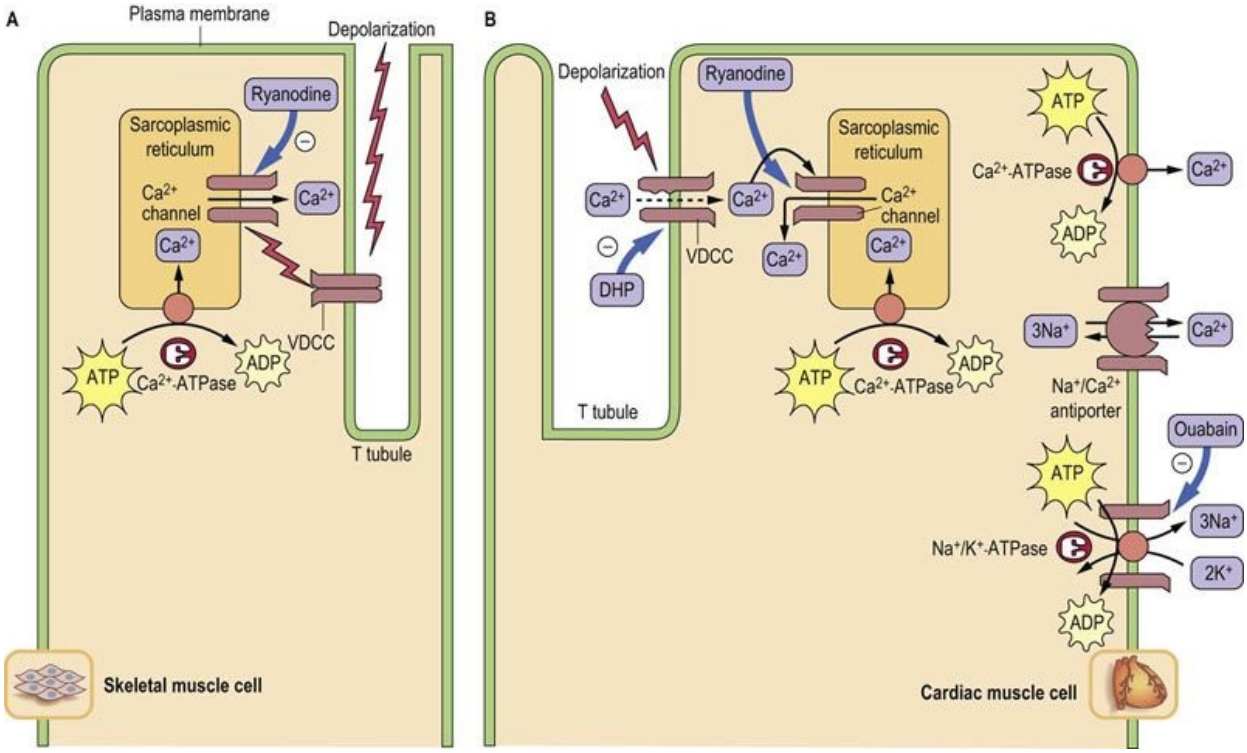


FIG. 8.4 Ca^{2+} movement in muscle contraction cycle.
Roles of transporters in Ca^{2+} movements in skeletal **(A)** and cardiac **(B)** muscle cells during contraction. Thick arrows indicate the binding sites for inhibitors. In **skeletal muscle**, VDCCs directly activate release of Ca^{2+} from the sarcoplasmic reticulum. The increased cytoplasmic Ca^{2+} concentration triggers muscle contraction. A Ca^{2+}-ATPase in the sarcoplasmic reticulum pumps Ca^{2+} back into the lumen, decreasing the cytoplasmic Ca^{2+} concentration, and the muscle relaxes. In **heart muscle**, VDCCs allow entry of a small amount of Ca^{2+}, which induces release of Ca^{2+} from the lumen of sarcoplasmic reticulum. Two types of Ca^{2+}-ATPases and an Na^{+}/Ca^{2+}-antiporter are responsible for pumping cytoplasmic Ca^{2+} out of the muscle cell. The Na^{+}/Ca^{2+}-antiporter uses the sodium (Na^{+}) gradient produced by Na^{+}/K^{+}-ATPase to antiport Ca^{2+}. DHP, dihydropyridine, nifedipine, a calcium channel blocker used for treatment of hypertension. Ryanodine is a potent inhibitor of the Ca^{++} channel in the sarcoplasmic reticulum.

In cardiac muscle, VDCCs permit the entry of a small amount of Ca^{2+}, which then stimulates Ca^{2+} release through the Ca^{2+} channel from the lumen of the sarcoplasmic reticulum (Ca^{2+}-induced Ca^{2+} release). Not only the sarcoplasmic reticulum Ca^{2+}-ATPase but also an Na^{+}/Ca^{2+}-antiporter and a plasma membrane Ca^{2+}-ATPase are responsible for pumping out cytoplasmic Ca^{2+} from heart muscle (Fig. 8.4, right). The rapid restoration of ion gradients allows for rhythmic contraction of the heart.

Active transport of glucose in the intestines

A Na^+/K^+-ATPase drives uptake of glucose into intestinal and renal epithelial cells

The transport of blood glucose into cells is generally by facilitated diffusion, as the intracellular concentration of glucose is typically less than that of blood (see Table 8.2). In contrast, the transport of glucose from the intestine into blood involves both facilitated diffusion and active transport processes (Fig. 8.5). Active transport is especially important for maximal recovery of sugars from the intestine when the intestinal concentration of glucose falls below that in the blood.

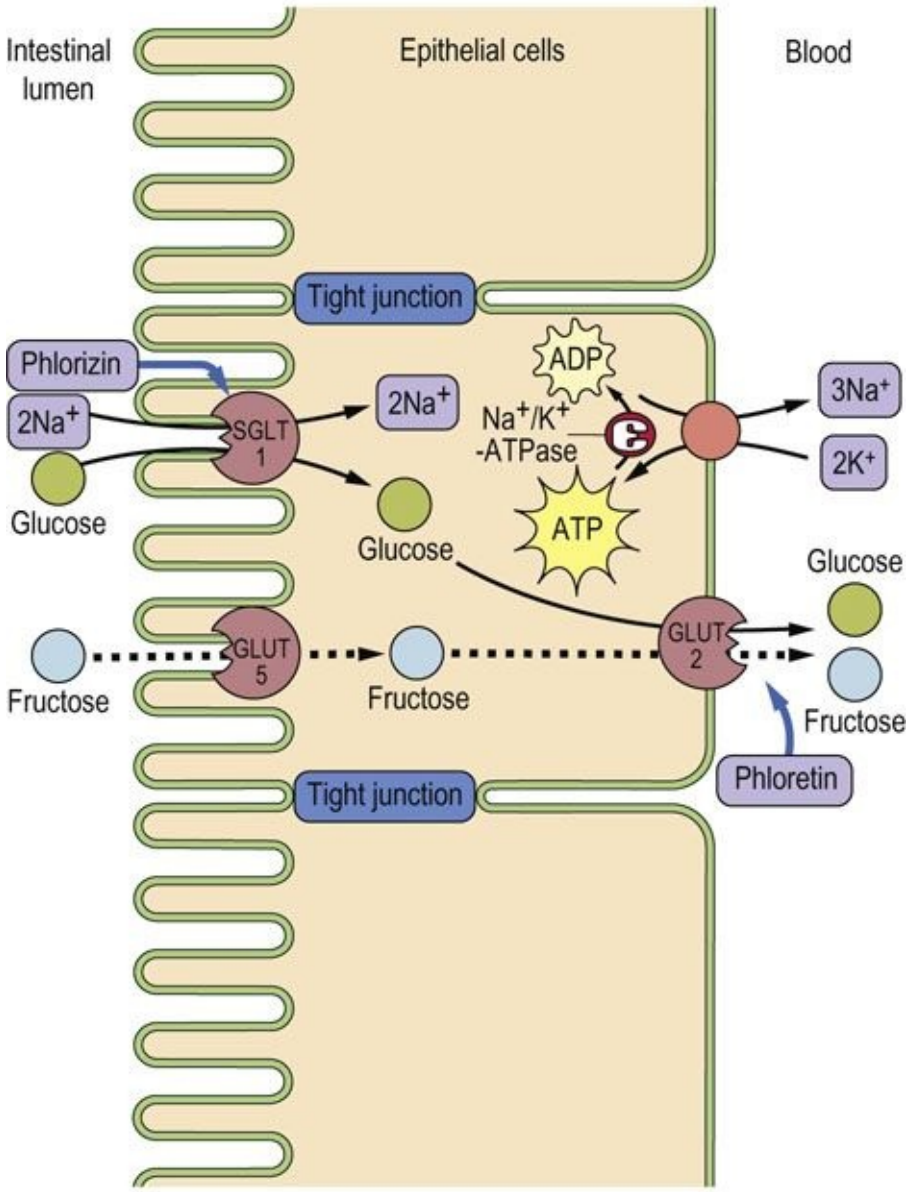


FIG. 8.5 Glucose transport from intestinal lumen into the blood.
Glucose is pumped into the cell through the Na^+-coupled glucose symporter (SGLT1), and passes out of the cell by facilitated diffusion mediated by the GLUT-2 uniporter. The Na^+ gradient for glucose symport is maintained by the Na^+/K^+-ATPase, which keeps the intracellular concentration of Na^+ low. SGLT1 is inhibited by phlorizin and GLUT-2 by phloretin. Phloretin-insensitive GLUT-5 catalyzes the uptake of fructose by facilitated diffusion. The fructose is then exported through GLUT-2. A defect of SGLT1 causes glucose/galactose malabsorption. Adjacent cells are connected by impermeable tight junctions, which prevent solutes from crossing the epithelium.

An Na^+-coupled glucose symporter SGLT1, driven by a Na^+ gradient

formed by Na^+/K^+-ATPase, transports glucose into the intestinal epithelial cell, while **GLUT-2** facilitates the downhill movement of glucose into the portal circulation (see Fig. 8.5). A similar pathway operates in the kidney.

The renal glomerulus is an ultrafiltration system that filters small molecules from blood. However, glucose, amino acids, many ions, and other nutrients in the ultrafiltrate are almost completely reabsorbed in the proximal tubules, by symport processes. Glucose is reabsorbed primarily **by sodium glucose transporter 2 (SGLT2**; one-to-one Na^+**:**Glc stoichiometry) into renal proximal tubular epithelial cells. Much smaller amounts of glucose are recovered by **SGLT1** in a later segment of the tubule, which couples transport of one molecule of glucose to two sodium ions. The concentration of Na^+ in the filtrate is 140 mmol/L (322 mg/dL), while that inside the epithelial cells is 30 mmol/L (69 mg/dL), so that Na^+ flows 'downhill' along its gradient, dragging glucose 'uphill' against its concentration gradient. As in intestinal epithelial cells, the low intracellular concentration of Na^+ is maintained by an Na^+/K^+-ATPase on the opposite side of the tubular epithelial cell, which antiports three cytoplasmic sodium ions for two extracellular potassium ions, coupled with hydrolysis of a molecule of ATP.

Clinical box Modulation of transporter activity in diabetes

The ATP-sensitive K^+ channel (K_{ATP}) participates in regulation of insulin secretion in pancreatic islet β-cells. When the blood concentration of glucose increases, glucose is transported into the β-cell through a glucose transporter (GLUT-2) and metabolized, resulting in an increase in cytoplasmic ATP concentration. The ATP binds to the regulatory subunit of the K^+ channel, K_{ATP}-β (called the sulfonylurea receptor, SUR1) causing structural change of a K_{ATP}-α subunit, which closes the K_{ATP} channel. This induces depolarization of the plasma membrane (decreased voltage gradient across the membrane) and activates voltage-dependent calcium (Ca^{2+}) channels (VDCCs). The entry of Ca^{2+} stimulates exocytosis of vesicles that contain insulin.

The binding of sulfonylureas such as tolbutamide and glibenclamide to K_{ATP}-β on the outside of the plasma membrane is thought to mimic the regulatory effect of intracellular ATP.

Sulfonylureas stimulate insulin secretion, which decreases blood glucose concentration in diabetes.

Defective K_{ATP} channels, which are unable to transport K^+, induce low blood glucose concentration – a condition called Persistent Hyperinsulinemic Hypoglycemia of Infancy (PHHI) that occurs in 1 per 50,000 persons – as a result of loss of K^+-channel function and continuous insulin secretion.

Advanced concept box Various drugs inhibit transporters in muscle

Phenylalkylamine (verapamil), benzothiazepine (diltiazem), and dihydropyridine (DHP; nifedipine) are **Ca^{2+}-channel blockers** that inhibit VDCCs (Fig. 8.4). Ryanodine inhibits the Ca^{2+}-release channel in the sarcoplasmic reticulum. These drugs are used as **antihypertensive agents** to inhibit the increase in cytoplasmic Ca^{2+} concentration and thus the force of muscle contraction. In contrast, cardiac glycosides such as ouabain and digoxin increase heart muscle contraction and are used for treatment of congestive heart failure. They act by inhibiting the Na^+/K^+-ATPase that generates the Na^+ concentration gradient used to drive export of Ca^{2+} by the Na^+/Ca^{2+} antiporter. Snake venoms such as α-bungarotoxin, and tetrodotoxin from the puffer fish inhibit voltage-dependent Na^+ channels. Lidocaine, a Na^+-channel blocker, is used as a local anesthetic and antiarrhythmic drug. Inhibition of Na^+ channels represses transmission of the depolarization signal.

Acidification of gastric juice by a proton pump in the stomach

A P-ATPase in gastric parietal cells maintains the low pH of

the stomach

The lumen of the stomach is highly acidic (pH ≈1) because of the presence of a proton pump (**H^+/K^+-ATPase**; P-ATPase in Table 8.3) that is specifically expressed in gastric parietal cells. The gastric proton pump is localized in intracellular vesicles in the resting state. Stimuli such as histamine and gastrin induce fusion of the vesicles with the plasma membrane (Fig. 8.6A). The pump antiports two cytoplasmic protons and two extracellular potassium ions, coupled with hydrolysis of a molecule of ATP; thus it is called an H^+/K^+-ATPase. The counter-ion $C1^-$ is secreted through a Cl^- channel, producing hydrochloric acid (HCl) (gastric acid) in the lumen (Fig. 8.6B).

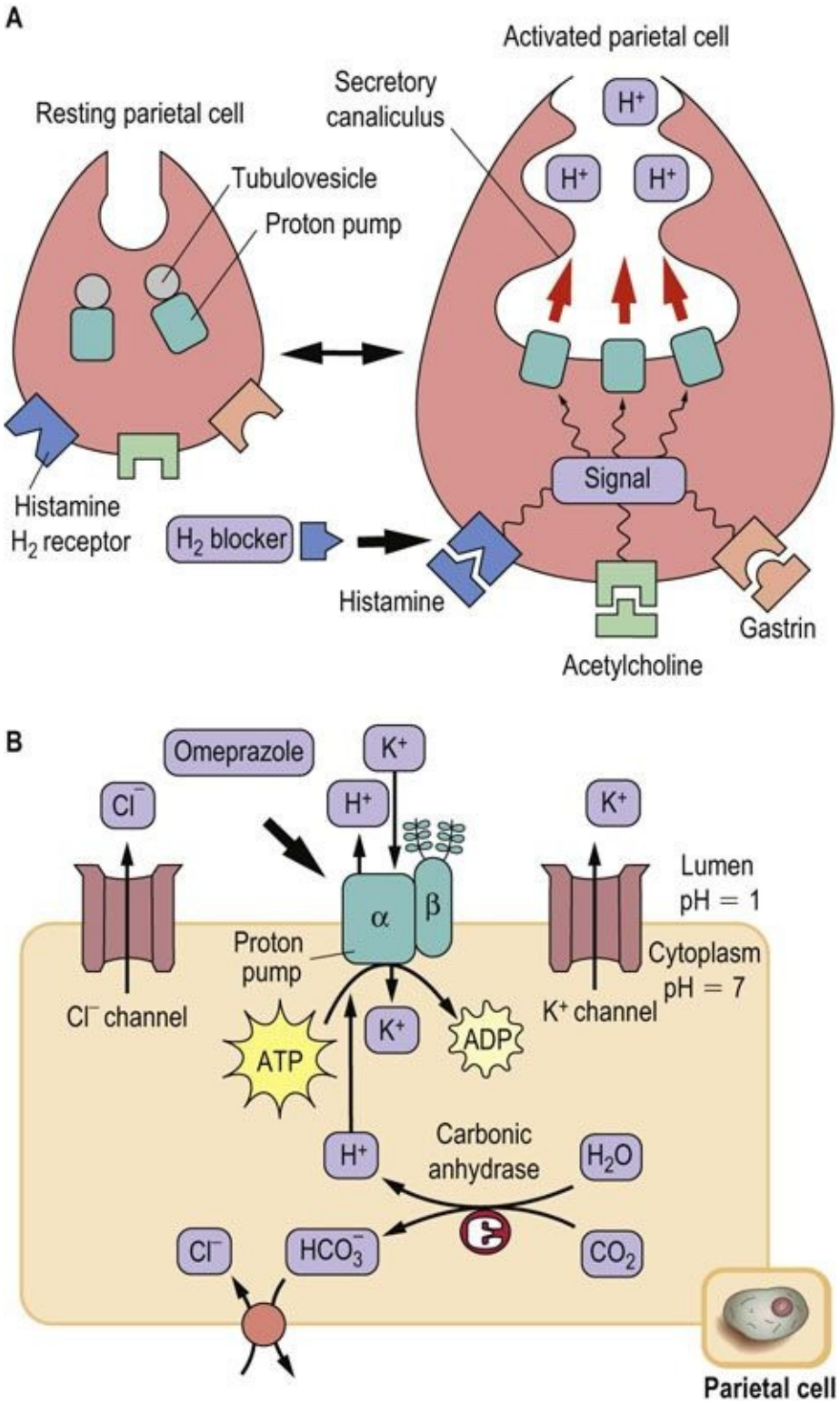


FIG. 8.6 Acid secretion from gastric parietal cells.
(A) Acid secretion is stimulated by extracellular signals and accompanied by morphologic changes in parietal cells, from resting (left) to activated (right). The proton pump (H^+/K^+-ATPase) moves to the secretory canaliculus (plasma membrane) from cytoplasmic

tubulovesicles. H_2-blockers compete with histamine at the histamine H_2-receptor. (B) Ion balance in the parietal cell. The H^+ transported by the proton pump are supplied by carbonic anhydrase. Bicarbonate, the other product of this enzyme, is antiported with Cl^-, which is secreted through a Cl^- channel. The potassium ions imported by the proton pump are again excreted by a K^+ channel. The proton pump has catalytic α-and glycosylated β-subunits. The drug omeprazole covalently modifies cysteine residues located in the extracytoplasmic domain of the α-subunit and inhibits the proton pump. Thick arrows indicate the binding sites for inhibitors.

Renal Na^+ transport and renin–angiotensin–aldosterone system

The kidney is a major site for regulation of Na^+ homeostasis; several transport systems participate in Na^+ handling in the kidney tubules

The kidney plays important parts in preservation of water and electrolyte homeostasis (Chapter 23). The primary urine is formed by glomerular filtration. Salts passing through the glomerular filter are reabsorbed in the renal tubule. Most of the Na^+ reabsorption in the tubules occurs in proximal segments. Although <10% of Na^+ reabsorption occurs in the distal nephron, this site is recognized to be essential for fine-tuning of plasma volume since Na^+ reabsorption in the distal nephron is sensitive to the steroid hormone aldosterone.

As indicated in Figure 8.7, several Na^+ transport proteins participate in Na^+ reabsorption in the kidney tubules. Among them, Na^+/K^+-ATPase expressed in the basolateral membranes of epithelial cells transports Na^+ across the membrane into the blood and generates a low intracellular Na^+ concentration. The Na^+ ions in the urine are transported into the epithelial cells through apical membranes by segment-specific Na^+ transport proteins. Na^+/H^+-antiporter (NHE3) mediates Na^+ entry in the proximal tubule. Furosemide-sensitive Na^+-K^+-$2Cl^-$-symporter (NKCC2) and thiazide-sensitive Na^+-Cl^--cotransporter (NCCT) are localized in thick ascending loop of Henle and distal convoluted tubules, respectively. Amiloride-sensitive epithelial Na^+ channel (ENaC) is expressed in the aldosterone-sensitive distal nephron. The ionic balance of epithelial cells is also regulated by the renal outer medullary K^+ channel (ROMK) and the Cl^- channel (ClC).

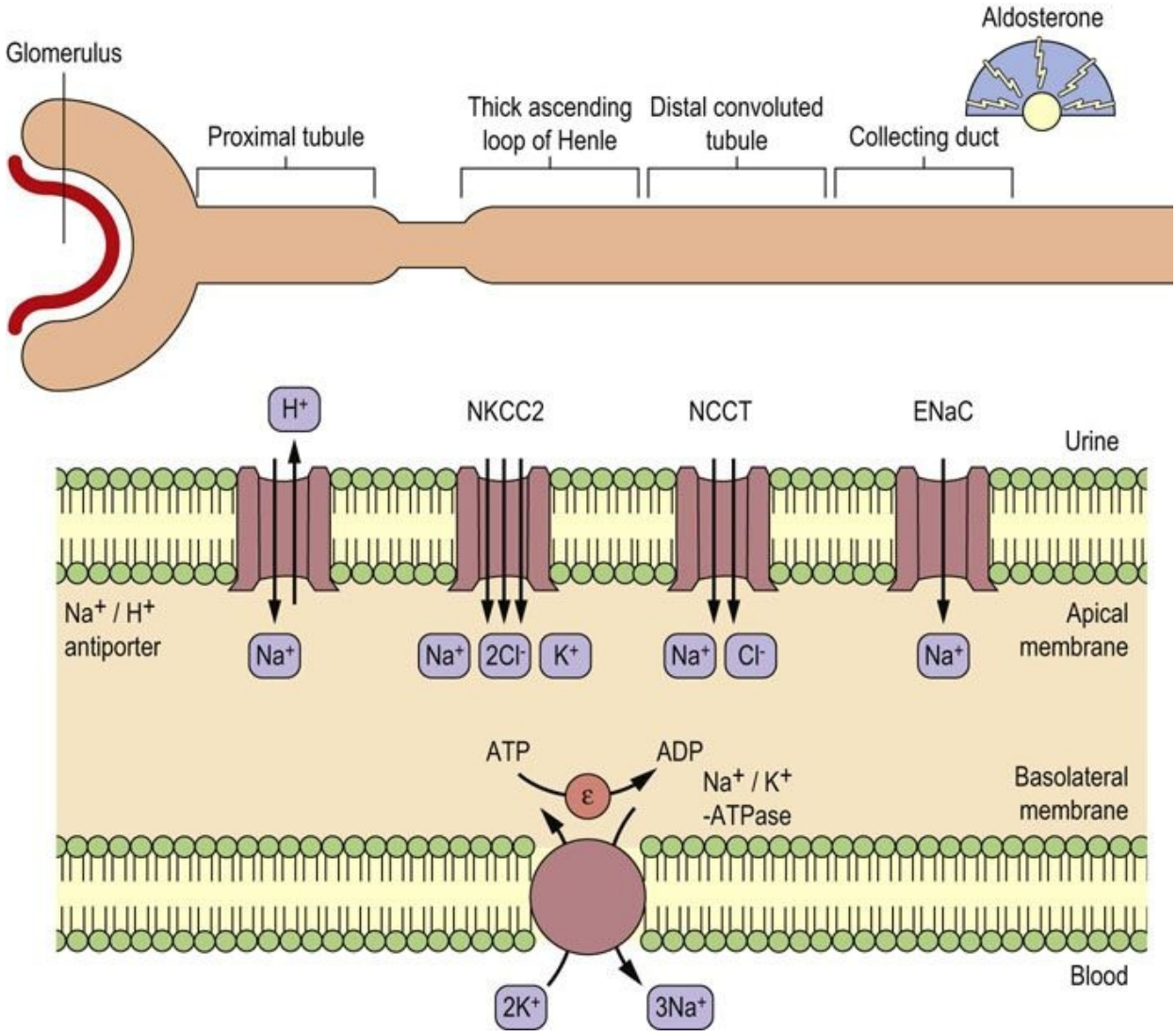


FIG. 8.7 Segment-specific expression of renal Na^+ transport proteins.
Various Na^+ transport proteins expressed in the apical membranes of renal epithelial cells transport Na^+ from the urine into the cells. Na^+ in the cells is excreted into the blood by an Na^+/K^+-ATPase expressed in the basolateral membranes. K^+ ions are recycled through basolateral ROMK in the thick ascending loop of Henle and aldosterone sensitive distal nephron. Chloride ions leaves the cells in the thick ascending loop of Henle and distal convoluted tubule through ClCs.

The renin-angiotensin-aldosterone system plays a central role in the regulation of circulating volume, extracellular fluid osmolality and blood pressure

As discussed in more detail in Chapters 24 and 25, renin is produced in the

juxtaglomerular apparatus of the kidney upon decrease of circulating volume, and it converts angiotensinogen produced in the liver to angiotensin I. Angiotensin I is converted to angiotensin II by angiotensin-converting enzyme (ACE). Angiotensin II produced in the blood promotes the release of aldosterone from the adrenal glands. Aldosterone promotes renal Na^+ and water retention and elevated blood pressure by increasing Na^+ reabsorption in the cortical collecting duct. Angiotensin II also promotes elevations in blood pressure by direct vasoconstriction and by increasing Na^+ reabsorption in the proximal tubule.

Clinical box Inhibiting the gastric proton pump and eradication of *helicobacter pylori*

Chronic strong acid secretion by the gastric proton pump injures the stomach and the duodenum, leading to gastric and duodenal ulcers. Proton pump inhibitors such as omeprazole are delivered to parietal cells from the circulation after oral administration. Omeprazole is a prodrug: it accumulates in the acidic compartment, as it is a weak base, and is converted to the active compound under the acidic conditions in the gastric lumen. The active form covalently modifies cysteine residues located in the extracytoplasmic domain of the proton pump. H_2-blockers (receptor antagonists) such as cimetidine and ranitidine indirectly inhibit acid secretion by competing with histamine for its receptor (Fig. 8.6).

Comment.

Infection of the stomach by *Helicobacter pylori* also causes ulcers and is associated with an increased risk of gastric adenocarcinoma. Recently, antibiotic treatment has been introduced to eradicate *H. pylori*. Interestingly, antibiotic treatment together with omeprazole is much more effective, possibly because of an increased stability of the antibiotic under the weakly acidic condition produced by proton pump inhibition.

Summary

■ Most of the permeability properties of the membrane are determined by transport proteins, which are integral membrane proteins.
■ Protein-mediated transport is a saturable process with high substrate specificity.
■ Facilitated diffusion is catalyzed by transporters that permit the movement of ions and molecules down concentration gradients, whereas uphill or active transport requires energy.
■ Primary active transport is catalyzed by pump ATPases that use energy produced by ATP hydrolysis.
■ Secondary active transport uses electrochemical gradients of Na^+ and H^+, or membrane potential produced by primary active transport processes. Uniport, symport, and antiport are examples of secondary active transport.
■ Numerous substrates such as ions, nutrients, small organic molecules including drugs and peptides, and proteins are transported by various transporters.
■ All these transporters are indispensable for homeostasis. The expression of unique sets of transporters is important for specific cell functions such as muscle contraction, nutrient and ion absorption by intestinal epithelial cells, resorption of nutrients by kidney cells, and secretion of acid from gastric parietal cells.

Active learning

1. Describe the similarities between the kinetics of enzyme action and transport processes. Compare the properties of various glucose transporters with those of hexokinase and glucokinase, both kinetically and in terms of physiologic function.
2. Identify a number of transport inhibitors used in clinical medicine, *e.g.* Ca^{2+}-channel blockers, laxatives and inhibitors of gastric acid secretion.
3. Investigate the process of glucose transport across the blood–brain barrier and explain the pathogenesis of hypoglycemic coma.
4. Study the role and specificity of ABC transporter in multidrug resistance to chemotherapeutic agents.

Further reading

Camargo, SM, Brockenhauer, D, Kleta, R. Aminoacidurias: clinical and molecular aspects. *Kidney Int*. 2008; 73:918–925.

Hinz, A, Tampe, R. ABC transporters and immunity: mechanism of self-defense. *Biochemistry*. 2012; 51:4981–4989.

Linton, KJ. Structure and function of ABC transporters. *Physiology*. 2007; 22:122–130.

Prassas, I, Diamandis, EP. Novel therapeutic applications of cardiac glycosides. *Nat Rev Drug Discov*. 2008; 7:926–935.

Seyberth, HW, Schlingmann, KP. Bartter-and Gitelman-like syndromes: salt-losing tubulopathies with loop or DCT defects. *Pediatr Nephrol*. 2011; 26:1789–1802.

Thorens, B, Mueckler, M. Glucose transporters in the 21st century. *Am J Physiol Endocrinol Metab*. 2010; 298:E141–E145.

Websites

General reviews

Transport across cell membrane. http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/D/Diffusion

Membrane transpor. www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb1/part2/carriers.htm

Human ABC transporter. http://nutrigene.4t.com/humanabc.htm

P-ATPas. www.traplabs.dk/patbase

Animation

www.stolaf.edu/people/giannini/biological%20anamations.html

www.wiley.com/legacy/college/boyer/0470003790/animations/membrane_transport/membrane_transport.ht

www.phschool.com/science/biology_place/biocoach/biomembrane1/intro.html

CHAPTER 9

Bioenergetics and Oxidative Metabolism

L.William Stillway

Learning objectives

After reading this chapter you should be able to:

- Describe how thermodynamics is related to nutrition and obesity.
- Outline the mitochondrial electron transport system showing eight major electron carriers.
- Explain how ubiquinone, heme and the iron–sulfur complexes participate in electron transport.
- Define membrane potential and explain its role in ATP synthesis and thermogenesis.
- Explain the role of uncoupling proteins in thermogenesis.
- Describe the mechanism of ATP synthase.
- Describe the effects of inhibitors such as rotenone, antimycin A, carbon monoxide, cyanide and oligomycin on oxygen uptake by mitochondria.

Introduction

ATP is the central metabolic currency

Oxidation of metabolic fuels is essential to life. In higher organisms, fuels such as carbohydrates and lipids are metabolized to carbon dioxide and water, generating a central metabolic currency, adenosine triphosphate (ATP). Most metabolic energy is produced by oxidation-reduction (redox) reactions in mitochondria. The regulation of energy metabolism is no small feat, because warm-blooded animals have such variable demands for energy from such processes as thermogenesis at low temperatures, stimulation of ATP synthesis during stress, and coupling of ATP synthesis with the rate of respiration during work and exercise. This chapter will provide an introduction to the concept of free energy, oxidative phosphorylation and the transduction of energy from fuels into useful work. The pathways and specific molecules through which electrons are transported to oxygen and the mechanism of generation of ATP will be described and related to the structure of the mitochondrion, the powerhouse of the cell and the major source of cellular ATP. Lastly, these biochemical processes will be applied to human health and disease.

Oxidation as a source of energy

Energy content of foods

Nutrition and disorders such as obesity, diabetes, and cancer all require an understanding of thermodynamics. Obesity, for example, is a disorder in which there is an imbalance between energy intake and expenditure. It is therefore important that the energy content of foods be known. The commonly accepted energy values for the four major food categories are shown in Table 9.1; alcohol is included because it is a significant dietary component for some people. These values are obtained by completely burning (oxidizing) samples of each food. Biologically, about 40% of food energy is conserved as ATP, and the remaining 60% is liberated as heat.

Table 9.1

Energy content of the major classes of food

	Metabolic fuel (kJ/g)	Energy content (kcal/g)
Fats	38	9
Carbohydrates	17	4
Proteins	17	4
Alcohol	29	7

Note that the thermodynamic term, kcal (energy required to increase the temperature of 1 kg (1 L) of water by 1°C) is equivalent to the common nutritional Calorie (capital C), *i.e.* 1 Cal = 1 kcal, 1 kcal = 4.2 kJ.

The basal metabolic rate (BMR)

The basal metabolic rate is a measure of the total daily energy expenditure by the body at rest

Virtually all of the reactions in the body are exothermic, and the sum of all reactions at rest is called the basal metabolic rate (BMR), which can be measured by two basic methods: **direct calorimetry**, where the total heat liberated by an animal is measured over time, and **indirect calorimetry**, where the BMR is calculated from the quantity of oxygen consumed, which is directly related to the BMR. Adult men (70 kg) have a BMR of about 7500 kJ (1800 kcal) and women about 5400 kJ (1300 kcal) per day; the BMR may vary

by a factor of two between individuals, depending on age, sex, body mass and composition. Heat production by mitochondria accounts for the largest portion of the BMR. The BMR is measured under controlled conditions: after an 8-hour sleep, in the reclining position, in the postabsorptive state, typically after a 12-hour fast.

Another measure often used is the RMR, or resting metabolic rate, which is virtually the same as the BMR but measured under less restrictive conditions. The RMR is a measure of minimum energy expenditure at rest; it is typically about 70% of total daily energy expenditure. Exercise scientists frequently use the term **MET, metabolic equivalent task**, as a measure of energy expenditure at rest. Slow to vigorous walking is a 2–4 MET/h activity; vigorous running on a treadmill may consume more than 15 MET/h.

Stages of fuel oxidation

The oxidation of fuels can be divided into two general stages: production of reduced nucleotide coenzymes during the oxidation of fuels, and use of the free energy from oxidation of the reduced coenzymes to produce ATP (Fig. 9.1).

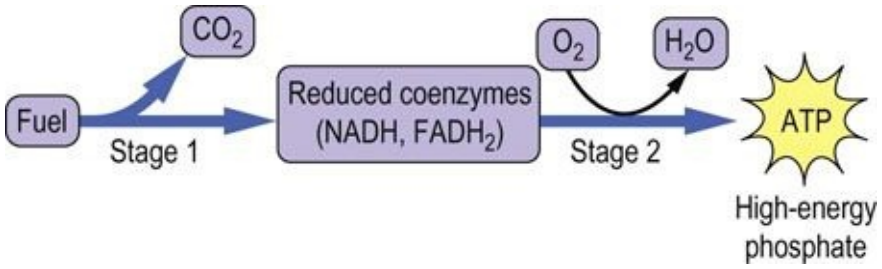


FIG. 9.1 Stages of fuel oxidation.
NADH, reduced nicotinamide adenine dinucleotide; $FADH_2$, reduced flavin adenine dinucleotide.

Free energy

The direction of a reaction depends on the difference between the free energy of reactants and products

The Gibbs' free energy (ΔG) of a reaction is the maximum amount of energy that can be obtained from a reaction at constant temperature and pressure. The units of free energy are kcal/mol (kJ/mol). It is not possible to measure the absolute free energy content of a substance directly, but when reactant A reacts to form product B, the free energy change in this reaction, ΔG, can be determined.

For the reaction A → B:

$$\Delta G = G_B - G_A$$

where G_A and G_B are the free energy of A (reactant) and B (product), respectively. All reactions in biological systems are considered to be reversible reactions, so that the free energy of the reverse reaction, B → A, is numerically equivalent, but opposite in sign to that of the forward reaction.

If there is a greater concentration of B than of A at equilibrium, *i.e.* Keq > 1, the conversion A → B is favorable – that is, the reaction tends to move forward from a standard state in which A and B are present at equal concentrations. In this case, the reaction is said to be a spontaneous or exergonic reaction, and the free energy of this reaction is defined as negative: that is, ΔG < 0, indicating that energy is liberated by the reaction. Conversely, if the concentration of A is greater than that of B at equilibrium, the forward reaction is termed unfavorable, nonspontaneous or endergonic, and the reaction has a positive free energy: that is, when starting concentrations are equal, B tends to form A, rather than A to form B. In this case, energy input would be required to push the reaction A → B forward from its equilibrium position to the standard state in which A and B are present at equal concentrations. The total free energy available from a reaction depends on both its tendency to proceed forward from the standard state (ΔG) and the amount (moles) of reactant converted to product.

The free energy of metabolic reactions is related to their

equilibrium constants by the Gibbs' equation

Thermodynamic measurements are based on standard-state conditions where reactant and product are present at 1 molar concentrations, the pressure of all gases is 1 atmosphere and the temperature is 25°C (298K). Most commonly, the concentrations of reactants and products are then measured after equilibrium is attained. Standard free energies are represented by the symbol ΔG° and biological standard free energy change by ΔG°′, with the accent symbol designating pH 7.0. The free energy available from a reaction may be calculated from its equilibrium constant by the Gibbs' equation:

$$\Delta G^{\circ\prime} = -RT\ln K'_{eq}$$

where T is absolute temperature (Kelvin), $\ln K'_{eq}$ is the natural logarithm of the equilibrium constant for the reaction at pH 7, and R is the ideal gas constant:

$$R = (8.3\,J\,mol^{-1}/K \text{ or } \sim 2\,cal\,mol^{-1}/K)\,.$$

Several common metabolic intermediates that you will encounter in your studies are listed in Table 9.2, along with the equilibrium constants and free energies for their hydrolysis reactions. Those intermediates with free energy changes equal to or greater than that of ATP, the central energy transducer of the cell, are considered to be **high-energy compounds**, and generally have either anhydride or thioester bonds. The lower-energy compounds listed are all phosphate esters and, in comparison, do not yield as much free energy on hydrolysis. The hydrolysis reaction of glucose-6-phosphate (Glc-6-P) is written as:

Table 9.2

Thermodynamics of hydrolysis reactions

Metabolite	K'eq	ΔG°' (kJ/mol)	(kcal/mol)
Phosphoenolpyruvate	1.2×10^{11}	−61.8	−14.8
Phosphocreatine	9.6×10^{8}	−50.2	−12.0
1,3-bisphosphoglycerate	6.8×10^{8}	−49.3	−11.8
Pyrophosphate	9.7×10^{5}	−33.4	−8.0
Acetyl coenzyme A	4.1×10^{5}	−31.3	−7.5
ATP	2.9×10^{5}	−30.5	−7.3
Glucose-1-phosphate	5.5×10^{3}	−20.9	−5.0
Fructose-6-phosphate	7.0×10^{2}	−15.9	−3.8
Glucose-6-phosphate	3.0×10^{2}	−13.8	−3.3

Equilibrium constants and free energy of hydrolysis of various metabolic intermediates at pH 7 (ΔG°').

$$\text{Glc-6-P} + H_2O \rightarrow \text{Glucose} + \text{Pi}$$

This reaction has a negative free energy and occurs spontaneously. The reverse reaction, synthesis of Glc-6-P from glucose and phosphate, would require input of energy.

Conservation of energy by coupling with adenosine triphosphate

ATP is a product of catabolic reactions and a driver of biosynthetic reactions

Living systems must transfer energy from one molecule to another without losing all of it as heat. Some of the energy must be conserved in a chemical form in order to drive nonspontaneous biosynthetic reactions. In fact, nearly half of the energy obtained from the oxidation of metabolic fuels is channeled into the synthesis of **ATP, a universal energy transducer in living systems**. ATP is often referred to as the common currency of metabolic energy, because it is used to drive so many energy-requiring reactions. ATP consists of the purine base adenine, the five-carbon sugar ribose, and α, β, and γ phosphate groups (Fig. 9.2). The two phosphoanhydride linkages are said to be high-energy bonds, because their hydrolysis yields a large negative change in free energy. When ATP is used for metabolic work, these high-energy linkages are broken and ATP is converted to ADP or to AMP.

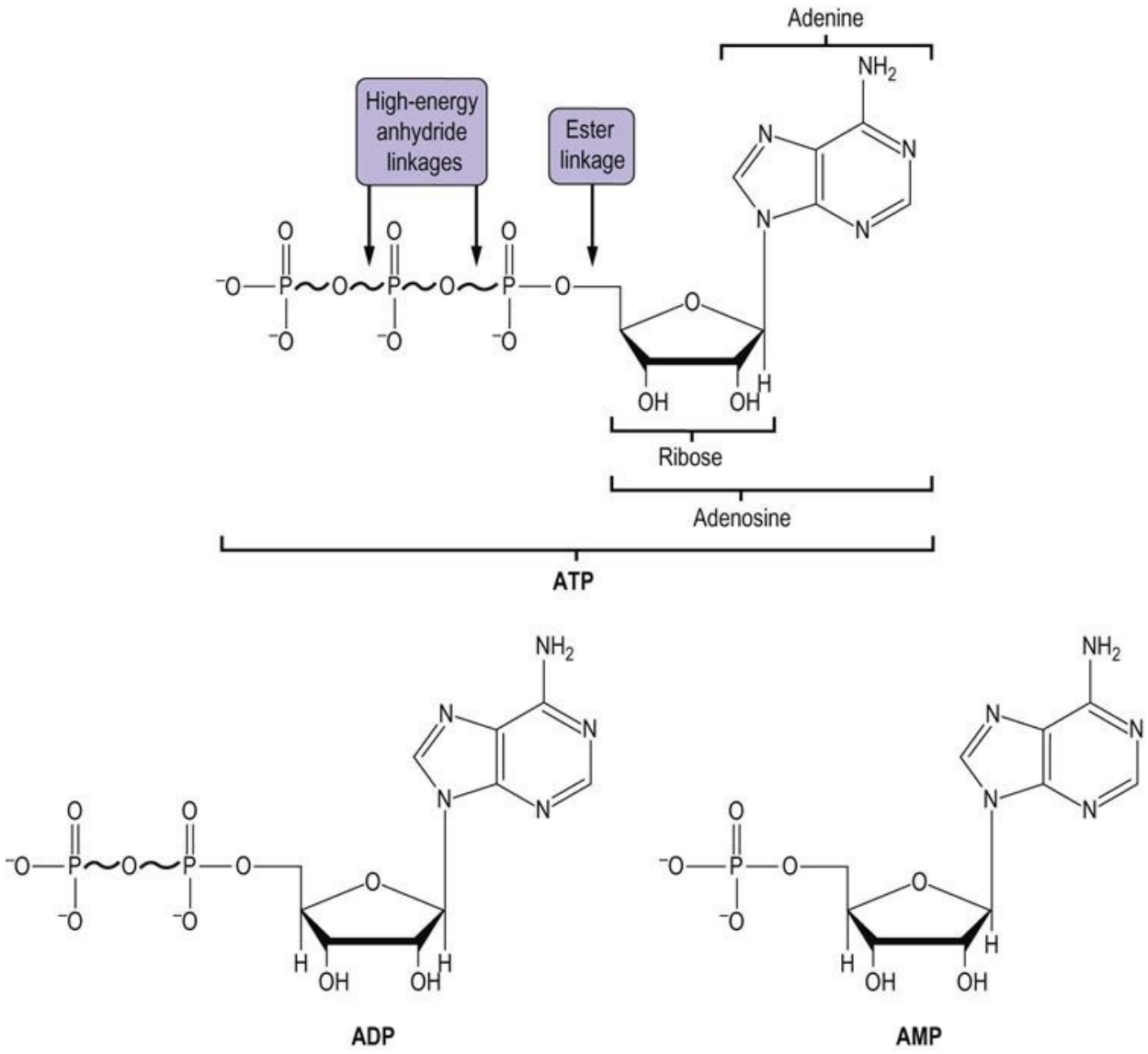


FIG. 9.2 Structures of adenine nucleotides.
ATP is shown, together with its hydrolysis products, adenosine diphosphate (ADP) and adenosine monophosphate (AMP). ATP has two high-energy phosphoanhydride bonds; ADP has one; and AMP has only a low-energy phosphoester bond.

The free energy of a high-energy bond, such as the phosphate anhydride bonds in ATP, can be used to drive or push forward reactions that would otherwise be unfavorable. In fact, nearly all biosynthetic pathways are thermodynamically unfavorable, but are made favorable by coupling various reactions with hydrolysis of high-energy compounds. For example, the first step in the metabolism of glucose is the synthesis of Glc-6-P (Fig. 3.4). As shown in Table 9.2, this is not a favorable reaction: the hydrolysis of Glc-6-P ($\Delta G^{\circ\prime}$ = −13.8 kJ/mol or −3.3 kcal/mol) is the favored reaction. However, as shown

below, the synthesis of Glc-6-P (reaction I) can be **energetically coupled** to the hydrolysis of ATP (reaction II), yielding a 'net reaction' III that is favorable for synthesis of Glc-6-P:

	$\Delta G^{\circ\prime}$
I : $Glc + Pi \rightarrow Glc\text{-}6\text{-}P + H_2O$	+ 3.3 kcal/mol
II : $ATP + H_2O \rightarrow ADP + Pi$	– 7.3 kcal/mol
Net : $Glc + ATP \rightarrow Glc\text{-}6\text{-}P + ADP$	– 4 kcal/mol

This is possible because of the high free energy or 'group transfer potential' of ATP. The physical transfer of the phosphate from ATP to glucose occurs in the active site of a kinase enzyme, such as glucokinase. This motif, in which ATP is used to drive biosynthetic reactions, transport processes or muscle activity, occurs commonly in metabolic pathways.

Mitochondrial synthesis of adenosine triphosphate from reduced coenzymes

Oxidative phosphorylation is the mechanism by which energy derived from fuel oxidation is conserved in the form of ATP

Metabolism of carbohydrates begins in the cytoplasm through the glycolytic pathway (see Chapter 12), whereas energy production from fatty acids occurs exclusively in the mitochondrion. Mitochondria are subcellular organelles, about the size of bacteria. They are essential for aerobic metabolism in eukaryotes. Their main function is to oxidize metabolic fuels and conserve free energy by synthesizing ATP.

Mitochondria are bounded by a dual membrane system (Fig. 9.3). The outer membrane (OMM) contains enzyme and transport proteins and via the pore-forming protein porin (P), it is permeable to virtually all ions, small molecules (S) and proteins less than 10,000 Da. Large proteins must be transported via the **TOM** (translocase in the outer mitochondrial membrane) and **TIM** (translocase in the inner mitochondrial membrane) complexes. This is especially vital to the cell, because almost all mitochondrial proteins are nuclear encoded and must be transported into the mitochondrion. The **mitochondrial genome**, mtDNA, encodes 13 vital subunits of the proton pumps and ATP synthase. The **inner membrane (IMM)** is pleated with structures known as **cristae**, and is impermeable to most ions and small molecules, such as nucleotides (including ATP), coenzymes, phosphate, and protons. Transporter proteins are required to selectively facilitate translocation of specific molecules across the inner membrane. The inner membrane also contains components of oxidative phosphorylation – the process by which the oxidation of reduced nucleotide coenzymes is coupled to the synthesis of ATP.

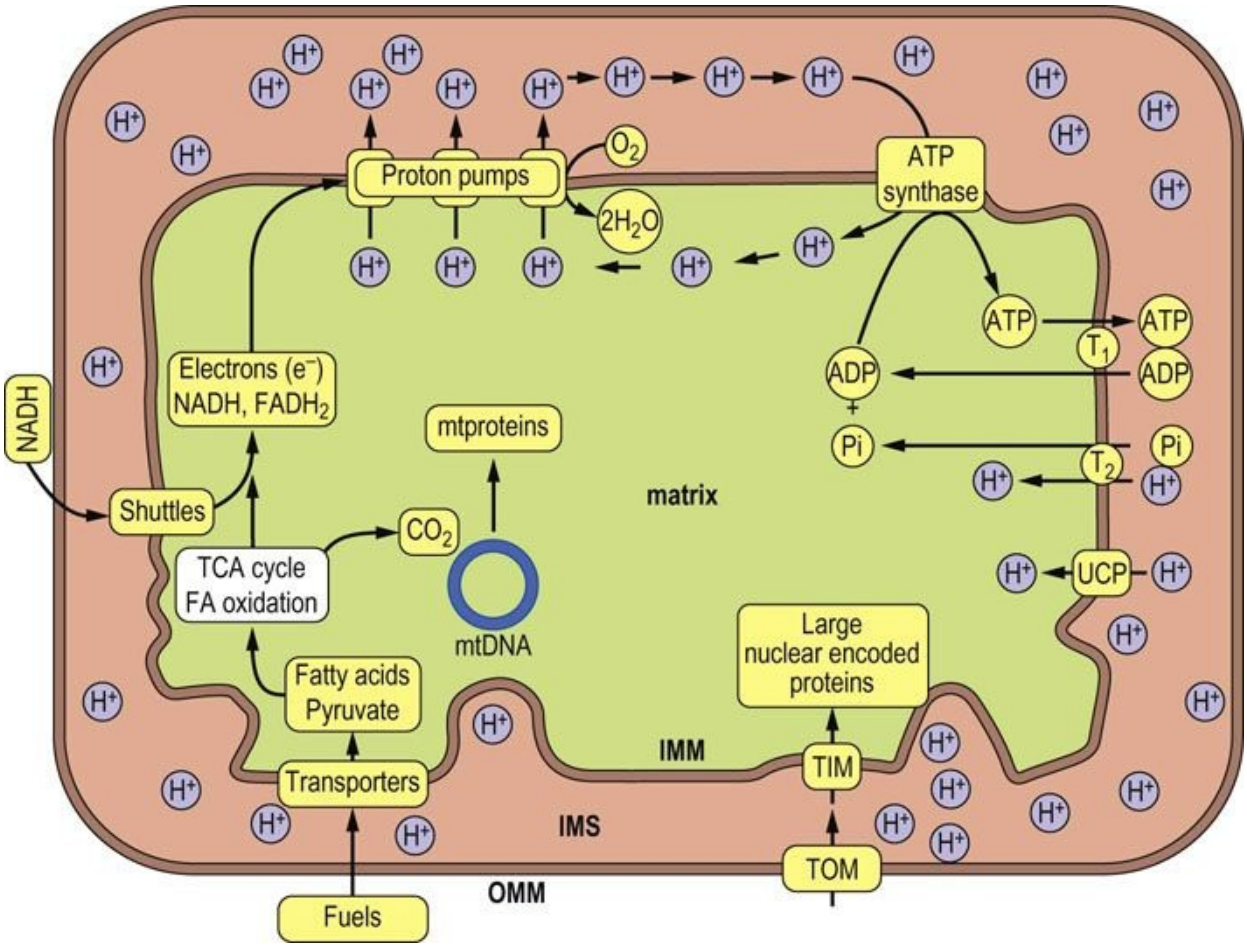


FIG. 9.3 Mitochondrial structure and pathways of energy transduction: the mechanism of oxidative phosphorylation.
Major fuels, such as pyruvate from carbohydrates and fatty acids (FA) from triglycerides, are transported into the matrix where they are oxidized to generate CO_2 and the reduced nucleotide coenzymes NADH and $FADH_2$. Oxidation of these nucleotides via the electron transport system reduces oxygen to water and pumps protons by three proton pumps out of the matrix and into the intermembrane space (IMS), creating a pH gradient, which is the major contributor to the membrane potential. It should be noted that protons in the intermembrane space freely diffuse through the outer membrane via the protein porin, a proton channel, so the intermembrane space is roughly equivalent to the cytosol. Although the membrane potential is mostly composed of the proton gradient it actually consists of several electrochemical gradients and is expressed as a voltage. Controlled influx of protons through ATP synthase powers the synthesis of ATP by ATP synthase (F-ATPase, Table 8.3). Mitochondrial ATP is then exchanged for cytoplasmic ADP through the ADP-ATP translocase (T_1). Phosphate (Pi), which is also required for ATP synthesis, is transported by the phosphate translocase (T_2). The inner membrane also contains uncoupling proteins (UCP) that may be used to allow the controlled leakage of protons back into the matrix. OMM, outer mitochondrial membrane; IMM, inner mitochondrial membrane; mtproteins, mitochondrial proteins; mtDNA, mitochondrial DNA; TOM and TIM, protein translocase complexes in outer and inner mitochondrial membrane; TCA, tricarboxylic acid cycle.

Advanced concept box Exercise and mitochondrial biogenesis

It has long been known that exercise increases the oxidative capacity of skeletal muscle by inducing mitochondrial biogenesis. Continued exercise results in energy consumption, and AMP accumulates. AMP-activated protein kinase is a fuel sensor, and it plays a critical role in initiating the production of new mitochondria and electron transport components such as heme. Such mechanisms are not only of importance in exercise training but also in the regeneration of tissues after tissue injury, such as trauma, heart attacks and strokes.

Transduction of energy from reduced coenzymes to high-energy phosphate

NAD^+, FAD, and FMN are the major redox coenzymes

The major redox coenzymes involved in transduction of energy from fuels to ATP are nicotinamide adenine dinucleotide (NAD^+), flavin adenine dinucleotide (FAD) and flavin mononucleotide (FMN) (Fig. 9.4). During energy metabolism, electrons are transferred from carbohydrates and fats to these coenzymes, reducing them to NADH, $FADH_2$ and $FMNH_2$. In each case, two electrons are transferred, but the number of protons transferred differs. NAD^+ accepts a hydride ion (H^-) that consists of one proton and two electrons; the remaining proton is released into solution. FAD and FMN accept two electrons and two protons.

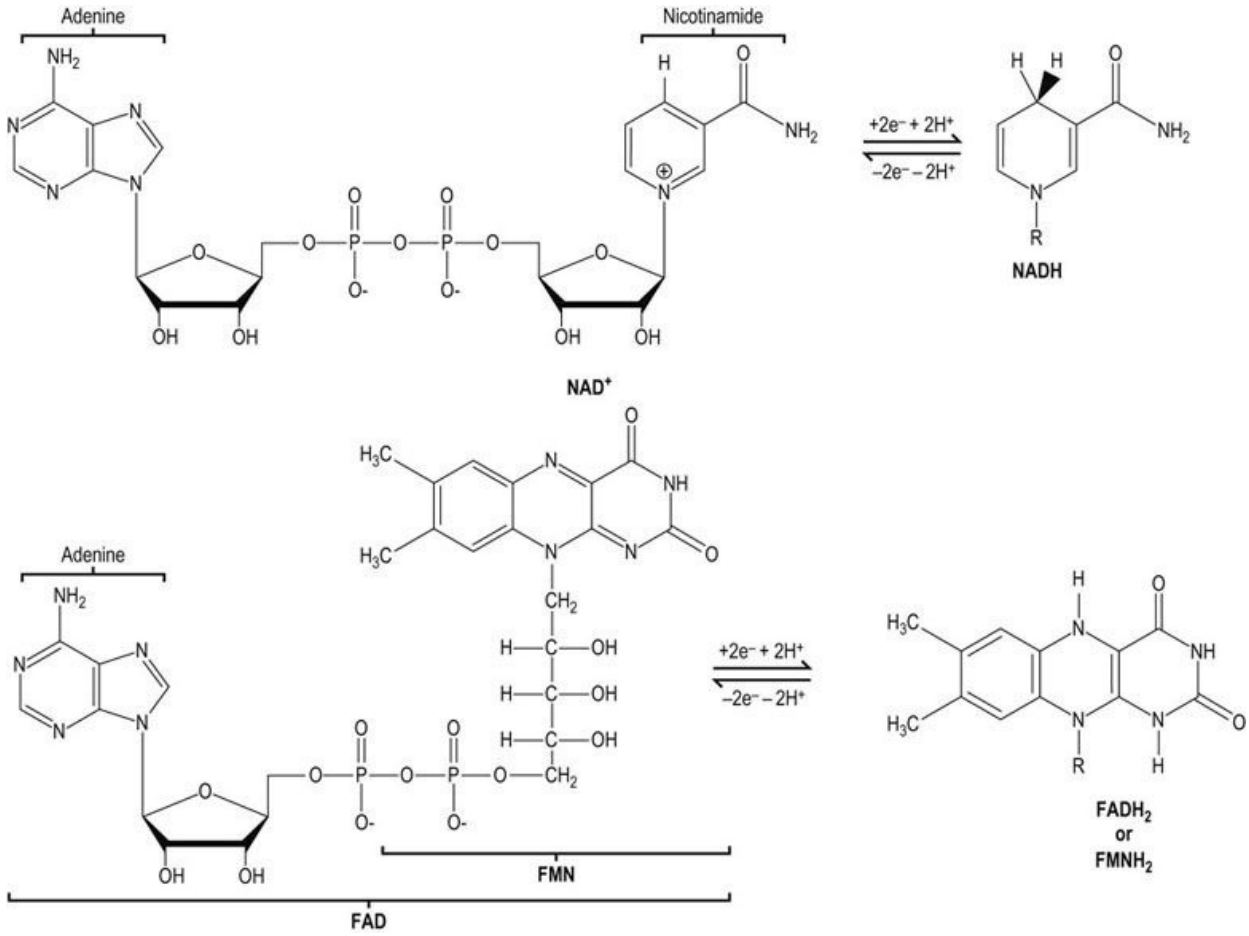


FIG. 9.4 The structure of redox coenzymes.
NAD^+ and its reduced form, NADH (nicotinamide adenine dinucleotide), consists of adenine, two ribose units, two phosphates, and nicotinamide. FAD and its reduced form, $FADH_2$ (flavin adenine dinucleotide), consists of riboflavin, two phosphates, ribose and adenine. FMN and $FMNH_2$ consist of riboflavin phosphate. The nicotinamide and riboflavin components of these coenzymes are reversibly oxidized and reduced during electron transfer (redox) reactions. NADH and $FADH_2$ are often called reduced nucleotides or reduced coenzymes.

The oxidation of reduced nucleotides by the electron transport system produces a large amount of free energy. When the oxidation of 1 mole of NADH is coupled to the reduction of 0.5 mole of oxygen to form water, the energy produced is theoretically sufficient to synthesize 7 moles of ATP:

$$NADH + H^+ + \tfrac{1}{2}O_2 \rightarrow NAD^+ + H_2O$$

$$\Delta G^{\circ\prime} = -220\ \text{kJ/mol}\ (-52.4\ \text{kcal/mol})$$

$$\text{ADP} + \text{Pi} \rightarrow \text{ATP} + \text{H}_2\text{O}$$

$$\Delta G^{\circ\prime} = -30.5\ \text{kJ/mol}\ (-7.3\ \text{kcal/mol})$$

Dividing 220 kJ/mol of $\Delta G^{\circ\prime}$ available from oxidation of NADH by $\Delta G^{\circ\prime}$ 30.5 required for synthesis of ATP yields, theoretically, 7 mol ATP/mol NADH. As discussed below, the actual yield is closer to 2.5 mol ATP/mol NADH oxidized.

The free energy of oxidation of NADH and $FADH_2$ is used by the electron transport system to pump protons into the intermembrane space. The energy produced when these protons reenter the mitochondrial matrix is used to synthesize ATP. This process is known as oxidative phosphorylation (see Fig. 9.3)

Advanced concept box Metabolic function of atp requires magnesium

ATP readily forms a complex with magnesium ion, and it is this complex that is required in all reactions in which ATP participates, including its synthesis. A magnesium deficiency impairs virtually all of metabolism, because ATP can neither be made nor utilized in adequate amounts.

The mitochondrial electron transport system

The mitochondrial electron transport chain transfers electrons in a defined multi-step sequence from reduced nucleotides to oxygen

The entire electron transport system, also known as the electron transport chain or respiratory chain, is located in the inner mitochondrial membrane (Fig. 9.5). It consists of several large protein complexes and two small, independent components – ubiquinone and cytochrome *c*. The protein components are very complex; complex I, for example, which accepts electrons from NADH, contains at least 46 subunits. Each step in the electron transport chain involves a redox reaction where electrons are transferred from components with more negative reduction potentials to components with more positive reduction potentials. Electrons are conducted through this system in a defined sequence from reduced nucleotide coenzymes to oxygen, and the free energy changes drive the transport of protons from the matrix into the intermembrane space via the three proton pumps. After each step, the electrons are at a lower energy state.

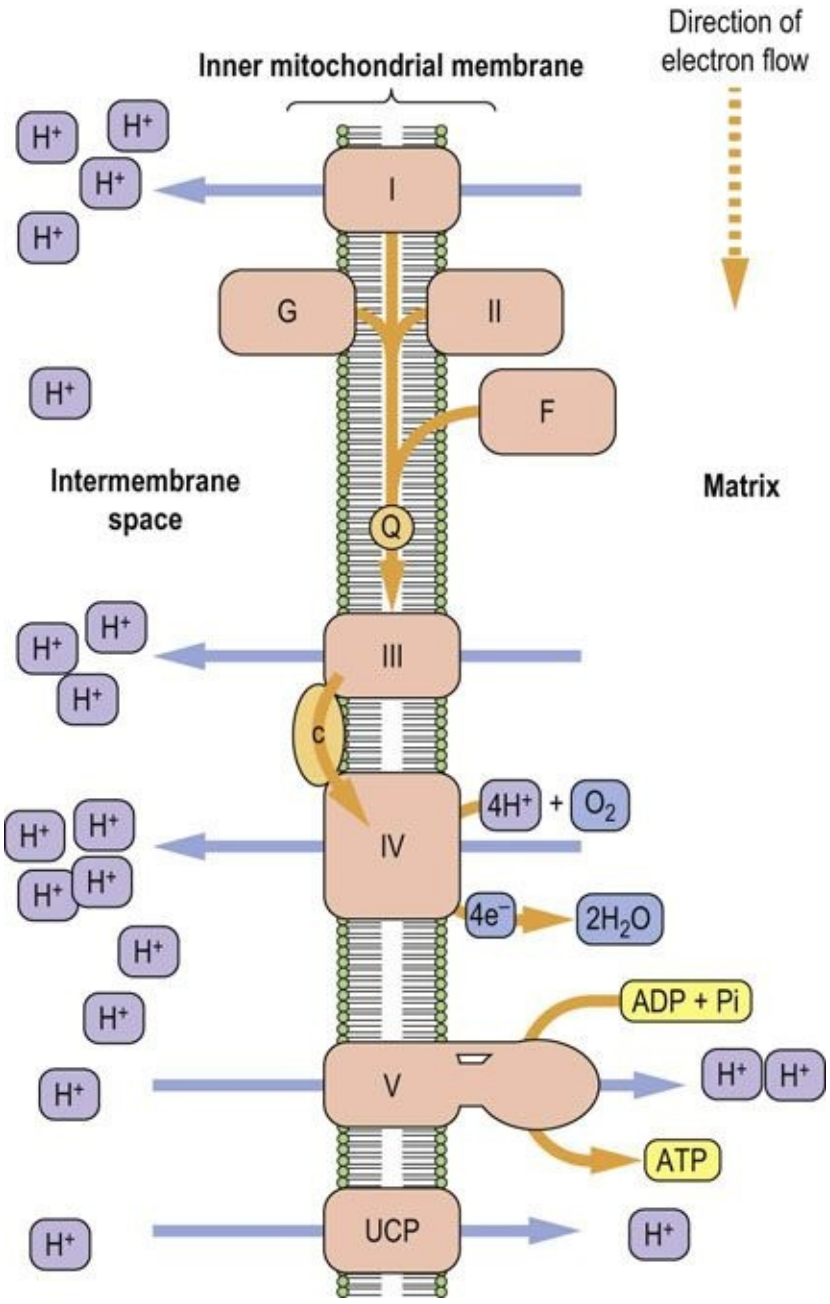


FIG. 9.5 A section of the mitochondrial inner membrane with the electron transport system and ATP synthase.
I, complex I; II, complex II (succinate dehydrogenase); III, complex III; IV, complex IV; V, complex V or ATP synthase; G, glycerol-3-phosphate dehydrogenase; F, fatty acyl CoA dehydrogenase; Q, ubiquinone; *c*, cytochrome *c*; UCP, uncoupling protein.

Electrons are funneled into the electron transport chain by several flavoproteins

There are four flavoproteins in the electron transport chain: Complex I contains FMN and the other three contain FAD. These flavoproteins all reduce the small, lipophilic molecule **ubiquinone (Q or coenzyme Q_{10})**, at the beginning of the common electron transport pathway, consisting of Q, complex III, cytochrome *c*, and complex IV.

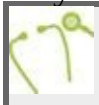
$$\text{Flavoprotein}_{(\text{reduced})} + \text{Q} \rightarrow \text{Flavoprotein}_{(\text{oxidized})} + \text{QH}_2$$

Protons are pumped from the matrix into the intermembrane space by complexes I, III, and IV. Oxygen (O_2) is the final electron acceptor at the end of the chain, and it is reduced to two water molecules by the transfer of four electrons from complex IV and four protons from the mitochondrial matrix compartment.

The efficiency of oxidative phosphorylation can be measured by dividing the amount of phosphate incorporated into ADP by the amount of atomic oxygen reduced. One atom of oxygen is reduced by two electrons (one electron pair).

$$\text{ADP} + \text{Pi} + \tfrac{1}{2}\,\text{O}_2 + 2\text{H}^+ + 2\text{e}^- \rightarrow \text{ATP} + \text{H}_2\text{O}$$

For each pair of electrons transported through complexes I, III or IV, a sufficient number of protons is pumped by each complex for the synthesis of approximately one mole of ATP/complex. If electron transport begins with an electron pair from NADH, approximately 2.5 moles of ATP are synthesized, whereas an electron pair from any of the other three $FADH_2$-containing flavoproteins yields about 1.5 moles of ATP, because the proton-pumping capability of complex I is bypassed.



Clinical box Iron deficiency leads to anaemia

A 45-year-old woman complains of tiredness and appears pale. She is a vegetarian and is experiencing a monthly menstrual flow that is heavy and prolonged. Her hematocrit is 0.32 (reference range 0.36–0.46) and her hemoglobin concentration 90 g/L (normal range 120–160 g/L; 12–16 g/dL).

Comment.

Iron deficiency anemia is a common nutritional problem and is especially common in menstruating and pregnant women because of their increased dietary requirement for iron. Men require about 1 mg iron/day, menstruating women about 2 mg/day, and pregnant women about 3 mg/day. Iron is required to maintain normal amounts of hemoglobin, the cytochromes, and iron–sulfur complexes that are central to oxygen transport and energy metabolism. All these processes are impaired in iron deficiency. Heme iron, which is found in meats, is absorbed much more readily than inorganic iron such as that found in egg yolks, vegetables, and nuts. For hematology reference values, see Table 5.2.

Flavoproteins contain FAD or FMN prosthetic groups

Complex I, also called NADH–Q reductase or NADH dehydrogenase, is a flavoprotein containing FMN. It oxidizes mitochondrial NADH, and transfers electrons through FMN and iron–sulfur (FeS) complexes to ubiquinone, providing enough energy to pump four protons from the matrix in the reaction:

$$NADH + Q + 5H^+_{\text{matrix}} \rightarrow NAD^+ + QH_2 + 4H^+_{\text{intermembrane space}}$$

Three other flavoproteins transfer electrons from oxidizable substrates via $FADH_2$ to ubiquinone (Q) (see Fig. 9.5):

■ Succinate – Q reductase (complex II or succinate dehydrogenase of the TCA cycle) (see Chapter 14) oxidizes succinate to fumarate and reduces FAD to $FADH_2$.

■ Glycerol-3-phosphate – Q reductase, a part of the glycerol-3-P shuttle (see below), oxidizes cytoplasmic glycerol-3-P to dihydroxyacetone phosphate (DHAP) and reduces FAD to $FADH_2$.

■ Fatty acyl CoA dehydrogenase catalyzes the first step in the mitochondrial oxidation of fatty acids and also produces $FADH_2$.

Both FMN and FAD contain the water-soluble vitamin riboflavin. A dietary deficiency of riboflavin can severely impair the function of these and other flavoproteins.

Clinical box A rare coenzyme Q_{10} deficiency

A 4-year-old boy presented with seizures, progressive muscle weakness, and encephalopathy. Accumulation of lactate, a product of anaerobic metabolism of glucose, in the cerebrospinal fluid (CSF) suggested a defect in mitochondrial oxidative metabolism. Muscle mitochondria were isolated for study. The activities of the individual complexes I, II, III, and IV were normal, but the combined activities of I + III and II + III were significantly decreased. Treatment with coenzyme Q_{10} improved the muscle weakness, but not the encephalopathy.

Comment.

Severe muscle weakness, encephalopathy, or both, may be caused in so-called mitochondrial myopathies by mitochondrial defects involving the electron transport system. The finding of increased lactate in the CSF suggests a defect in oxidative phosphorylation. The decreased activities of complexes I + III and II + III suggested a deficiency in coenzyme Q_{10}, which was confirmed by direct measurements.

Advanced concept box Iron–sulfur complexes

Iron–sulfur complexes participate in redox reactions.

Iron is an important constituent of heme proteins, such as

hemoglobin, myoglobin, cytochromes, and catalase, but it is also associated with iron–sulfur (FeS) complexes or nonheme iron proteins that function as electron transporters in the mitochondrial electron transport system. The Fe_2S_2 and Fe_4S_4 types are shown in Figure 9.6. In each case, the iron–sulfur center is bound to a peptide through cysteine residues. The FeS complexes undergo reversible distortion and relaxation during redox reactions. The redox energy is said to be conserved in the 'conformational energy' of the protein.

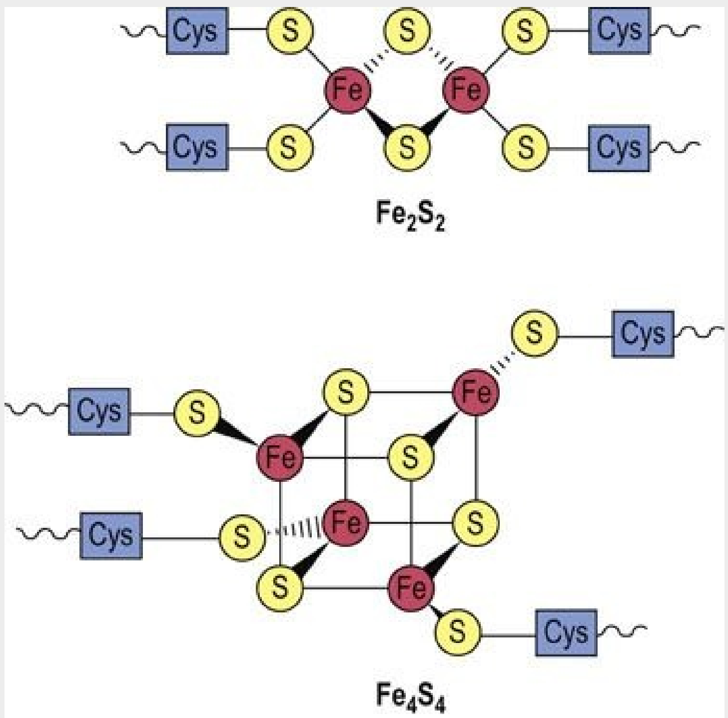


FIG. 9.6 Iron–sulfur complexes.
Cys, cysteine.

Transfer of electrons from NADH into mitochondria

Electron shuttles

Electron shuttles are required for mitochondrial oxidation of NADH produced in the cytoplasmic compartment

NADH is produced in the cytosol during carbohydrate metabolism. NADH cannot cross the inner mitochondrial membrane, and therefore it cannot be oxidized by the electron transport system. Two redox shuttles permit the oxidation of cytosolic NADH without its physical transfer into the mitochondrion. A characteristic feature of these shuttles is that they are powered by cytoplasmic and mitochondrial isoforms of the same enzyme, which catalyze opposing reactions on opposite sides of the membrane. The glycerol-3-P shuttle is the simpler of the two (Fig. 9.7, top). It transfers the electrons of NADH from the cytoplasm to the mitochondrion by reducing FAD to $FADH_2$. Cytoplasmic glycerol-3-P dehydrogenase catalyzes reduction of dihydroxyacetone-P (DHAP) with NADH to glycerol-3-P, regenerating NAD^+. The cytoplasmic glycerol-3-phosphate is oxidized back to DHAP by another glycerol-3-phosphate dehydrogenase isoform facing the outer surface of the inner mitochondrial membrane; this enzyme is a flavoprotein in which FAD is reduced to $FADH_2$. The electrons are then transferred to the common pathway via ubiquinone. Because the electrons are transferred to FAD, the yield of ATP from cytoplasmic NADH by this pathway is approximately 1.5 moles, rather than the 2.5 moles available from mitochondrial NADH via the NADH-Q reductase complex (complex I).

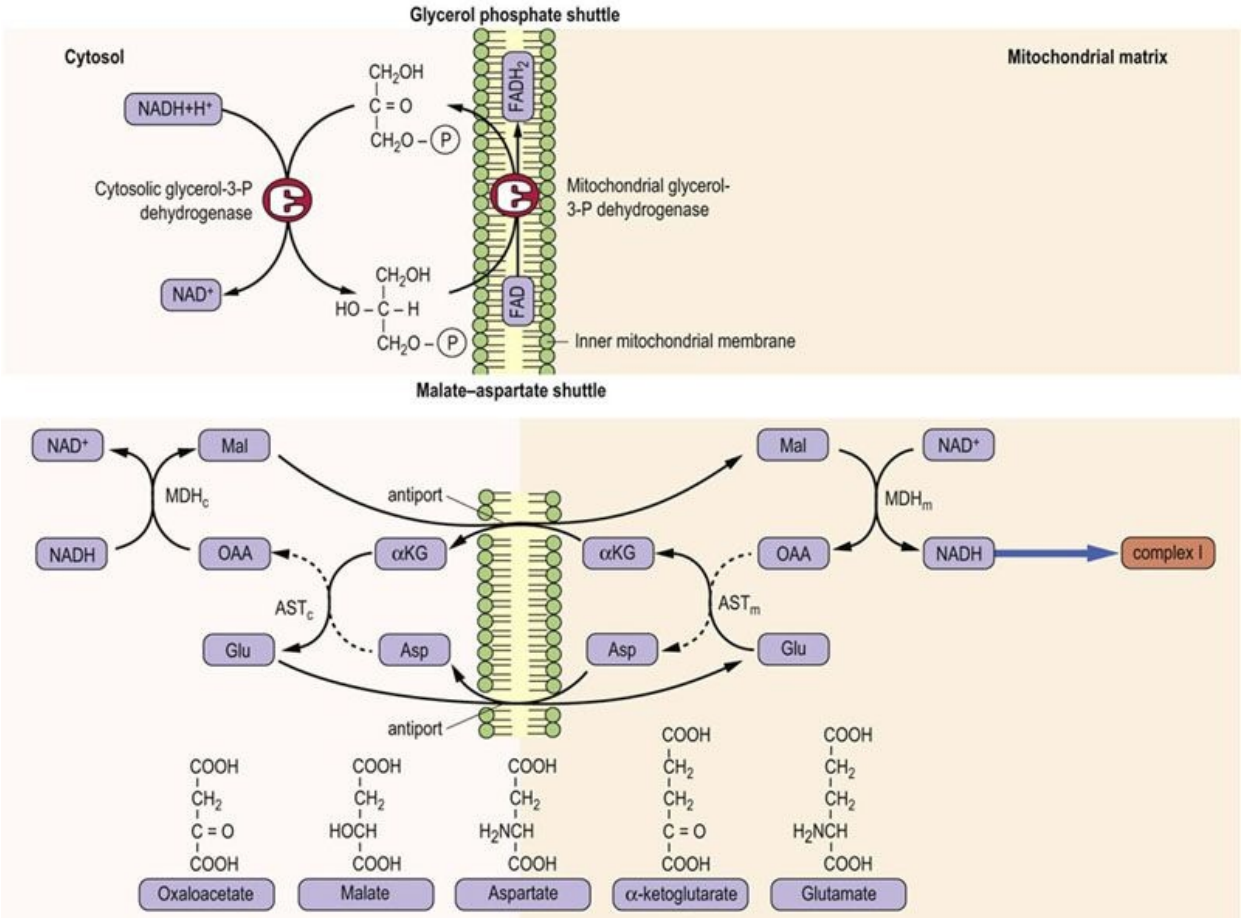


FIG. 9.7 Redox shuttles in the inner mitochondrial membrane. (Top) The glycerol phosphate shuttle. (Bottom) The malate–aspartate shuttle MDH, malate dehydrogenase; AST, aspartate aminotransferase. Subscripts c and m refer to cytosolic and mitochondrial isozymes.

Many cells, *e.g.* in skeletal muscle, use the glycerol-3-P shuttle, but heart and liver rely on the malate–aspartate shuttle (Fig. 9.7, bottom), which yields 2.5 moles of ATP per mole of NADH. This shuttle is more complicated, because the substrate, malate, can cross the inner mitochondrial membrane, but the membrane is impermeable to the product, oxaloacetate – there is no oxaloacetate transporter. The exchange is therefore accomplished by interconversion between α-keto- and α-amino acids, involving cytoplasmic and mitochondrial glutamate and α-ketoglutarate, and isozymes of glutamate-oxaloacetate transaminase (aspartate aminotransferase).

Ubiquinone (coenzyme Q_{10})

Ubiquinone tranfers electrons from flavoproteins to complex III

Ubiquinone is so named because it is ubiquitous in virtually all living systems. It is a small, lipid-soluble compound found in the inner membrane of animal and plant mitochondria and in the plasma membrane of bacteria. The primary form of mammalian ubiquinone contains a side chain of 10 isoprene units and is often called CoQ_{10}. It diffuses within the inner membrane, accepts electrons from the four major mitochondrial flavoproteins, and transfers them to complex III (QH_2–cytochrome *c* reductase). Ubiquinone can carry either one or two electrons (Fig. 9.8) and is also thought to be a major source of superoxide radicals in the cell (see Chapter 37).

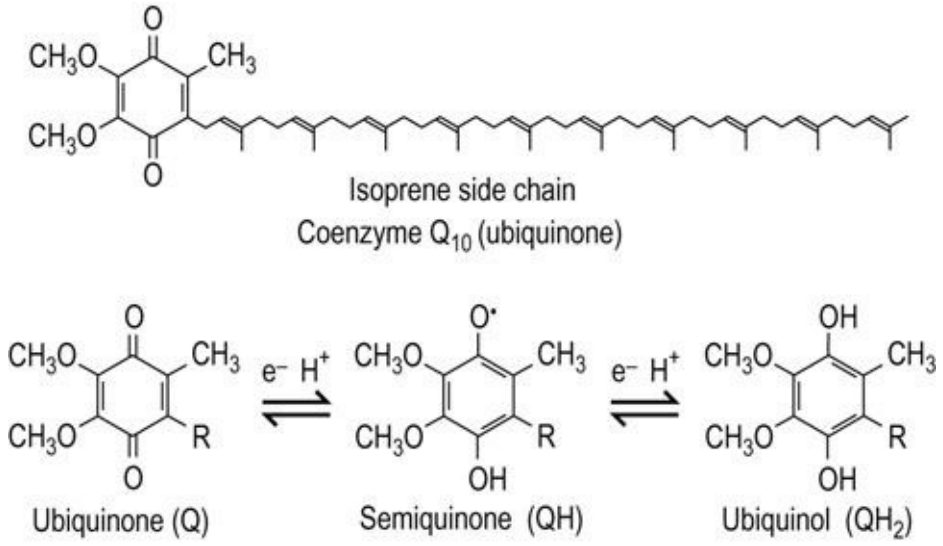


FIG. 9.8 Coenzyme Q10, or ubiquinone, accepts one or two electrons, transferring them from flavoproteins to complex III. The semiquinone form is a free radical.

Complex III – cytochrome *c* reductase

Complex III accepts electrons from ubiquinone and pumps four hydrogen ions across the inner mitochondrial membrane

This enzyme complex, also known as ubiquinone–cytochrome *c* reductase or QH_2–cytochrome *c* reductase, oxidizes ubiquinone and reduces cytochrome *c*. Reduced ubiquinone funnels electrons that it gathers from mitochondrial flavoproteins and transfers them to complex III. Electrons from ubiquinone are transferred through two species of cytochrome *b*, to an FeS center, to cytochrome c_1, and finally to cytochrome *c*. Transport of two electrons to cytochrome *c* yields sufficient free energy change and protons pumped to synthesize about one mole of ATP. The overall reaction is:

$$QH_2 + 2\text{cyt c}_{\text{oxidized}} + 2H^+_{\text{matrix}} \rightarrow$$
$$2Q + 2\text{cyt c}_{\text{reduced}} + 4H^+_{\text{intermembrane space}}$$

Four protons are pumped during this reaction, two from fully reduced ubiquinone and two from the matrix.

Advanced concept box

Cytochromes

Cytochromes, found in the mitochondrion and endoplasmic reticulum, are proteins that contain heme groups, but which are not involved in oxygen transport (Fig. 9.9). The core structure of these heme groups is a tetrapyrrole ring similar to that of hemoglobin, sometimes differing only in the composition of the side chains. The heme group of cytochromes *b* and c_1 is known as iron protoporphyrin IX and is the same heme that is found in hemoglobin, myoglobin, and catalase. Cytochrome *c* contains heme C that is covalently bound to the protein through cysteine residues. Cytochromes *a* and a_3 contain heme A, which, in common with ubiquinone, contains a hydrophobic isoprene side chain. In hemoglobin and myoglobin, heme must remain in the ferrous (Fe^{2+}) state; in cytochromes, the heme iron is reversibly reduced and oxidized between the Fe^{2+} and Fe^{3+} states as electrons are shuttled from one protein to another.

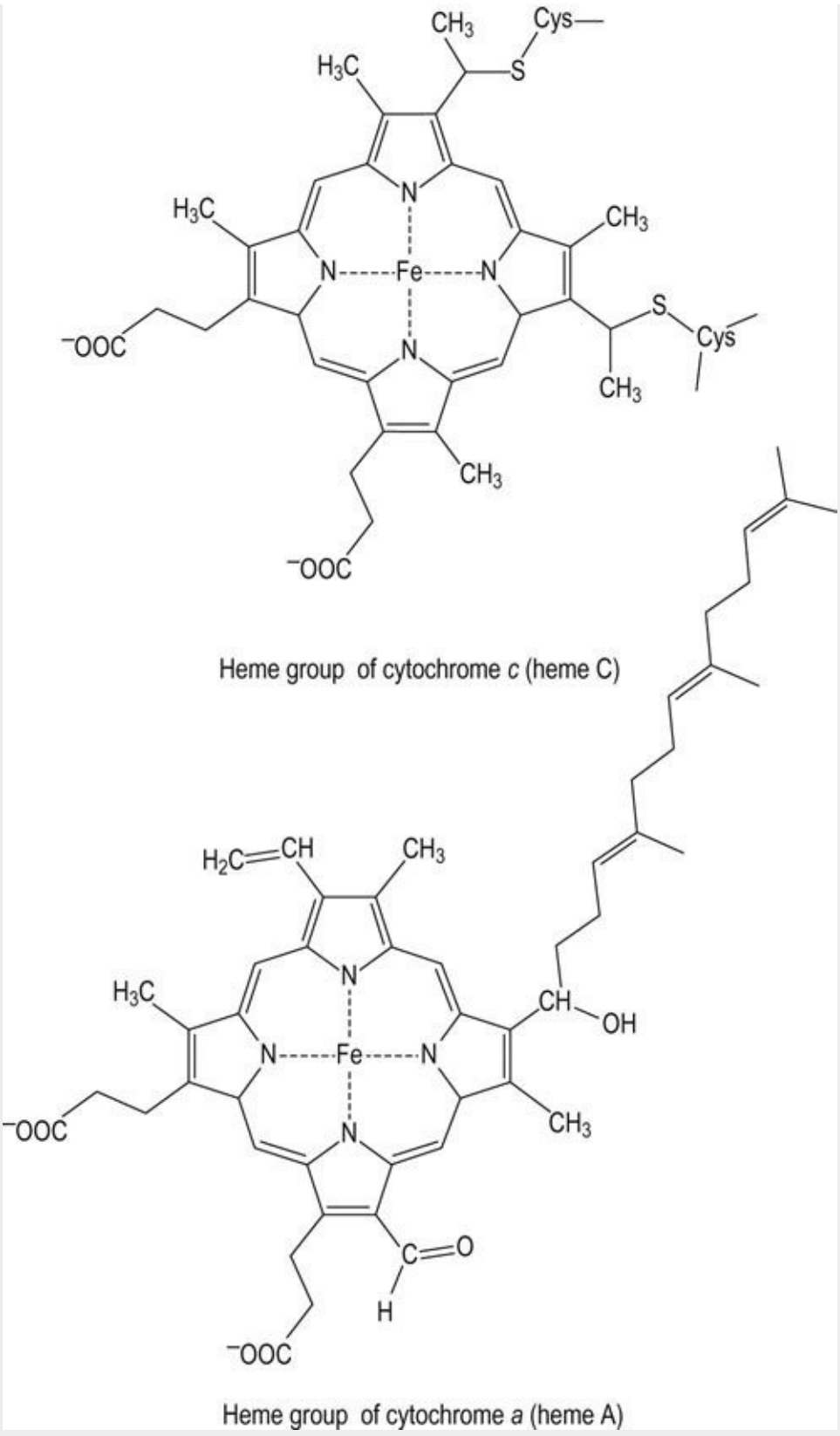


FIG. 9.9 Variations in heme structures among cytochromes.

The cytochromes are proteins that contain heme groups.

Cytochrome *c*

Cytochrome c is a peripheral membrane proteins, shuttling electrons from complex III to complex IV

Cytochrome *c*, a small heme protein that is loosely bound to the outer surface of the inner membrane, shuttles electrons from complex III to complex IV. Each cytochrome *c* carries only one electron, so the reduction of O_2 to $2H_2O$ by complex IV requires four reduced cytochrome *c* molecules. The binding of cytochrome *c* to complexes III and IV is largely electrostatic, involving a number of lysine residues on the protein surface. Reduction of ferricytochrome *c* (Fe^{3+}) to ferrocytochrome *c* (Fe^{2+}) by cytochrome c_1 leads to a change in the three-dimensional structure, charge distribution and dipole moment of the protein, promoting transfer of electrons to cytochrome *a* in complex IV (see Fig. 9.5). In response to oxidative stress and cell injury (Chapter 37), cytochrome *c* may be released from the inner mitochondrial membrane and leak into the cytosol, inducing apoptosis (cell death).

Clinical box Copper deficiency in neonates

Copper is required in trace amounts for optimal human nutrition. Although copper deficiency is rare in adults, premature infants have low stores of copper and may suffer from its deficiency. This may lead to anemia and cardiomyopathy, because of failure to synthesize adequate amounts of cytochrome *c* oxidase and other enzymes, including several cuproenzymes involved in the synthesis of heme.

Comment.

Copper deficiency can impair ATP production by inhibiting the terminal reaction of the electron transport chain, leading to

pathology in the heart, where energy demand is high. Dietary formulas for premature infants must contain adequate copper; cow's milk alone is unsuitable, because it is low in copper.

Complex IV

Complex IV, at the end of the electron transport chain, transfers electrons to oxygen, producing water

Complex IV, known as cytochrome *c* oxidase or cytochrome oxidase, exists as a dimer in the IMM. It oxidizes the mobile cytochrome *c*, and conducts electrons through cytochromes *a* and a_3, finally reducing oxygen to water in a four-electron transfer reaction (Fig. 9.10). Copper is a common component of this and other oxidase enzymes. Small molecule poisons, such as **azide**, **cyanide**, and **carbon monoxide**, bind to the heme group of cytochrome a_3 in cytochrome *c* oxidase and inhibit complex IV. In common with complexes I and III, the cytochrome oxidase complex pumps protons out of mitochondria, providing for the synthesis of about one mole of ATP per pair of electrons transferred to oxygen. The actual number of protons pumped is four. In addition, another four are required in the reduction of O_2 to water. The overall reaction catalyzed by complex IV is:

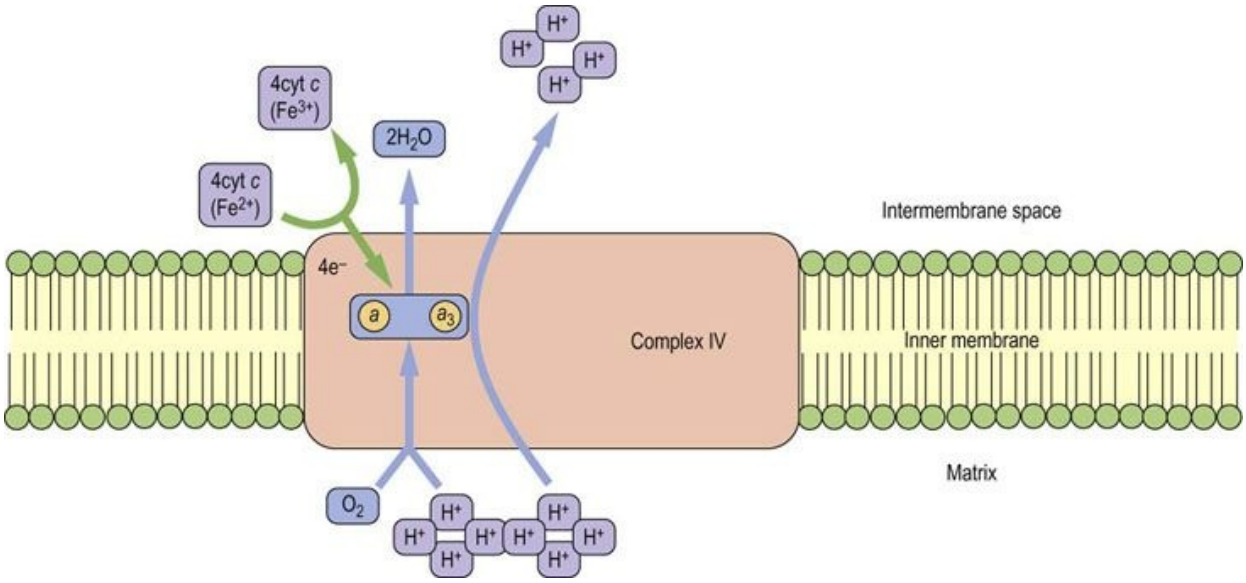


FIG. 9.10 Complex IV.
Complex IV utilizes four electrons from cytochrome *c* and eight protons from the matrix. Four protons and electrons reduce oxygen to water. Four additional protons are pumped out of the matrix. Complex IV is regulated allosterically by ATP, by reversible phosphorylation/dephosphorylation, and by thyroid hormone (T_2 or diiodothyronine). *a*, cytochrome *a*; a_3, cytochrome a_3.

$$4\text{cyt c}_{\text{reduced}} + 4\text{H}^{+}{}_{(\text{matrix})} + \text{O}_2 \rightarrow 4\text{cytc}_{\text{oxidized}} + 2\text{H}_2\text{O}$$

Synthesis of adenosine triphosphate – the chemiosmotic hypothesis

According to the chemiosmotic hypothesis, mitochondria produce ATP using the free energy from the proton gradient generated during oxidation of NADH and $FADH_2$. This energy is described as a **proton motive force**, an **electrochemical gradient** created by the proton concentration gradient and a proton charge differential (outside positive) across the inner mitochondrial membrane. To operate, it requires an inner membrane system that is impermeable to protons, except through ATP synthase or other complexes in a regulated fashion. When protons are pumped out of the matrix, the intermembrane space becomes more acidic and more positively charged than the matrix.

The ATP synthase complex (complex V) is an example of rotary catalysis

Lining the inner matrix face of the inner membrane of each mitochondrion are thousands of copies of the ATP synthase complex, also called complex V or F_0F_1-ATP synthase (F = coupling factor; see Inhibitors of ATP synthase, below). ATP synthase is also called an ATPase, because it can hydrolyze ATP. ATP synthase consists of two major complexes (Fig. 9.11). The inner membrane component, termed F_0, is the proton-driven motor with the stoichiometry of *a*, b_2 and $c_{10–14}$. The *c*-subunits form the *c*-ring, which rotates in a clockwise direction in response to the flow of protons through the complex. Since the γ-and ε-subunits are bound to the *c*-ring, they rotate with it, inducing large conformational changes in the three-αβ dimers of the F_1 complex. The two β-proteins immobilize the second complex (F_1-ATP synthase).

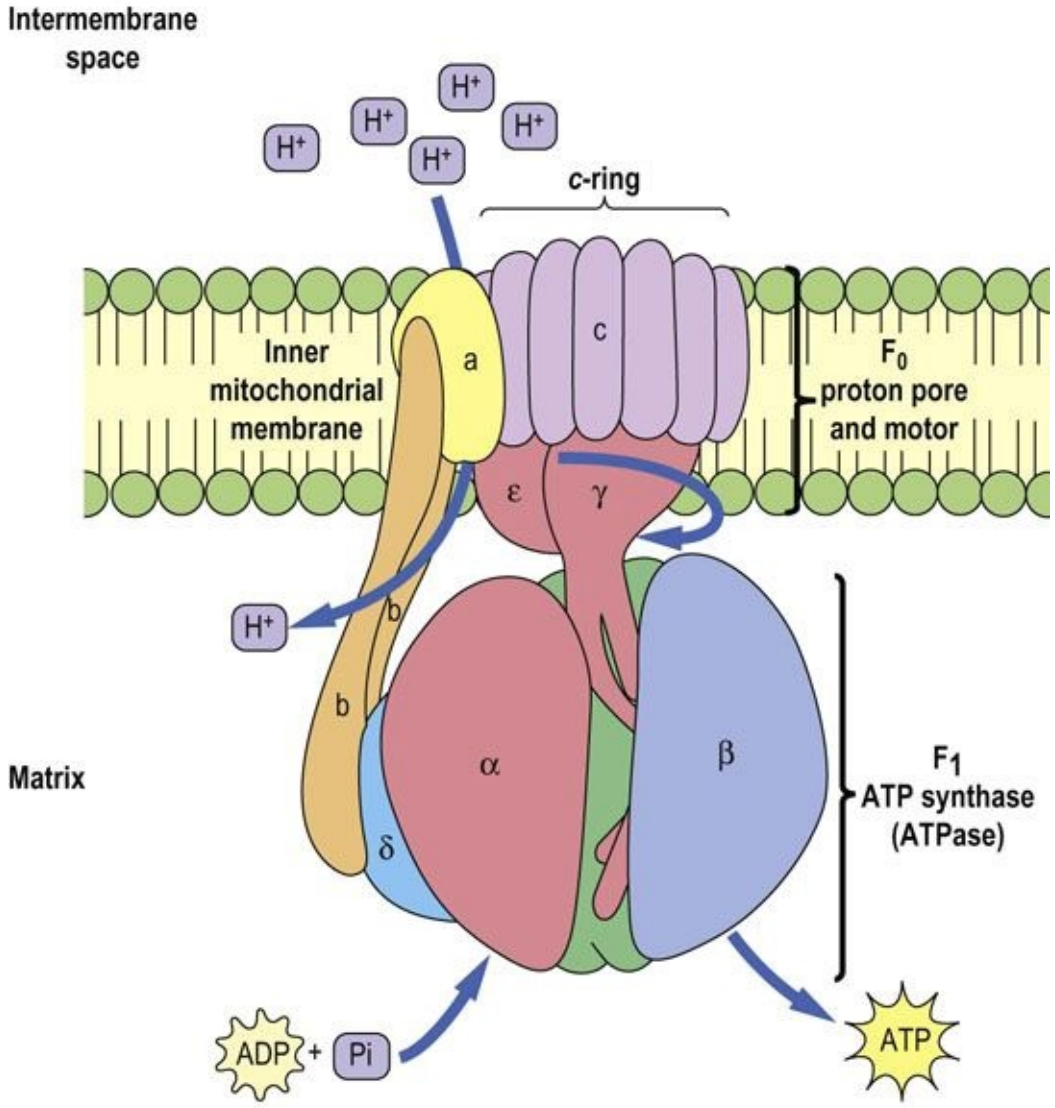


FIG. 9.11 ATP synthase complex.
The ATP synthase complex consists of a motor (F_0) and generator (F_1). The proton pore involves the *c*-ring and the *a*-protein. The rotary component is the coiled-coil γ-subunit, which is bound to the ε-subunit and to the *c*-ring. The stationary component is the hexameric $\alpha_3\beta_3$ unit, which is held in place by the δ, *b* and *a*-proteins.

F_1 has a stoichiometry of α_3, β_3, γ, δ, ε. The major part of F_1 consists of three αβ dimers arranged like slices of an orange, with the catalytic activity residing on the β-subunits. Each 120° rotation of the γ-subunit induces conformational changes in the αβ-dimeric subunits such that the nucleotide-binding sites alternate between three states: the first binds ADP and Pi, the second synthesizes ATP, and the third releases ATP, so each complete turn produces 3ATP. This is

known as the **binding-change mechanism** (Fig. 9.12). Surprisingly, the proton-motive free energy used by ATP synthase is not for ATP synthesis itself, but for its release; when the proton gradient is too low to support ATP release, ATP remains stuck to ATP synthase and further ATP production ceases. ADP and Pi are bound to the complex as soon as ATP leaves. The αβ-dimers are asymmetrical, because each is in a different conformation at any given moment. This complex is a proton-driven motor, and it is an example of rotary catalysis. About three protons are required for the synthesis of each ATP. This complex acts independently of the electron transport chain; addition of a weak acid, such as acetic acid, to a suspension of isolated mitochondria is sufficient to induce the biosynthesis of ATP in vitro.

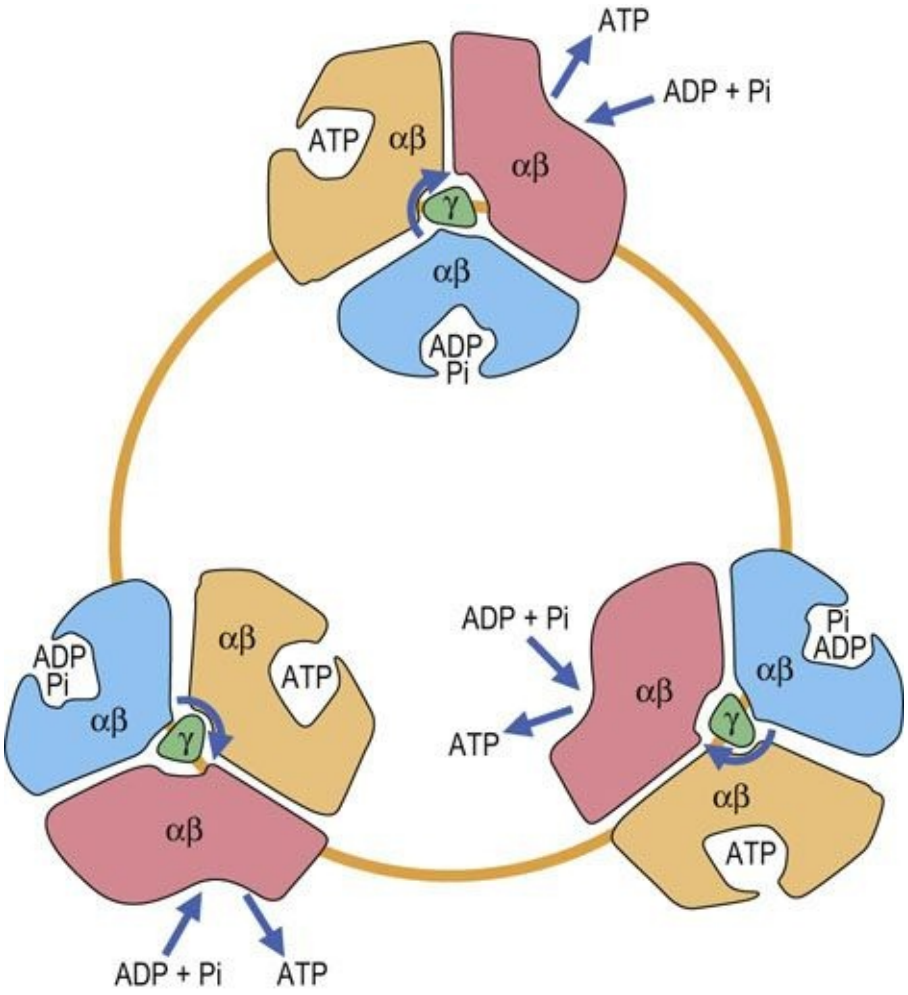


FIG. 9.12 Binding-change mechanism of ATP synthase.
Powered by protons, the rotation of the γ-subunit of ATP synthase induces simultaneous

conformational changes in all three αβ-dimers. Each 120° rotation results in ejection of an ATP, binding of ADP and Pi and ATP synthesis.

P : O ratios

The P : O ratio is a measure of the number of high-energy phosphates (i.e. amount of ATP) synthesized per atom of oxygen ($\frac{1}{2}O_2$) consumed, or per mole of water produced. The P : O ratio can be calculated from the moles of ADP used to synthesize ATP and the atoms of oxygen taken up by mitochondria. For example, if 2 mmol of ADP is converted to ATP and 0.5 mmol of oxygen (1.0 milliatom of oxygen) is taken up, the P : O ratio is 2.0. As discussed earlier, the theoretical yield of ATP per mole of NADH is about 7 moles; however, by actual measurement with isolated mitochondria, **the P : O ratio for oxidation of metabolites that yield NADH is about 2.5 and the ratio for those that yield $FADH_2$ is about 1.5**. The remainder of the energy available from the oxidation of NADH and $FADH_2$ is released in the form of heat.

‘Respiratory control’ is the dependence of oxygen uptake by mitochondria on the availability of ADP

Normally, oxidation and phosphorylation are tightly coupled: substrates are oxidized, electrons are transported, and oxygen is consumed only when synthesis of ATP is required (coupled respiration). Thus, resting mitochondria consume oxygen at a slow rate, which can be greatly stimulated by addition of ADP (Fig. 9.13). ADP is taken up by the mitochondria and stimulates ATP synthase, which lowers the proton gradient. Respiration increases, because the proton pumps are stimulated to reestablish the proton gradient. When the ADP is depleted, ATP synthesis terminates and respiration returns to the original rate. Oxygen uptake declines to the original rate when the concentration of ADP is depleted and ATP synthesis terminates.

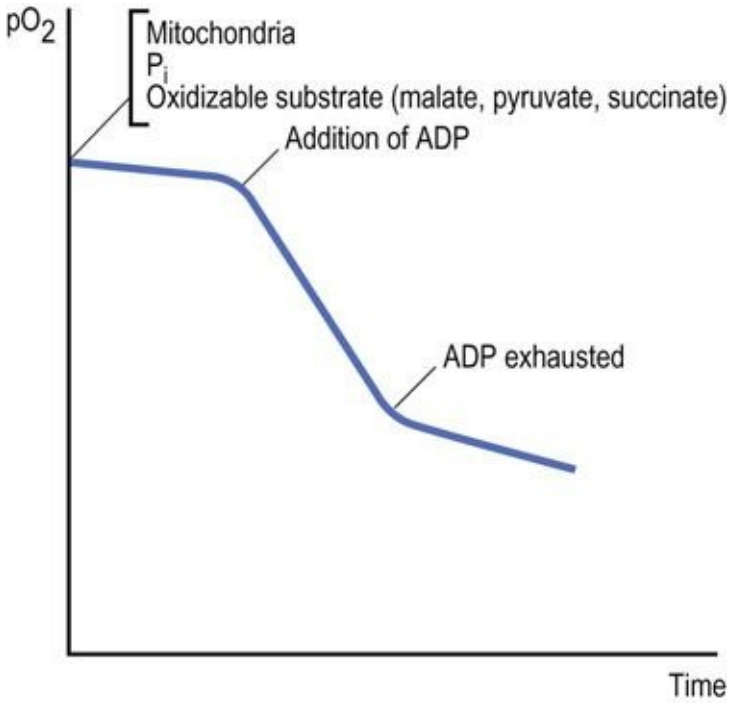


FIG. 9.13 Effect of ADP on the uptake of oxygen by isolated mitochondria. This may be studied in an isolated (sealed) system with an oxygen electrode and a recording device. The graph shows a typical recording of oxygen consumption (pO_2, partial pressure of oxygen) by normal mitochondria on introduction of ADP.

Mitochondria can become partially uncoupled if the inner membrane loses its structural integrity. They are said to be ‘leaky’, because protons can diffuse through the inner membrane without involving ATP synthase. This occurs if isolated mitochondria are treated with mild detergents that disrupt the inner membrane, or if they have been stored for a period of time. Such mitochondria are said to be ‘uncoupled’; oxidation proceeds without production of ATP and uncoupled mitochondria lose respiratory control because protons pumped by the electron chain bypass the ATPase and leak unproductively back into the matrix. The P : O ratio declines under these conditions.

The mechanism of respiratory control depends on the requirement for ADP and Pi binding to the ATP synthase complex: in the absence of ADP and Pi, protons cannot enter the mitochondrion through this complex and oxygen consumption markedly decreases, because the proton pumps cannot pump protons against a high proton back-pressure. This happens because the free energy of the electron transport reactions is sufficient to generate a pH gradient of only two units across the membrane. If the pH gradient cannot be discharged for production of ATP, the two pH unit gradient is established and the pumps

grind to a halt and stall. The electron transport chain becomes reduced and substrate oxidation and oxygen consumption decrease. A little physical activity, with consumption of ATP and generation of ADP and Pi, opens up the ATPase channels, discharging the proton gradient and activating the electron transport chain, and fuel and oxygen consumption. At a whole-body level, we breathe faster during exercise to provide the additional oxygen needed for increased oxidative phosphorylation.

Uncouplers

Uncouplers and uncoupling proteins are thermogenic

Uncouplers of oxidative phosphorylation dissipate the proton gradient by transporting protons back into mitochondria, bypassing the ATP synthase. Uncouplers stimulate respiration, because the system makes a futile attempt to restore the proton gradient by oxidizing more fuel and pumping more protons out of mitochondria. **Uncouplers are typically hydrophobic compounds and either weak acids or bases, with pK_a near pH 7.** The classic uncoupler, 2,4-dinitrophenol (DNP) (Fig. 9.14), is protonated in solution on the outer, more acidic side of the inner mitochondrial membrane. Because of its hydrophobicity, it may then freely diffuse through the inner mitochondrial membrane. When it reaches the matrix side, it encounters a more basic pH and the proton is released, effectively discharging the pH gradient. Other uncouplers include preservatives and antimicrobial agents, such as pentachlorophenol and *p*-cresol.

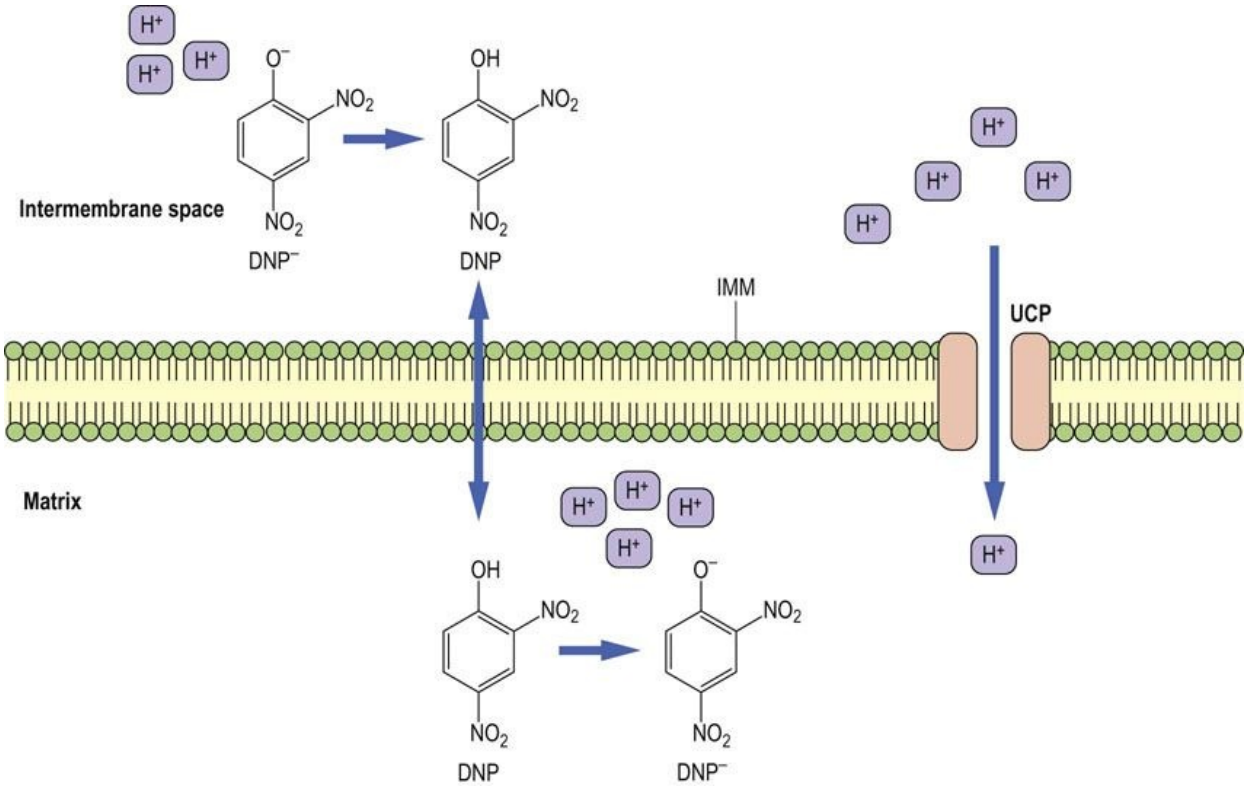


FIG. 9.14 Proton transport by uncouplers.
Uncouplers transport protons into the mitochondrion, dissipating the proton gradient. DNP is an example of an exogenous uncoupler. The uncoupling proteins (UCP) are endogenous uncouplers in the IMM and are regulated by hormones. The gradient consisting of protons and other factors constitute the mitochondrial membrane potential (MMP), which is expressed in millivolts (mV). DNP, 2,4-dinitrophenol; IMM, inner mitochondrial membrane.

Uncoupling proteins (UCP)

According to the chemiosmotic hypothesis, the inner mitochondrial membrane is topologically closed. However, it has long been known that protons are transported into the matrix from the intermembrane space by routes other than the ATP synthase complex and inner membrane transporters. Much of the BMR is now thought to be mainly due to inner membrane components called uncoupling proteins (UCP). The first discovered was uncoupling protein-1 (UCP1), formerly known as **thermogenin**, which is found exclusively in brown adipose tissue. **Brown adipose tissue** is abundant in the newborn and in some adult mammals, and it is brown because of its high content of mitochondria. In humans, brown adipose tissue is abundant in infants, but it gradually diminishes

and is barely detectable in adults. UCP1 provides body heat during cold stress in the young and in some adult animals. It accomplishes this by uncoupling the proton gradient, thereby generating heat (thermogenesis) instead of ATP. Uncoupling proteins are expressed at high levels in hibernating animals, permitting them to maintain body temperature without movement or exercise.

Four additional uncoupling proteins are expressed by the human genome: UCP2, UCP3, UCP4 and UCP5. While UCP1 is exclusive to brown adipose tissue, UCP2 is expressed ubiquitously, UCP3 is mainly expressed in skeletal muscle, and UCP4 and UCP5 are expressed in the brain. Except for UCP1, the physiologic functions of these proteins are not well understood, but could be of profound significance in our understanding of such health issues as diabetes, obesity, cancer, thyroid disease and aging. As uncouplers, they have been linked to a number of fundamental functions. For example, there is strong evidence that obesity induces the synthesis of UCP2 in β-cells of the pancreas. This may play a role in the β-cell dysfunction found in type 2 diabetes, because it lowers the intracellular concentration of ATP, which is required for secretion of insulin. The thyroid hormone (T_3) has been shown to stimulate thermogenesis in rats by promoting the synthesis of UCP3 in skeletal muscle. The common fever that is induced by infectious organisms is probably also due to uncoupling by the UCP, but the mechanism is unknown.

Inhibitors of oxidative metabolism

Electron transport system inhibitors

Inhibitors of electron transport selectively inhibit complexes I, III or IV, interrupting the flow of electrons through the respiratory chain. This stops proton pumping, ATP synthesis, and oxygen uptake. Several inhibitors are readily available poisons that could be encountered in the practice of medicine. It is noteworthy that genetic defects in respiratory chain components often mimic the effects of these inhibitors.

Rotenone inhibits complex I (NADH–Q reductase)

Rotenone, a common insecticide, and some barbiturates (e.g. amytal) inhibit complex I. Because malate and lactate are oxidized by NAD^+, their oxidation will be decreased by rotenone. However, substrates yielding $FADH_2$ can still be oxidized, because complex I is bypassed and electrons are donated to ubiquinone from $FADH_2$. Addition of ADP to a suspension of mitochondria supplemented with malate and phosphate (Fig. 9.15) markedly stimulates oxygen uptake as ATP synthesis occurs. Oxygen uptake is markedly inhibited by rotenone, but when succinate is added, ATP synthesis and oxygen consumption resume until the supply of ADP is exhausted. Rotenone inhibition of complex I causes reduction of all components prior to the point of inhibition, because they cannot be oxidized, whereas those after the point of inhibition become fully oxidized. This is known as a crossover point, and it can be determined spectrophotometrically, because light absorption by respiratory chain components changes according to redox state. Such analyses were used to define the sequence of components in the respiratory chain.

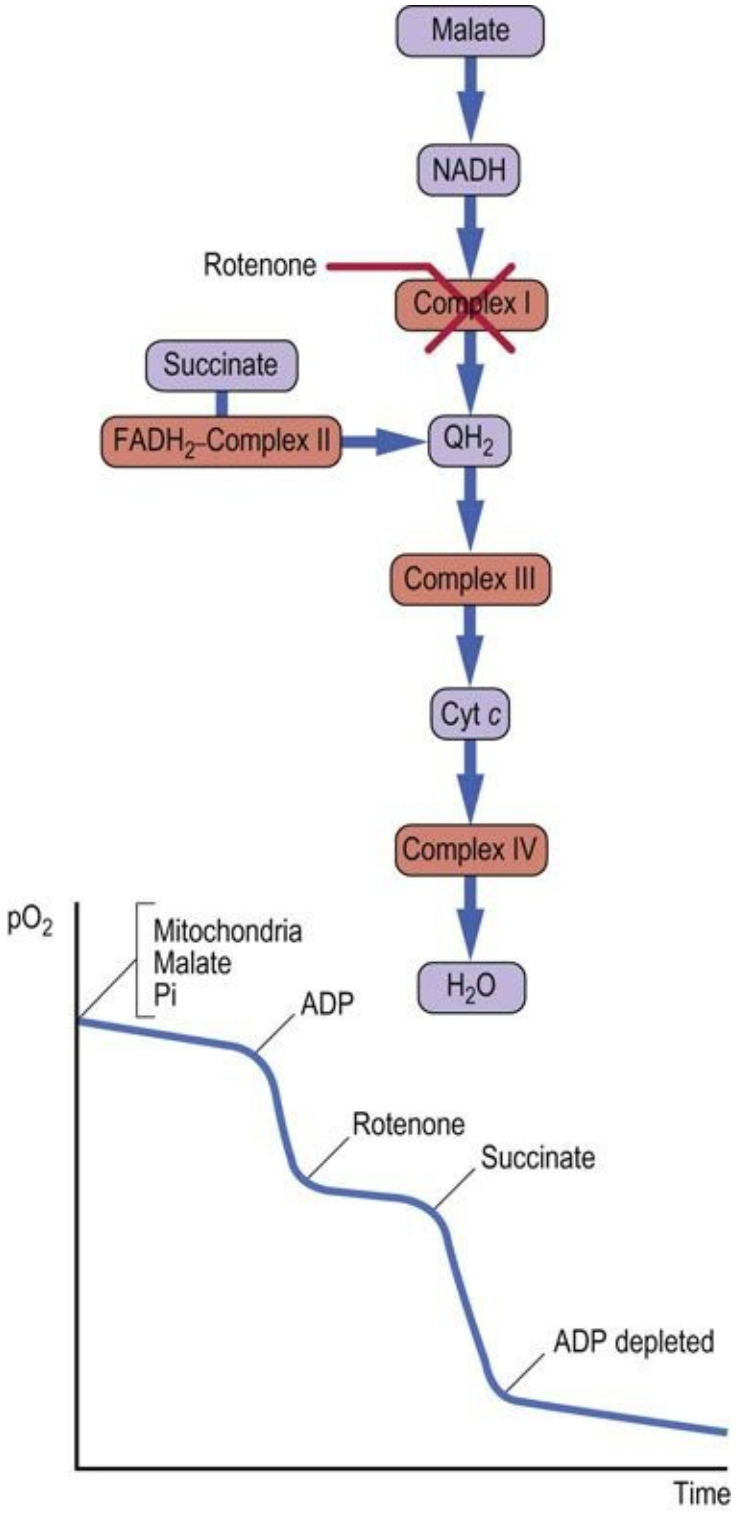


FIG. 9.15 Inhibition of complex I.
Inhibitors such as rotenone retard oxygen uptake by mitochondria when NADH-producing substrates are being oxidized.

Antimycin A inhibits complex III (QH_2–cytochrome c reductase)

The inhibition of complex III by antimycin A prevents transfer of electrons from either complex I or $FADH_2$-containing flavoproteins to cytochrome *c*. In this case, components preceding complex III become fully reduced, and those after it become oxidized. The oxygen uptake curve (Fig. 9.16) shows that the stimulation of respiration by ADP is inhibited by antimycin A, but that the addition of succinate does not relieve the inhibition. Ascorbic acid can reduce cytochrome *c*, and addition of ascorbic acid restores respiration, illustrating that complex IV is unaffected by antimycin A.

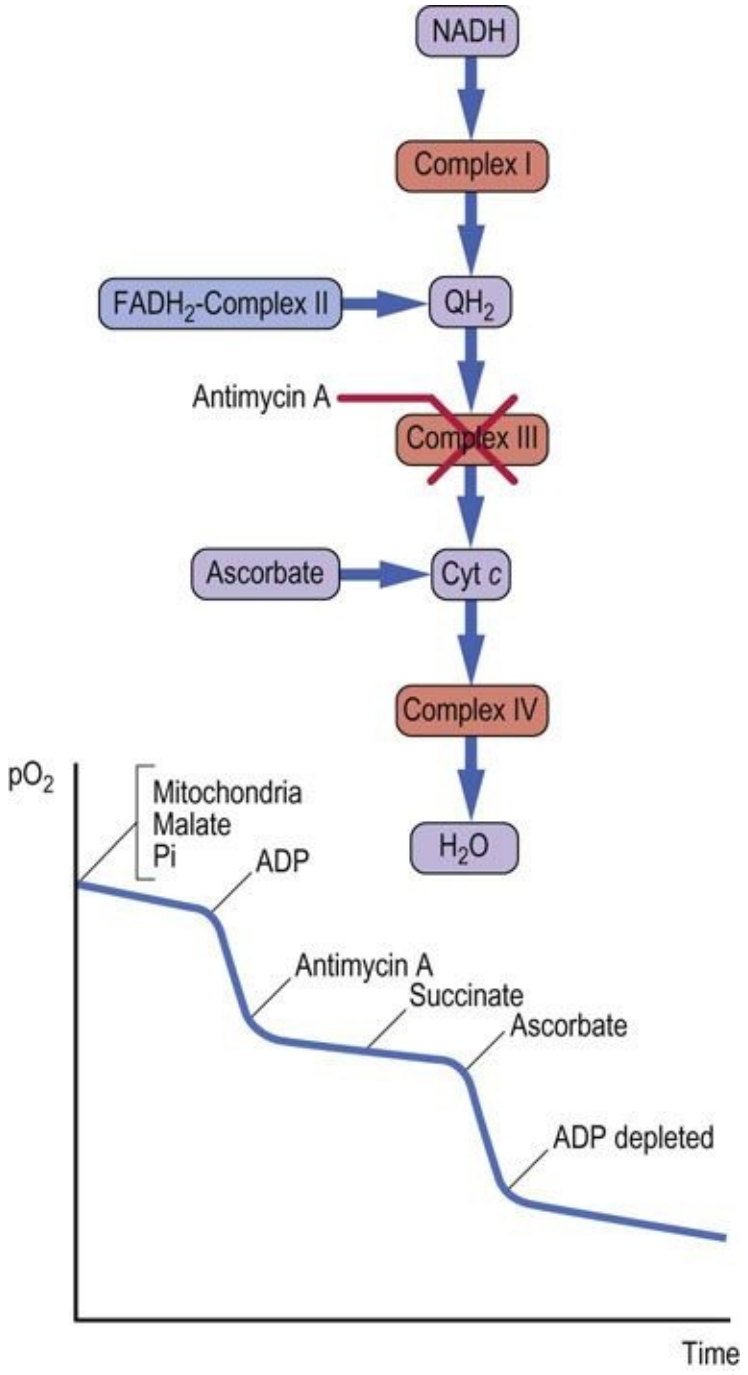


FIG. 9.16 Inhibition of complex III by antimycin.
Antimycin A inhibits complex III, blocking transfer of electrons from both complex I and flavoproteins, such as complex II.

Cyanide and carbon monoxide inhibit complex IV

Azide (N_3^-), cyanide $(CN)^-$; and carbon monoxide (CO) inhibit complex IV (cytochrome *c* oxidase) (Fig. 9.17). Because complex IV is the terminal electron transfer complex, its inhibition cannot be bypassed. All components preceding complex IV become reduced, oxygen cannot be reduced, none of the complexes can pump protons, and ATP is not synthesized. Uncouplers such as DNP have no effect, because there is no proton gradient. Cyanide and carbon monoxide also bind to hemoglobin, blocking oxygen binding and transport (see Chapter 5). In these poisonings, both the ability to transport oxygen and to synthesize ATP are impaired. The administration of oxygen is used for the treatment of such poisonings. It is of interest that sodium azide is the nitrogen source for the inflation of air bags; it may pose an environmental problem if it is accidentally released, *i.e.* nonexplosively.

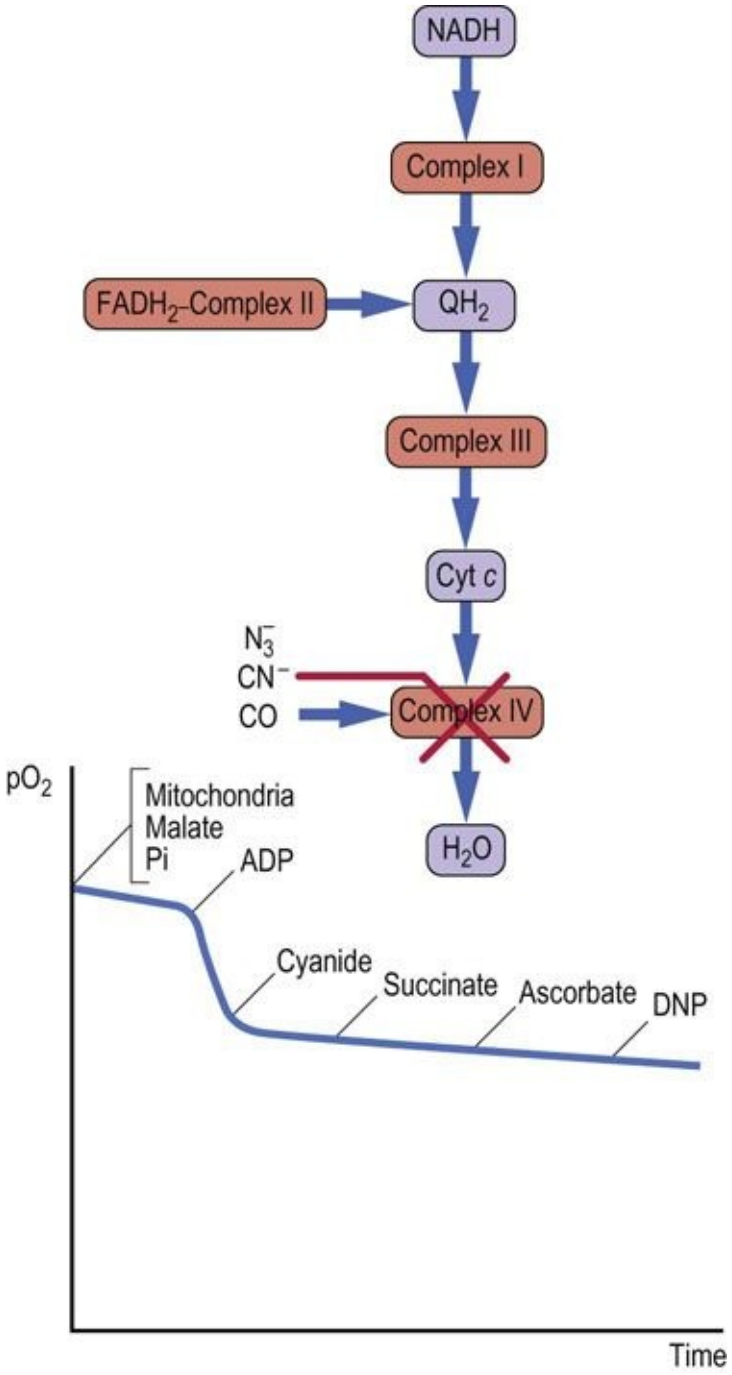


FIG. 9.17 Inhibition of complex IV.
The inhibition of complex IV interrupts the transfer of electrons, in the final step of electron transport. Electrons cannot be transferred to oxygen, and the synthesis of ATP is halted.

Oligomycin inhibits ATP synthase

Oligomycin inhibits respiration but, in contrast to electron transport inhibitors, it is not a direct inhibitor of the electron transport system. Instead, it inhibits the proton channel of ATP synthase. It causes an accumulation of protons outside the mitochondrion, because the proton pumping system is still intact but the proton channel is blocked. The addition of the uncoupler DNP after oxygen uptake has been inhibited by oligomycin illustrates this point: DNP dissipates the proton gradient and stimulates oxygen uptake as the electron transport system attempts to reestablish the proton gradient (Fig. 9.18).

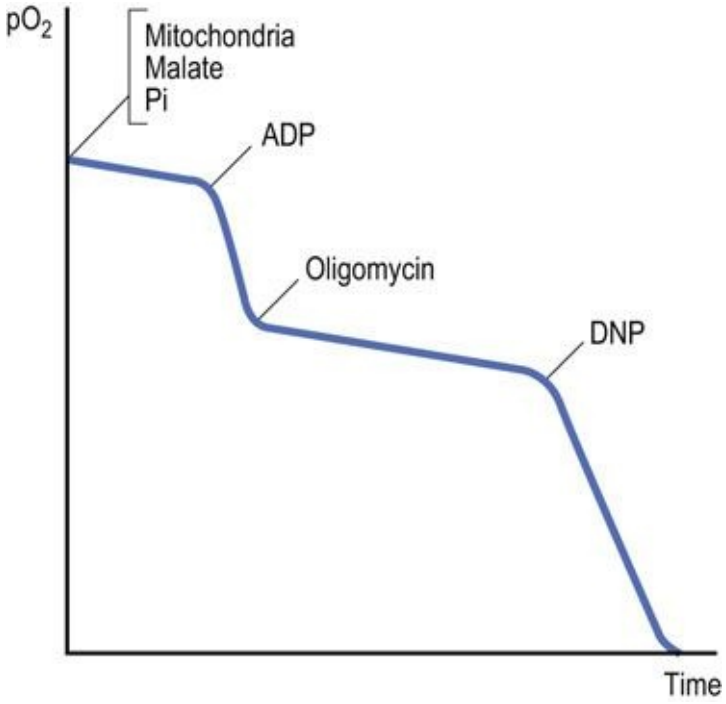


FIG. 9.18 Oligomycin inhibition of oxygen uptake.
Oligomycin inhibits oxygen uptake in ATP-synthesizing mitochondria. Oligomycin inhibits ATP synthase and oxygen uptake in coupled mitochondria. However, DNP stimulates oxygen uptake after oligomycin inhibition, by dissipating the proton gradient.

Inhibitors of the ADP–ATP translocase

Most ATP is synthesized in the mitochondrion, but used in the cytosol for biosynthetic reactions. Newly synthesized mitochondrial ATP and spent cytosolic ADP are exchanged by a mitochondrial ADP–ATP translocase, representing about 10% of the protein in the inner mitochondrial membrane (see

Fig. 9.3). This translocase can be inhibited by unusual plant and mold toxins, such as **bongkrekic acid and atractyloside**. Their effects are similar to those of oligomycin in vitro – a proton gradient builds up and electron transport stops, but, as with oligomycin, respiration can be reactivated by uncouplers.

Regulation of oxidative phosphorylation

Respiratory control and feedback regulation

ADP is the key feedback regulator of oxidative phosphorylation

The oldest and simplest known mechanism of respiratory control is accomplished by the supply of ADP. This is based on the fact that when added to isolated mitochondria, ADP stimulates respiration and ATP synthesis. When ADP is completely converted to ATP, respiration returns to the initial rate. Oxidative phosphorylation is also tightly coupled to other fundamental pathways such as glycolysis, fatty acid oxidation and the tricarboxylic acid cycle (see Chapters 12, 14 and 15) through feedback regulatory mechanisms. The ratios of NADH/NAD and ATP/ADP feedback on key enzymes, such as PFK-1, pyruvate dehydrogenase and isocitrate dehydrogenase. If, for example, oxidative phosphorylation is switched off because of high ATP, both NADH and ATP will negatively feedback on other extramitochondrial pathways, limiting the flux of fuel to mitochondria. Since oxidative phosphorylation depends on the supply of $FADH_2$, NADH, ADP, and Pi as well as the ATP/ADP ratio, magnitude of the membrane potential, uncoupling and hormonal factors, its modes of regulation are clearly complex.

Regulation by covalent modification and allosteric effectors (ATP–ADP)

The main target for regulating oxidative phosphorylation appears to be complex IV. It is phosphorylated in response to hormone action (see Chapter 13) by cyclic-3′5′-adenosine monophosphate (cAMP)-dependent protein kinase (PKA) and dephosphorylated by a Ca^{2+}-stimulated protein phosphatase. Phosphorylation enables allosteric regulation by ATP (ATP/ADP ratio). A high ATP/ADP ratio inhibits and a low ratio stimulates oxidative phosphorylation. It is thought that the complex is normally phosphorylated and inhibited by ATP. With high Ca^{2+} levels, *e.g.* in muscle during exercise (see Chapter 20), the enzyme is dephosphorylated, the inhibition by ATP is abolished, and its activity greatly stimulated, increasing ATP.

Based on the observation that in type 2 diabetes, ATP production is decreased when the β-subunit of ATP synthase is phosphorylated, it has been proposed that this complex is also regulated by phosphorylation/dephosphorylation. Note that the secretion of insulin in β-cells of the pancreas is ATP dependent, because ATP binds to the ATP-sensitive potassium channel (see Insulin Secretion, Chapter 21).

Regulation by thyroid hormones

Thyroid hormones act at two levels in mitochondria. In rats, T_3 stimulates the synthesis of UCP2 and UCP3, which can uncouple the proton gradient, but this has not been documented in humans. Additionally, T_2 binds to complex IV on the matrix side, inducing slip in cytochrome *c* oxidase. The term 'slip' means that complex IV pumps fewer protons per electron transported through the complex, resulting in thermogenesis. The action of T_3 could explain, in part, the long-term and T_2 the short-term thermogenic effects of thyroid hormones (see also Chapter 39).

Mitochondrial permeability transition pore (MPTP)

Located in the inner mitochondrial membrane, the MPTP is a nonselective pore that is a critical factor in cell death. It is normally closed, but will open when cells are reperfused after a period of ischemia (**ischemic reperfusion injury, IRI**; Chapter 37), and small molecules will leave the mitochrondrial matrix. Opening of the MPTP is now considered a key feature of IRI in which cellular damage is much greater than that produced by ischemia alone. Cascades of reactions occuring in response to IRI lead to apoptosis and necrosis, and cell death.

Ischemia, such as that found in heart attacks, is usually caused by a clot that blocks an artery. Clot busters such as streptokinase can be administered to dissolve clots and reperfuse ischemic cells. But, if the ischemic state has been prolonged before administration of a clot buster, death may result from reperfusion injury and opening of the MPTP. This occurs all too often in heart attack patients. Several drugs, such as cyclosporin A, inhibit the MPTP from opening and may protect cells from necrosis or apoptosis after the administration of a clot buster. In fact, a large clinical trial of cyclosporin A is currently being conducted with heart attack patients in conjunction with the administration of a clot buster.

Summary

■ The electron transport system consists of electron carriers that are located in the inner mitochondrial membrane, each of which is isolable as a complex or as a single molecule.
■ Electrons from four major flavoproteins feed electrons to ubiquinone, the first member of the common pathway. Energy derived from the conductance of electrons through the electron transport system is used by three of the complexes to pump protons into the intermembrane space, creating an electrochemical gradient or proton motive force.
■ The proton gradient is used to power ATP synthase for synthesis of ATP by rotary catalysis as well as transport of intermediates across the inner membrane.
■ Numerous toxins can severely impair the electron transport system, the ATP synthase and the translocase that exchanges ATP and ADP across the inner mitochondrial membrane.
■ The rate of ATP production by the electron transport system is regulated by modulation of the proton gradient, by allosteric modification and phosphorylation–dephosphorylation and by thyroid hormones.
■ At least five uncoupling proteins (UCP) with specific tissue distributions are found in the inner mitochondrial membrane, and they all regulate the membrane potential, energy expenditure and thermogenesis.
■ Chronic diseases or conditions such as diabetes, cancer, obesity, and aging all have metabolic links to dysregulation of oxidative phosphorylation through effects on the electron transport system and ATP synthase.
■ The integrity of mitochondria and cells may be disrupted by ischemic reperfusion injury and opening of the mitochondrial permeability transition pore, leading to death and tissue damage.

Further reading

Cao, XH, Zhao, SS, Liu, DY, *et al.* ROS-Ca(2+) is associated with mitochondria permeability transition pore involved in surfactin-induced MCF-7 cells apoptosis. *Chem Biol Interact*. 2011; 190:16–27.

Picard, M, Taivassalo, T, Gouspillou, G, *et al.* Mitochondria: isolation, structure and function. *J Physiol*. 2011; 589:4413–4421.

Pinadda, V, Halestrap, AP. The roles of phosphate and the phosphate carrier in the mitochondrial permeability transition pore. *Mitochondrion*. 2012; 12:120–125.

Ramsden, DB, Ho, PW, Ho, JW, *et al.* Human neuronal uncoupling proteins 4 and 5 (UCP4 and UCP5):

structural properties, regulation, and physiological role in protection against oxidative stress and mitochondrial dysfunction. *Brain Behav*. 2012; 2:468–478.

Ruiz-Meana, M, Fernandez-Sanz, C, Garcia-Dorado, D. The SR-mitochondria interaction: a new player in cardiac pathophysiology. *Cardiovasc Res*. 2010; 88:30–39.

Shanbhag, R, Shi, G, Rujiviphat, J, *et al.* The emerging role of proteolysis in mitochondrial quality control and the etiology of Parkinson's disease. *Parkinsons Dis*. 2012; 2012:382175.

Vonck, J, Schäfer, E. Supramolecular organization of protein complexes in the mitochondrial inner membrane. *Biochim Biophys Acta*. 2009; 1793:117–124.

Wang, WH, Chiang, IT, Ding, K, *et al.* Curcumin-induced apoptosis in human hepatocellular carcinoma j5 cells: critical role of Ca(2+)-dependent pathway. *Evid Based Complement Alternat Med*. 2012; 2012:512907.

Websites

ATP synthase movies

www.cnr.berkeley.edu/~hongwang/Project/ATP_synthase/

www.youtube.com/watch?v=PjdPTY1wHdQ

ATP synthas. http://vcell.ndsu.nodak.edu/animations/atpgradient/index.htm

Virtual Cell Animation Cente. http://vcell.ndsu.nodak.edu/animations/home.htm

Bioenergetic. www.bmb.leeds.ac.uk/illingworth/oxphos/

The Children's Mitochondrial Disease Networ. www.emdn-mitonet.co.uk/

United Mitochondrial disease Foundatio. www.umdf.org/mitodisease/

CHAPTER 10

Digestion and Absorption of Nutrients

The Gastrointestinal Tract

Marek H. Dominiczak, Matthew Priest, Utkarsh V. Kulkarni and John I. Broom

Learning objectives

After reading this chapter you should be able to:

- Describe the main stages of digestion.
- Discuss mechanisms involved in the absorption of nutrients from the digestive tract.
- Discuss the role of digestive enzymes.
- Discuss digestion of the main classes of nutrients: carbohydrates, proteins and fats.
- Identify compounds arising from the digestion of carbohydrates, proteins and fats that become substrates for further metabolism.

Introduction

The gastrointestinal tract and the organs functionally associated with it, are responsible for digestion and absorption of food

All organisms require sources of energy and other materials to enable function and growth. Their survival depends on the ability to extract and assimilate these resources from the ingested food. Also, the intestinal epithelium and the tight junctions between the enterocytes form the most important barrier between the organism and its external environment. The barrier has selective absorption and secretion functions and also may become a scene of immune (or autoimmune) response.

The gastrointestinal (GI) tract and the organs functionally associated with it, principally the liver and pancreas, are responsible for digestion and absorption of food. Digestion is the process by which food is broken down into components simple enough to be absorbed in the intestine. Absorption is the uptake of the products of digestion by intestinal cells (enterocytes) and their delivery to blood or lymph. Digestion and absorption of nutrients are closely linked and are regulated by the nervous system, hormones and paracrine factors. The physical presence of food particles in the GI tract also stimulates these processes.

Importantly, the absorption and secretion of ions such as sodium, chloride potassium, and bicarbonate, and the absorption of water, are also essential functions of the gastrointestinal tract. Therefore many clinical problems associated with digestion and absorption are closely linked to fluid and electrolyte disorders (Chapter 24). It would be a mistake to regard these two fields as separate: a clinician must see them in an integrated way.

Impairment of digestion and absorption results in maldigestion and malabsorption syndromes, respectively. **Maldigestion** denotes impaired breakdown of nutrients to their absorbable products. **Malabsorption** is the defective absorption, uptake and transport of nutrient products that were adequately digested.

The key clinical signs of signs of malabsorption and/or maldigestion are diarrhea, steatorrhea and loss of weight. In children there is failure to thrive. While acute diarrhea carries a risk of fast dehydration and electrolyte depletion,

chronic diarrhea is associated with progressive malnutrition. Worldwide, diarrheal disease is, according to the WHO data (2011), a 5th leading cause of death. Malabsorption and maldigestion may also develop as consequences of surgical intervention such as gastrectomy, small bowel resection or colectomy.

The overall function of the GI tract is to break down food into components that can be absorbed and utilized by the body (Fig. 10.1), and then to excrete the nonabsorbable material. Its different anatomical segments have specific functions relating to digestion and absorption:

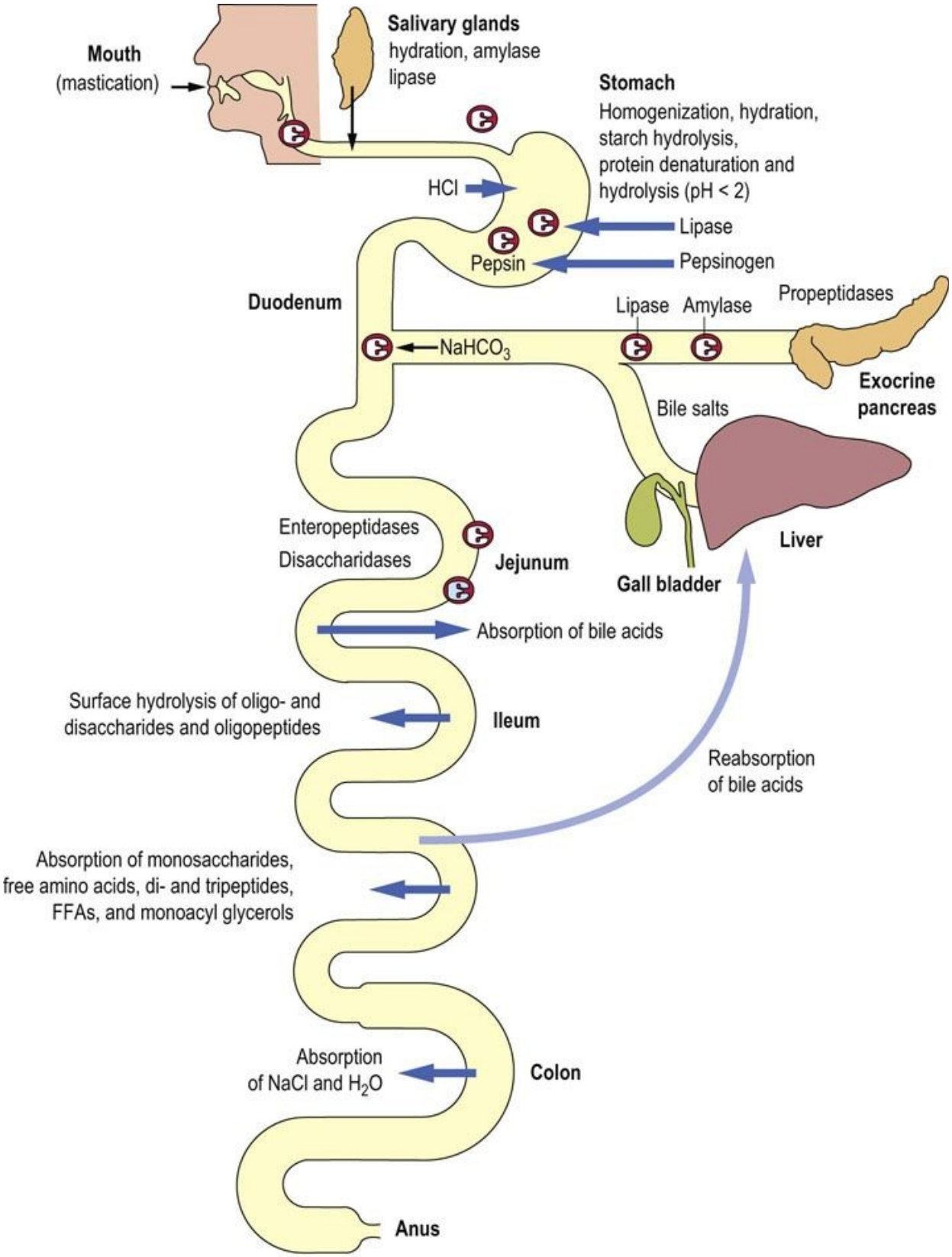


FIG. 10.1 The gastrointestinal (GI) tract.

Digestion and absorption of nutrients require integrated function of several organs. Mixing of food and initiation of digestion take place in the stomach. The absorptive processes start in the jejunum. However, the bulk of nutrients are absorbed in the ileum. The large intestine is involved in the absorption of water and electrolytes, and participates in the recirculation of the bile acids to the liver (Chapter 30). Taking into account all intake and secretions, a large amount of fluid (approximately 10 liters) passes through the GI tract every day. FFA, free fatty acids. Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012, Card 34.

■ The mouth, stomach and duodenum deal with the initial process of mixing ingested food and initiating digestion.
■ In the duodenum, bile and pancreatic secretions enter through the common bile duct.
■ The small intestine is the main digestive area: in the jejunum digestive processes continue and absorption is initiated; it continues in the ileum.
■ The large intestine (cecum, colon and rectum; primarily the colon) is involved in the absorption and secretion of electrolytes and water.

Water and electrolyte handling in the gastrointestinal tract

Handling of electrolytes and water by the GI tract is one of its main functions

Handling of electrolytes and water by the GI tract is one of its main functions, and includes not only their absorption and secretion but also cell volume maintenance, and affects cell proliferation and differentiation, as well as apoptosis and carcinogenesis.

Large volume of fluid is secreted and reabsorbed by the GI tract

In a 24 h period, around 10.0 L of fluid enter and leave the GI tract. One liter of saliva is secreted per day; it contains electrolytes, protein and mucus. Intestinal secretions total about 7.0 liters every day, over and above an average water intake of about 2.0 liters. Most of this fluid is reabsorbed by the small intestine. The colon is particularly active in reabsorption: it absorbs around 90% of the fluid that passes through it, and only about 150–250 mL of water are normally contained in the stool.

Electrolytes are secreted by the salivary glands, stomach and the pancreas

Several secretory processes take place in the GI tract. Salivary glands, stomach and pancreas secrete digestive enzymes in the form of inactive zymogens. There is the hydrogen ion secretion in the stomach. Secretion of the bicarbonate ion takes place throughout the GI tract, with particularly large amounts being secreted in the pancreatic juice. Potassium secretion occurs predominantly in the colon and is regulated by aldosterone.

Impaired intestinal function leads to fluid-electrolyte and acid–base disorders

All this has clinical implications: diseases of the GI tract, and surgical removal of large segments of small and large intestine, carry a risk of major water and

electrolyte disorders. Before treatment for cholera was known, a person with fulminant diarrhea caused by *Vibrio cholerae* infection could die of dehydration within a few hours. Acidosis due to bicarbonate loss can also be a feature of bowel disease (Chapter 25).

Clinical box Causes of fluid and electrolyte loss from the gastrointestinal tract

Prolonged **vomiting** causes loss of water, hydrogen and chloride ions, and a further loss of potassium due to the body's compensatory mechanisms. **Diarrhea** may be caused by increased intestinal secretion due to, for instance, inflammation, or may be caused by malabsorption of nutrients. Severe diarrhea, leading to the loss of alkaline intestinal contents may lead to dehydration and metabolic acidosis. It also results in the loss of sodium, potassium and other minerals (Chapters 11, 23 and 24). Patients with a short gut syndrome resulting from extensive small bowel resection, may develop severe fluid balance problems due to inability to reabsorb water.

Mechanisms of water and electrolyte transport in the intestine

Sodium-potassium ATPase is the driving force for transport processes in the enterocytes

Enterocytes possess an array of transporters and ion channels (Fig. 10.2). The sodium-potassium ATPase, described in more detail in Chapter 24, is located on the basolateral membrane (the 'blood side') and transports the sodium ion outside the cell in exchange for the potassium ($3Na^+$ to every $2K^+$ ions). This creates a sodium concentration gradient, and hyperpolarizes the membrane, increasing the intracellular negative potential and driving the passive transport systems (and thus transcellular ion transport). Moreover, transport of sodium (and chloride) is accompanied by passive transport of water, which is both

paracellular (through the tight junctions) and transcellular (by membrane water transporters, the aquaporins).

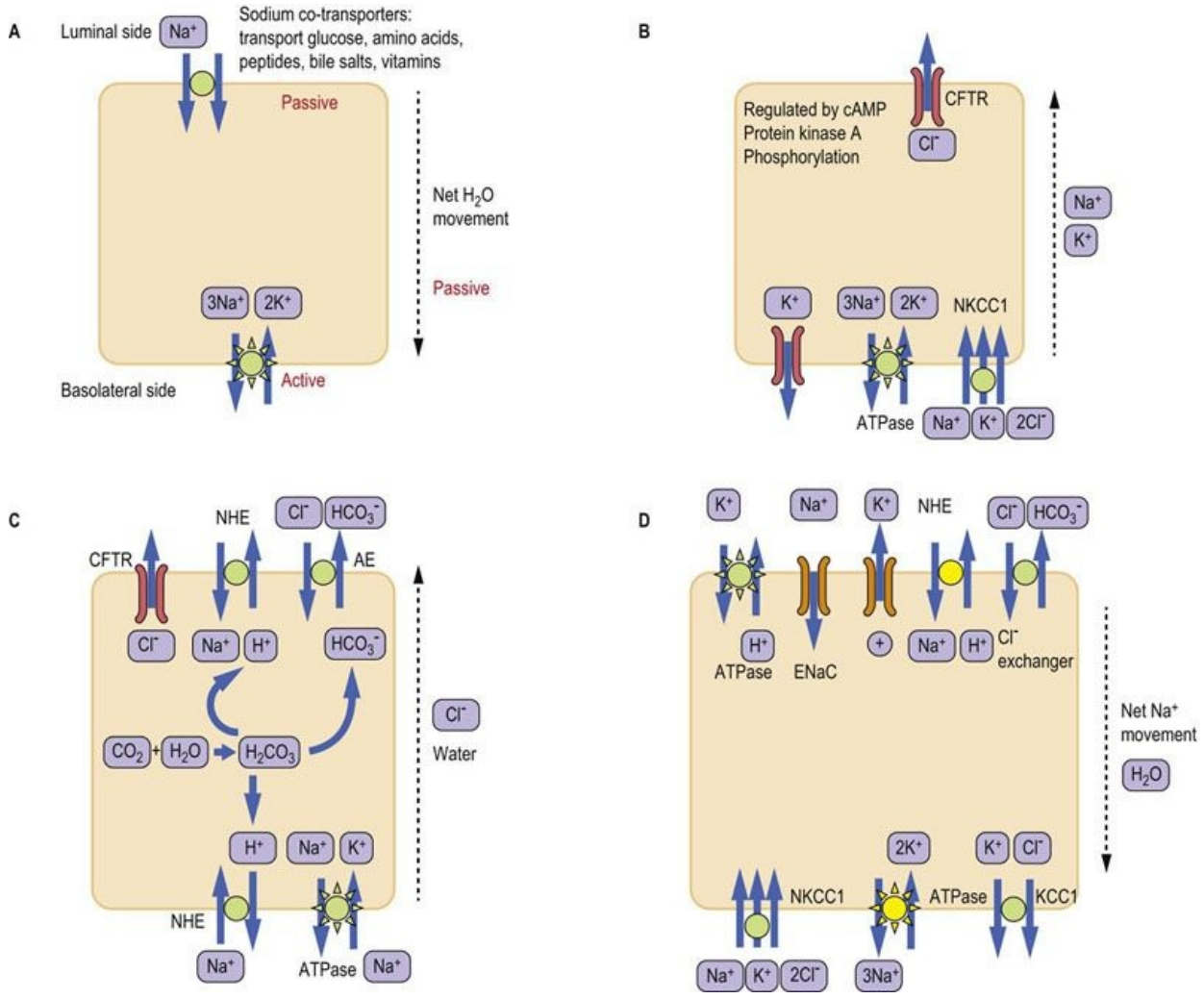


FIG. 10.2 Intestinal electrolyte transport systems.
(A) Sodium cotransporters transport a wide range of substrates, including glucose. Sodium gradient is created by the Na^+/K^+-ATPase located in the basolateral membrane. **(B)** CFTR transporter secretes chloride ion and is regulated by the cAMP-PKA signaling cascade. Sodium and potassium may also be secreted as counterions. Note the potassium 'leak' channel in the basolateral membrane. The NKCC1 transporter supplies chloride to the cell. **(C)** The electroneutral sodium absorption and the secretion of bicarbonate. **(D)** Electrogenic sodium absorption and potassium secretion in the distal colon. Transporters marked **yellow** are regulated by **aldosterone** in the distal colon. See text for details. CFTR, cystic fibrosis transmembrane conductance regulator; NHE, sodium/hydrogen exchanger; ENaC, epithelial sodium channel; AE, anion exchanger (chloride/bicarbonate exchanger); NKCC1, Na^+ K^+ Cl^- cotransporter; KCC1, K^+ Cl^- cotransporter.

Sodium cotransporters are a common mode of intestinal transport

Sodium cotransporters transport the sodium ion together with another molecule (Fig. 10.2A). For instance, glucose is absorbed together with sodium by the sodium-glucose cotransporter present in the luminal membrane, known as **SGLT-1** (Sodium/Glucose Linked Transport-1). Glucose is subsequently extruded into plasma at the basolateral membrane by the **GLUT2** transporter (Fig. 10.7 below). The discovery of the link between intestinal transport of sodium and glucose had enormous clinical consequences. It was during a cholera epidemic in Manila, in the late 1960s, that researchers observed that patients who had been dehydrated because of diarrhea did not absorb oral sodium chloride well during attempts at oral rehydration. However, they started to do so when glucose was also provided. This observation led to the formulation of the **WHO oral rehydration solution**, which subsequently saved the lives of millions of children affected by severe diarrhea worldwide.

Other modes of sodium transport are the electroneutral (Fig. 10.2B) and electrogenic transport

The electroneutral sodium transport is through sodium/hydrogen exchanger (**NHE**), usually combined with chloride transport via the chloride/bicarbonate exchanger known as AE exchanger (Fig. 10.2C). The exchangers are present on both luminal and basolateral membranes. This type of transport is responsible for most of the sodium chloride reabsorption in the colon. In the distal colon the NHE exchanger is upregulated by glucocorticoids.

The electrogenic absorption of sodium occurs through the epithelial sodium channels (**ENaCs**, also known as amiloride-sensitive sodium channels), which are present on the luminal side of the epithelium (Fig. 10.2D). ENaCs are regulated by steroids (aldosterone) and are important particularly in the distal colon. Absorption of Na^+ is accompanied by Cl^- following through a chloride channel (which could be the CFTR – see below). Aldosterone also upregulates the Na^+/K^+-ATPase.

Chloride transport: the cystic fibrosis transmembrane conductance regulator (CFTR)

Luminal secretion of chloride occurs via the cystic fibrosis transmembrane conductance regulator (CFTR; Fig. 10.2B). The CFTR is a single-polypeptide membrane ion channel. It is also present in the epithelia of the lung and sweat glands. Its function is controlled by the G-protein–cAMP-protein kinase A (PKA) signaling cascade (Chapter 40). Because the CFTR is activated by cAMP, prostaglandin E_2 (PGE_2), serotonin, as well as the cholera toxin and the *E. coli* heat-stable enterotoxin, all activate chloride secretion. On the other hand, the loss-of-function mutations of CFTR are the cause of **cystic fibrosis**, where the chloride transport is impaired or inhibited. CFTR also has regulatory function: its phosphorylation inhibits the NHE exchanger, thus decreasing Na^+ absorption. Interestingly CFTR is also able to transport chloride in the opposite direction, aiding chloride reabsorption (above).

The basolateral $C1^-$ uptake occurs through the Na^+ K^+ $C1^-$ cotransporter (known as NKCC1) and through chloride/bicarbonate exchangers.

Clinical box Cystic fibrosis

Cystic fibrosis, a monogenic autosomal recessive disorder, involves inhibition of chloride transport due to the absence of the CFTR. Different mutations of the *CFTR* gene lead to either complete absence of the transporter or impair its functionality.

The prevalence of CF is 1 : 3000 live births in the USA and northern Europe. In the USA, cystic fibrosis is the No. 1 cause of **malabsorption**. It manifests itself predominantly in childhood. The main problems are usually respiratory. Chloride secretion is decreased and the Na^+ reabsorption is accelerated. This results in decreased hydration of epithelial secretions. In the respiratory tract there is **decreased hydration of the airway mucus** and thus failure of its clearance, with ensuing bacterial infections. Gastrointestinal problems include **meconium ileus** and **intestinal obstruction**. The absence of the CFTR also affects functioning of the Cl^-/HCO_3^- exchanger (and thus the passive secretion of Na^+) – this impairs pancreatic enzyme secretion. Thickened biliary secretions may be a cause of focal **biliary cirrhosis** and **chronic cholelithiasis**. There also is impairment of mucus secretion in the colonic crypts, with enhanced Na^+ reabsorption through Na^+ channels and Na^+/H^+ transporters.

Potassium absorption and potassium secretion in the colon is aided by a several potassium channels

Potassium absorption is mediated by **H^+/K^+ ATPases** (belonging to the family of P-type ATPases) in the luminal membrane. On the other hand, the basolateral potassium transport is by potassium channels and the K^+ and Cl^- cotransporter (**KCC1**). Both basolateral and luminal K^+ channels are necessary to hyperpolarize the membrane to establish a driving force for the ENaC transporter.

K^+ secretion through luminal K^+ channels parallels the Cl^- secretion through CFTR, and is similarly stimulated by cAMP, cGMP and protein kinase C (PKC). The expression of luminal K^+ channels is also stimulated by aldosterone and glucocorticoids.

Reabsorption of short-chain fatty acids occurs together with bicarbonate secretion

The colon reabsorbs short-chain fatty acids (SCFAs) derived from bacterial fermentation of fiber, and this is combined with secretion of bicarbonate. Bicarbonate is secreted using luminal anion exchangers SCFA/HCO_3^- or $C1^-/HCO_3^-$.

Aquaporins control colonic water reabsorption

Water reabsorption in the colon is mediated by ion channels known as aquaporins (AQP). AQP1,3 and 4 are located on basolateral and AQP8 on luminal membranes. The intestinal electrolyte transport systems are summarized in Figure 10.2.

The pH of intestinal secretions vary

Hydrogen ion concentration varies widely in different parts of the GI tract. This is essential for the digestive process and for protection of the underlying tissues in the stomach and intestines. Saliva secreted into the mouth is alkaline due to its bicarbonate content. On the other hand, the contents of the stomach are strongly acidic, but the mucus protecting its walls is alkaline. Thus while the parietal cells of the stomach secrete large amounts of hydrogen ion (principally through the action of the luminal H^+/K^+-ATPase), the gastric surface cells secrete mucus containing bicarbonate ion; they employ the Cl^-/HCO_3^- exchanger. On entry to

the duodenum the acidic content of the stomach is neutralized by strongly alkaline (due to the bicarbonate content) pancreatic secretions.

Digestion

There is a significant mechanical component to digestion. Chewing breaks down food to enable it to be swallowed, whilst the addition of saliva in the mouth begins the digestive process and acts as lubrication to facilitate swallowing. The food is then moved into the esophagus by a process driven by the esophageal reflex. As it transfers into the stomach, it is broken down into smaller particles. The presence of the digest triggers peristalsis, which further helps mixing and stimulates digestive secretions. Major stimuli to peristalsis are mediated through the parasympathetic nervous system. Absorption of nutrients depends on the rate of transit: thus, increased motility may lead to malabsorption.

The stomach and intestines are lined by epithelium, which has an invaginated surface that greatly increases its area. The small intestine is lined by enterocytes arranged in intestinal villi, and in addition each cell contains microvilli. The total absorptive surface area of the intestine is approximately 250 m^2, roughly the area of a junior basketball court.

Advanced concept box Digestive function of the stomach

There are different cell types in the mucosal wall of the stomach, performing different digestive functions. Cells called ‘chief cells’ secrete **pepsinogen**, which is a precursor of **pepsin**. Pepsinogen is activated to pepsin in the acidic environment of the stomach lumen. Parietal cells generate **hydrogen ions** through the action of carbonic anhydrase and then pump them into the lumen by an ATP-dependent proton pump on the luminal membrane. The H^+ secretion is dependent on the parallel export of K^+ through luminal K^+ channel.

Parietal cell activity is stimulated by the action of **histamine** acting on H_2 receptors. (Chapter 41), produced by histamine-secreting cells. The hormone **gastrin** is secreted by G-cells, and is triggered by food entering the stomach. Stomach cells also secrete the **intrinsic factor (IF)**, which facilitates absorption of vitamin B_{12} in the intestine (Chapter 11). Last, but not least, epithelial cells secrete alkaline mucus, which protects the stomach lining from the

effects of the strong acid.

Damage to the lining of the stomach or duodenum leads to **ulceration**. Treatment of acid-related symptoms such as **dyspepsia** or **gastroesophageal reflux**, can be achieved with **antacids** which simply neutralize the pH, **H_2 antagonists**, *e.g.* cimetidine or ranitidine, which prevent histamine release, or **proton pump inhibitors**, *e.g.* omeprazole, which block H^+ secretion by the parietal cells.

Digestion is a sequential, ordered series of processes

The carbohydrates, proteins and fats are split to absorbable products. Some ingested material, such as complex carbohydrates of plant origin, is indigestible and constitutes fiber.

The process of digestion is characterized by a number of stages that occur in a sequence, allowing the interaction of fluid, pH, emulsifying agents and enzymes. This, in turn, requires concerted secretory action from the salivary glands, liver and gall bladder, the pancreas, and the intestinal mucosa. The processes involved are outlined in Figure 10.1 and can be summarized as follows:

- Lubrication and homogenization of food with fluids secreted by glands of the intestinal tract, starting in the mouth.
- Secretion of enzymes that break down macromolecules to a mixture of oligomers, dimers and monomers.
- Secretion of electrolytes, hydrogen ions and bicarbonate within different parts of the GI tract to optimize the conditions for enzymic hydrolysis.
- Secretion of bile acids to emulsify dietary lipid, facilitating enzymic hydrolysis and absorption.
- Further hydrolysis of oligomers and dimers by membrane-bound enzymes (jejunum).
- Specific transport of digested material into enterocytes and thence to blood or lymph.
- Recycling of bile acids. Absorption of the SCFAs produced by colonic bacterial fermentation.
- Reabsorption of water and electrolytes.

Spare capacity in the intestinal tract

There is considerable functional reserve in all aspects of digestion and absorption

Because if this, minor functional loss may go unnoticed, allowing pathology to progress for some time before being diagnosed. A considerable impairment of structure/function needs to be present before signs and symptoms of GI maldigestion or malabsorption occur. Each of the organs involved in digestion and absorption has the capacity to increase its activity several fold in response to specific stimulation; this adds to the reserve capacity. For example, pancreatic disease manifests itself only after 90% of the pancreatic function is destroyed.

Note also that digestion of a particular nutrient takes place at several points in the GI tract. Lipids, carbohydrates and proteins can be digested at multiple points along the GI tract. Therefore disruption of digestive mechanisms at a single point is unlikely to cause a complete inability to digest a nutrient group (Fig. 10.3).

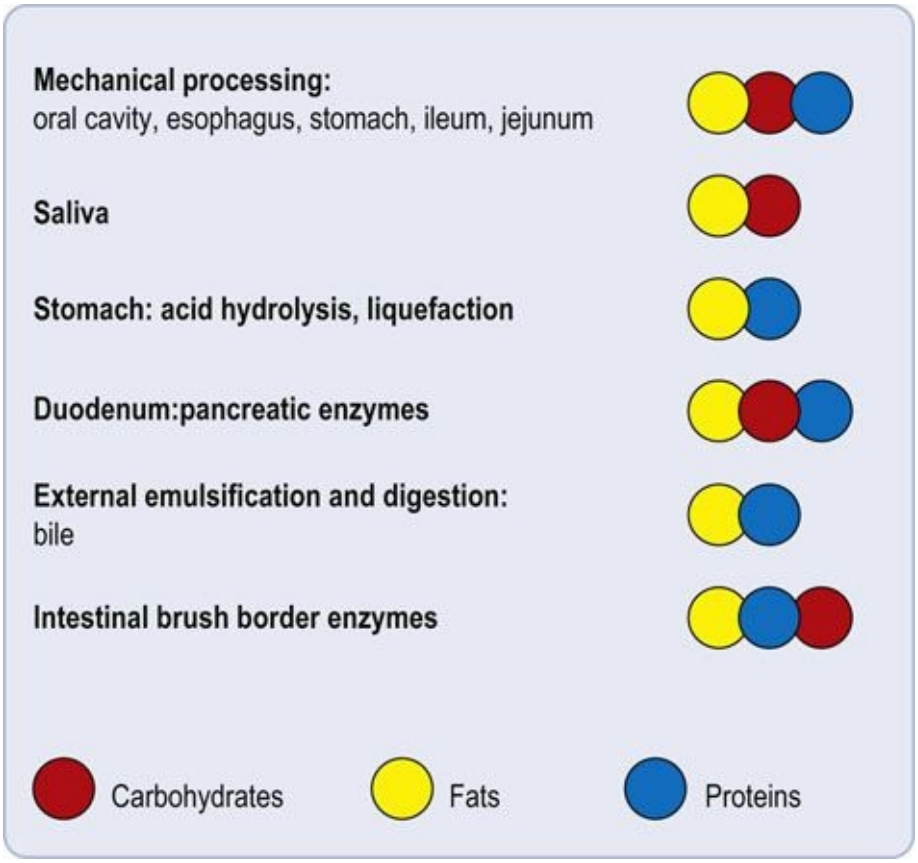


FIG. 10.3 Digestion as a multiorgan process.
Each of the main group of nutrients (carbohydrates, proteins and fats) undergoes digestion at multiple points.

The GI tract can also accommodate loss of function of one particular constituent organ. For example, if the stomach is surgically removed, the pancreas and small intestine can compensate for the loss of gastric digestion. On the other hand, in pancreatic disease, lingual lipases can accommodate, some loss of pancreatic lipase.

Digestive enzymes and zymogens

Most digestive enzymes are secreted as inactive precursors

With the exception of salivary amylase and lingual (associated with the tongue; hence oral) lipases, digestive enzymes are secreted into the gut lumen as inactive precursors termed zymogens (Chapter 6). Secretion of enzymes is similar in the salivary glands, gastric mucosa and pancreas. These organs contain specialized cells for the synthesis, packaging and transport of zymogen granules to the cell

surface and thence to the intestinal lumen. These secretions are termed exocrine, *i.e.* 'secreting to the outside', as opposed to the endocrine secretion of hormones.

Enzymes involved in digestion of protein (proteases) and fat (lipase: phospholipase A_2) are synthesized as inactive zymogens and are only activated on their release to the gut lumen. In general, these enzymes, once in their active form, can activate their own precursors. Activation of the precursors can also occur by change in pH (e.g. pepsinogen in the stomach is converted at pH below 4.0 into pepsin) or by the action of specific enteropeptidases bound to the mucosal membrane of the duodenum (Fig. 10.1).

All digestive enzymes are hydrolases

All digestive enzymes hydrolyze their substrates. The products of such hydrolytic procedures are oligomers, dimers and monomers of the parent macromolecule. Thus, carbohydrates are hydrolyzed into a mixture of dissacharides and monosaccharides. Proteins are broken down to a mixture of di- and tripeptides and amino acids. Lipids are broken down to a mixture of fatty acids (FAs), glycerol and mono-and diacylglycerols (Fig. 10.4).

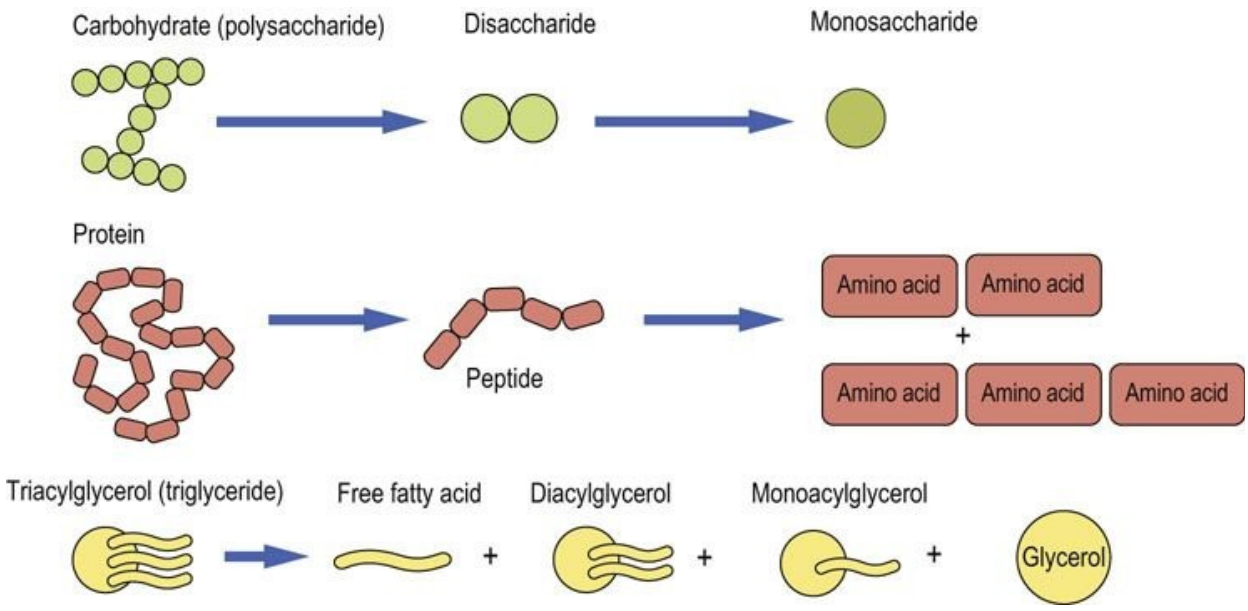


FIG. 10.4 Digestion of dietary polymers.
Polysaccharides are digested to yield di-and monosaccharides, and proteins to yield the component amino acids. Fat (triacylglycerols) digestion is a stepwise removal of fatty acid molecules, yielding di-and monoacylglycerols.

Digestion and absorption of carbohydrates

Dietary carbohydrates enter the GI tract as mono-, di-and polysaccharides

Dietary carbohydrates consist of mainly plant and animal starches (polysaccharides), the disaccharides sucrose and lactose, and the monosaccharides (Fig. 10.5). Monosaccharides include glucose, fructose and galactose, which are either present in the diet or are generated by digestion of di- and polysaccharides. Lactose, for instance, is a disaccharide derived from dairy products, and is hydrolyzed to the monosaccharides glucose and galactose by lactase and β-galactosidase. These sugar monomers can then be absorbed from the GI tract.

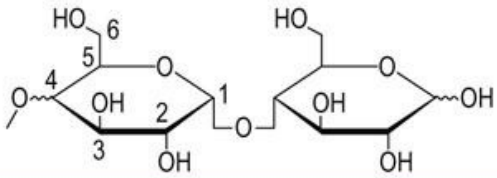
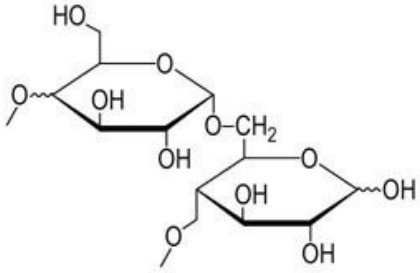
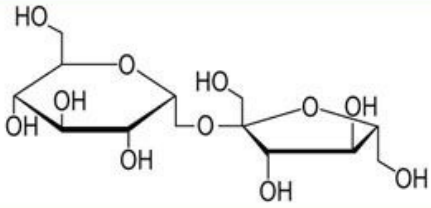
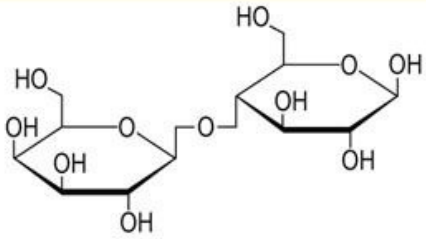
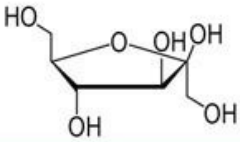
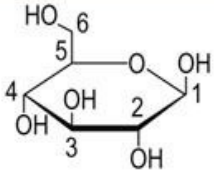
Carbohydrate	Food source	Structure
Starch (amylose) [plant]	Potatoes, rice, bread, onions	
Amylopectin (glycogen) [plant, animal]	Potatoes, rice, bread, muscle, liver	
Sucrose	Desserts, sweets, 'sugar'	
Lactose	Milk	
Fructose	Fruits, honey	
Glucose	Fruits, honey	

FIG. 10.5 Structure of the key dietary carbohydrates.
Starch and amylopectin are polysaccharides, and only two component sugar molecules are shown for each to illustrate the intermolecular linkages. Sucrose and lactose are the most common disaccharides, and fructose and glucose the most common monosaccharides. Refer to the glucose molecule for the standard numbering of carbon atoms (see also Chapter 3).

Advanced concept box Role of amylase, α-glucosidases and isomaltase in polysaccharide digestion

During eating, homogenization of food occurs by chewing. It is aided by contractions of the stomach wall muscles and gastric folds. One consequence of this is that dietary polysaccharides become hydrated. This is necessary for the action of **amylase**, which is specific for internal α-1,4-glycosidic linkages and not the α-1,6 linkages. Amylase also does not act on α-1,4 linkages of glycosyl residues serving as branching units. Thus, the cleaved units formed by its action are the disaccharide maltose, the trisaccharide maltotriose, and an oligosaccharide with one or more α-1,6 branches and containing on average eight glycosyl units, termed the 'α-limit dextrin'. These compounds are further cleaved to glucose by **oligosaccharidase** and **α-glucosidase**, the latter removing single glucose residues from α-1,4-linked oligosaccharides (including maltose). A **sucrase–isomaltase** complex is secreted as a single polypeptide precursor molecule and is activated into two separate enzymes, one of which (isomaltase) is responsible for the hydrolytic cleavage of α-1,6-glycosidic bonds. Thus the final product of digestion of starches is glucose. The amylase occurs free in the lumen, whereas α-glucosidases and **isomaltase** are attached to the membrane of the enterocyte.

Disaccharides, and polysaccharides such as starch and glycogen, require hydrolytic cleavage into monosaccharides before absorption

Disaccharides are broken down by membrane-bound disaccharidases present on the intestinal mucosal surface. Starch and glycogen require additional hydrolytic capacity of the enzyme amylase found in the secretions of the salivary glands and pancreas (Fig. 10.6).

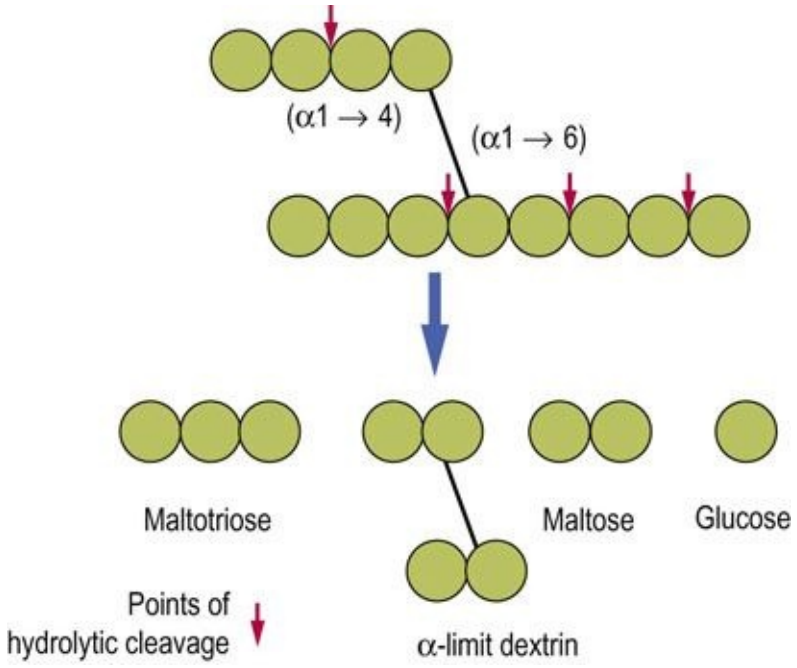


FIG. 10.6 Hydrolytic cleavage of polysaccharides.
Enzymatic hydrolysis is the mechanism of digestion of polysaccharides and disaccharides. The arrows illustrate points of cleavage and the type of hydrolyzed bond. Note that α-limit dextrin still contains both α-1,4 and α-1,6 bonds.

Starch is a plant polysaccharide and glycogen is its animal equivalent. Both contain a mixture of linear chains of glucose molecules linked by α-1,4-glycosidic bonds (amylose) and by branched glucose chains with α-1,6 linkages (amylopectin). Glycogen has a far more branched structure than starch. Digestion of these polysaccharides is promoted by the endosaccharidases, and amylase.

The products of hydrolysis of starch are the disaccharide maltose, the trisaccharide maltotriose and a branched unit, termed the α-limit dextrin. These products are further hydrolyzed by enzymes bound to the enterocytes, finally yielding the monosaccharide glucose (Fig. 10.7A).

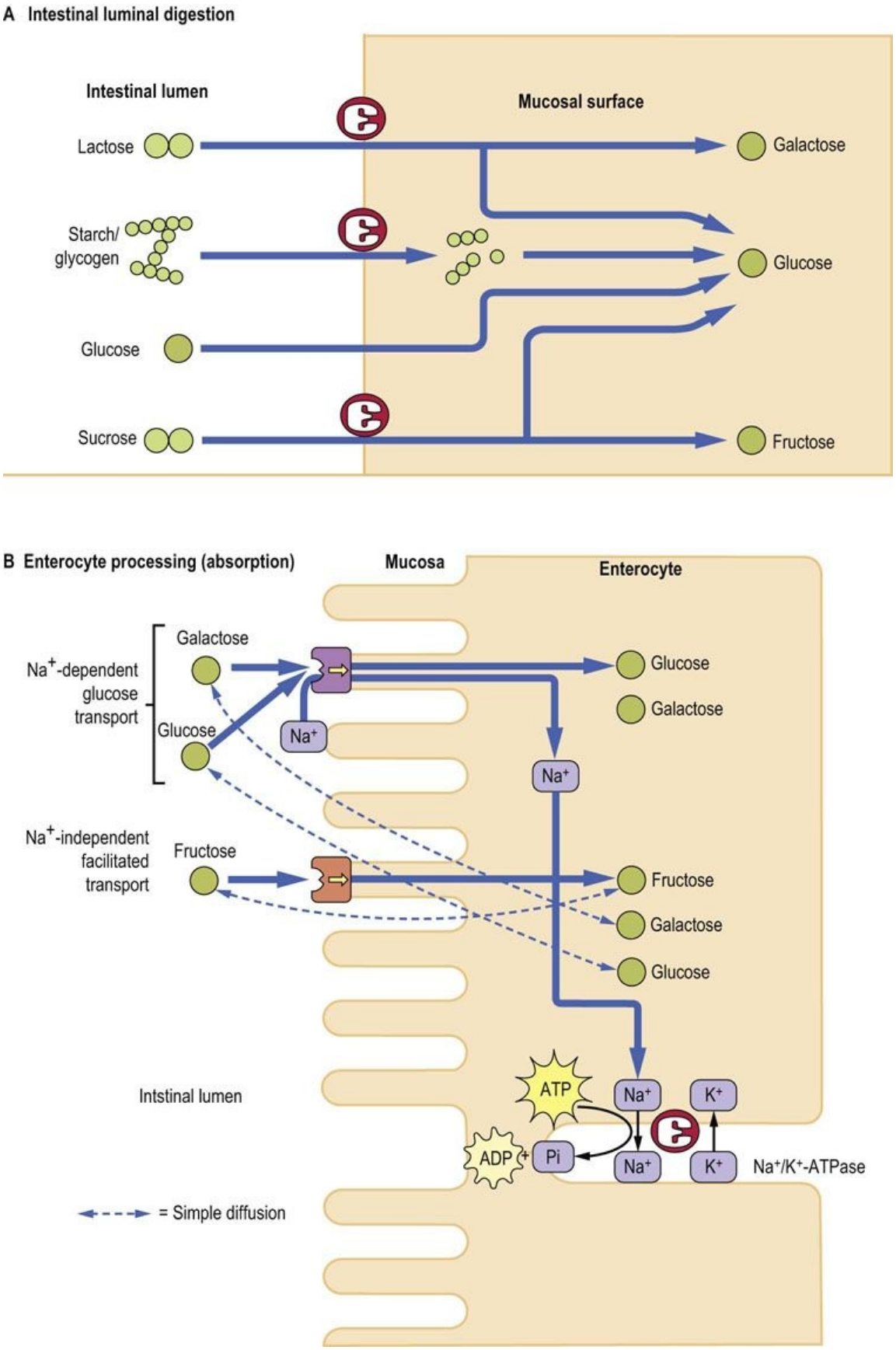

A Intestinal luminal digestion
Intestinal lumen
Mucosal surface
Lactose
Galactose
Starch/
glycogen
Glucose
Glucose
Sucrose
Fructose
B Enterocyte processing (absorption)
Mucosa
Enterocyte
Na+-dependent
glucose
transport
Galactose
Glucose
Na+
Glucose
Galactose
Na+
Na+-independent
facilitated
transport
Fructose
Fructose
Galactose
Glucose
Intstinal lumen
ATP
Na+
K+
ADP
Pi
Na+
K+
Na+/K+-ATPase
= Simple diffusion

FIG. 10.7 Digestion and absorption of dietary carbohydrates.
(A) Monosaccharides are released as a result of hydrolysis of different polysaccharides. Note that preliminary digestion occurs in the gut lumen, and the final stage takes place on the mucosal surface. **(B)** Links between absorption of monosaccharides and sodium, and their relationship to the activity of Na^+/K^+-ATPase. ADP, adenosine diphosphate; ATP, adenosine triphosphate; Pi, inorganic phosphate. Compare with Figure 8.5.

Dietary disaccharides such as lactose, sucrose and trehalose (a disaccharide made up of two glucose molecules joined by an α-1,1 linkage) are hydrolyzed to their constituent monosaccharides by specific disaccharidases attached to the small intestinal brush border membrane. The catalytic domains of these enzymes project into the gut lumen to react with their specific substrates, whilst their noncatalytic, structural domain(s) are attached to the enterocyte membrane.

Disaccharidases are inducible, with the exception of lactase

The greater the amount of a disaccharide (e.g. sucrose) present in the diet or produced by digestion, the greater is the amount of the relevant specific disaccharidase (e.g. sucrase) produced by the enterocytes. The rate-limiting step in the absorption of dietary disaccharides is thus the transport of the resultant monosaccharides. Lactose, on the other hand, is a noninducible brush border disaccharidase, and therefore the rate-limiting factor in lactose absorption is its hydrolysis.

Clinical box Types of diarrhea

Diarrhea can be caused by the nonabsorbable solutes present in the gut (osmotic diarrhea), by the failure to either digest or absorb nutrients, and also by the secretory agonists (secretory diarrhea).

Osmotic diarrhea may be caused by malabsorption, digestive enzyme deficiencies or short bowel, and inflammatory disease.

Secretory diarrhea may be caused by infections, malabsorption of bile salts, malabsorption of fat or by endocrine causes such as carcinoid syndrome or Zollinger–Ellison syndrome. Main causes of inflammatory diarrhea are Crohn's disease, ulcerative colitis, and irritable bowel syndrome.

Absorption can be impaired, and secretion increased, in conditions that lead to the inflammation of the bowel

(**inflammatory diarrhea**). Characteristically, the secretory diarrhea but not osmotic one, persists on fasting.

Active and passive transport systems transfer monosaccharides across the brush border membrane

The process of digestion results in a large increase in the number of osmotically active monosaccharide particles within the gut lumen. This leads to water being drawn into the lumen from the GI tract mucosa and vascular compartment. Increased brush border hydrolysis increases the osmotic load, while increased monosaccharide transport across the brush border enterocyte decreases it. For most oligo-and disaccharidases, the transport of the resulting monomers is rate limiting. As monomeric sugar concentrations (and osmolality) increase in the gut lumen, there is a compensatory decrease in the activity of brush border disaccharidases. This controls the osmotic load and prevents fluid shifts.

Glucose, fructose and galactose are the primary monosaccharides resulting from the digestion of dietary carbohydrates. Absorption of these sugars and other minor monosaccharides occurs by means of specific carrier-mediated mechanisms (Fig. 10.7B), all of which demonstrate substrate specificity and stereospecificity, show saturation kinetics and can be specifically inhibited. In addition, all monosaccharides can cross the brush border membrane by a simple diffusion, although this is extremely slow.

Clinical box A boy with abdominal discomfort, bloating and diarrhea: lactose intolerance

A 15-year-old African-American boy came across to the UK on an exchange visit for 2 months. After 2 weeks in the UK, he complained of abdominal discomfort, a feeling of being bloated, increased passage of urine and, more recently, the development of diarrhea. His only change in diet noted at the time was the introduction of milk. He had developed a considerable liking for milk and was consuming 1–2 large cartons per day. A lactose tolerance test was performed, whereby the young man was given

50 g lactose in an aqueous vehicle to drink. Plasma glucose levels did not rise by more than 1 mmol/L (18 mg/dL) over the next 2 hours, with sampling at 30-minute intervals. A diagnosis of lactose intolerance was made.

Comment.

Lactose intolerance results from acquired lactase deficiency. Lactase activity decreases with increasing age in children, but the extent of the decline in activity is genetically determined and demonstrates ethnic variation. Lactase deficiency in the adult black population varies from 45% to 95%. If symptoms of malabsorption occur after the introduction of milk to adult diets, the diagnosis of acquired lactase deficiency should be considered. The diagnosis is made by challenging the small bowel with lactose and monitoring the rise in plasma glucose. An increase of more than 1.7 mmol/L (30 mg/dL) is considered normal. A rise of less than 1.1 mmol/L (20 mg/dL) is diagnostic of lactase deficiency. A rise of 1.1–1.7 mmol/L (20–30 mg/dL) is inconclusive.

There are at least two carrier-mediated transport mechanisms for monosaccharides

At the brush border membrane both glucose and galactose are transported by the sodium-dependent glucose transporter mentioned above. This membrane-linked protein binds with glucose (or galactose) and Na^+ at separate sites, and transports both into the enterocyte cytosol. Na^+ is transported down its concentration gradient (the concentration within the gut lumen exceeding that inside cells), and carries glucose along *against* the glucose concentration gradient. This transport is linked to Na^+/K^+-ATPase). The transport of glucose or galactose is thus an indirect active process.

Fructose is transported across the brush border membrane by a sodium-independent facilitated diffusion involving the membrane-associated glucose transporter GLUT-5 present on the brush border side of the enterocyte, and GLUT-2, which transfers monosaccharides out of the enterocyte into the circulation (Table 8.2).

An incomplete digestion of carbohydrates (the components of fiber) leads to their conversion to short-chain fatty acids (acetate, propionate, butyrate) by the colonic bacteria.

Clinical box A young man with weight loss, diarrhea, abdominal bloating and anemia: celiac disease

A 22-year-old man presented with a history of weight loss, diarrhea, abdominal bloating and anemia. He described his stools as pale and bulky. Laboratory features included hemoglobin of 90 g/L (9 g/dL) (reference range 130–180 g/L; 13–18 g/dL). Biopsy of his small bowel demonstrated flattening of the mucosal surface, villous atrophy and disappearance of microvilli. A diagnosis of gluten-induced enteropathy (celiac disease) was made. All wheat products were removed from the patient's diet and the symptoms resolved.

Comment.

Celiac disease is an autoimmune condition precipitated by sensitivity to gluten resulting in inflammation of the small bowel mucosa. Gluten is a storage protein of wheat, barley and rye. It is actually a mixture of proteins, which includes the gliadins (the alcohol-soluble fraction of gluten) and glutelins. The gliadins pass through the intestinal barrier during, for instance, infections, and trigger the immune response. The inflammatory reaction ensues. The result is villous athrophy and hyperplasia of the crypts. Since the absorptive surface is markedly reduced, the resulting malabsorption can be severe.

Circulating antibodies to wheat gluten and its fractions are frequently present in cases of celiac disease. The diagnosis involves duodenal biopsy and testing the response to a gluten-free diet. The autoantibodies are tested for are **antigliadin antibodies** and/or tissue **transglutaminase antibodies** (transglutaminase is an enzyme that deamidates gliadin in the intestinal wall). Celiac disease is underdiagnosed, especially in patients with unexplained anemia.

For hematology reference values, refer to Table 5.2.

Advanced concept box Short-chain fatty acids are produced in the large bowel from undigested carbohydrates

Decreased absorption of dietary starch leads to the bacterial production of short-chain fatty acids (SHFAs) by the colonic bacteria.

SHFAs can be produced from fermentable oligosaccharides, disaccharides, monosaccharides and polyols (FODMAPs). Animal studies showed the presence of the short chain fatty acids receptor 2 (FFA2), a G-protein-coupled receptor present on intestinal endocrine cells., Binding of SCFA releases serotonin, leading to the increase in intestinal motility.

Digestion and absorption of lipids

Approximately 90% of fat in the diet is **triacylglycerol (TAG; also termed triglyceride)**. The remainder consists of cholesterol, cholesteryl esters, phospholipids and nonesterified fatty acids (NEFAs).

Fats need to be emulsified before digestion

The hydrophobic nature of fats prevents the access of water-soluble digestive enzymes. Furthermore, fat globules present only a limited surface area for enzyme action. These issues are overcome by the emulsification process. The change in the physical nature of lipids begins in the stomach: the core body temperature helps to liquefy dietary lipids, and the peristaltic movements of the stomach facilitate formation of a lipid emulsion. The emulsification process is also aided by the acid-stable salivary and gastric lipases. The initial rate of hydrolysis is slow, due to the separate aqueous and lipid phases and limited lipid–water interface. Once hydrolysis begins, however, the water-immiscible TAGs are degraded to fatty acids, which act as surfactants. They confer a hydrophilic surface to lipid droplets and break them down into smaller particles, thus increasing the lipid–water interface and facilitating hydrolysis. The lipid phase disperses throughout the aqueous phase as an emulsion. Dietary phospholipids, fatty acids and monoacyl glycerols also act as surfactants.

Bile salts and pancreatic enzymes act on the lipid emulsion in the duodenum

The lipid emulsion passes from the stomach into the duodenum where the further digestion occurs, driven by enzymes secreted by the pancreas. Solubilization is aided by the release of bile salts from the gall bladder, stimulated by the hormone cholecystokinin.

The major enzyme secreted by the pancreas is pancreatic lipase. Lipase, however, remains inactive in the presence of bile salts normally secreted into the small intestine. This inhibition is overcome by the concomitant secretion of co-lipase by the pancreas. Co-lipase binds to both the water–lipid interface and to pancreatic lipase, simultaneously anchoring and activating the enzyme. As shown in Figure 10.8, only a small proportion of dietary TAGs become completely hydrolyzed to glycerol and fatty acids. The ‘second’ and ‘third’ fatty

acids in TAGs are hydrolyzed with increasing difficulty: the pancreatic lipase produces mainly 2-monoacyl glycerols (2-MAGs), which are absorbed into enterocytes.

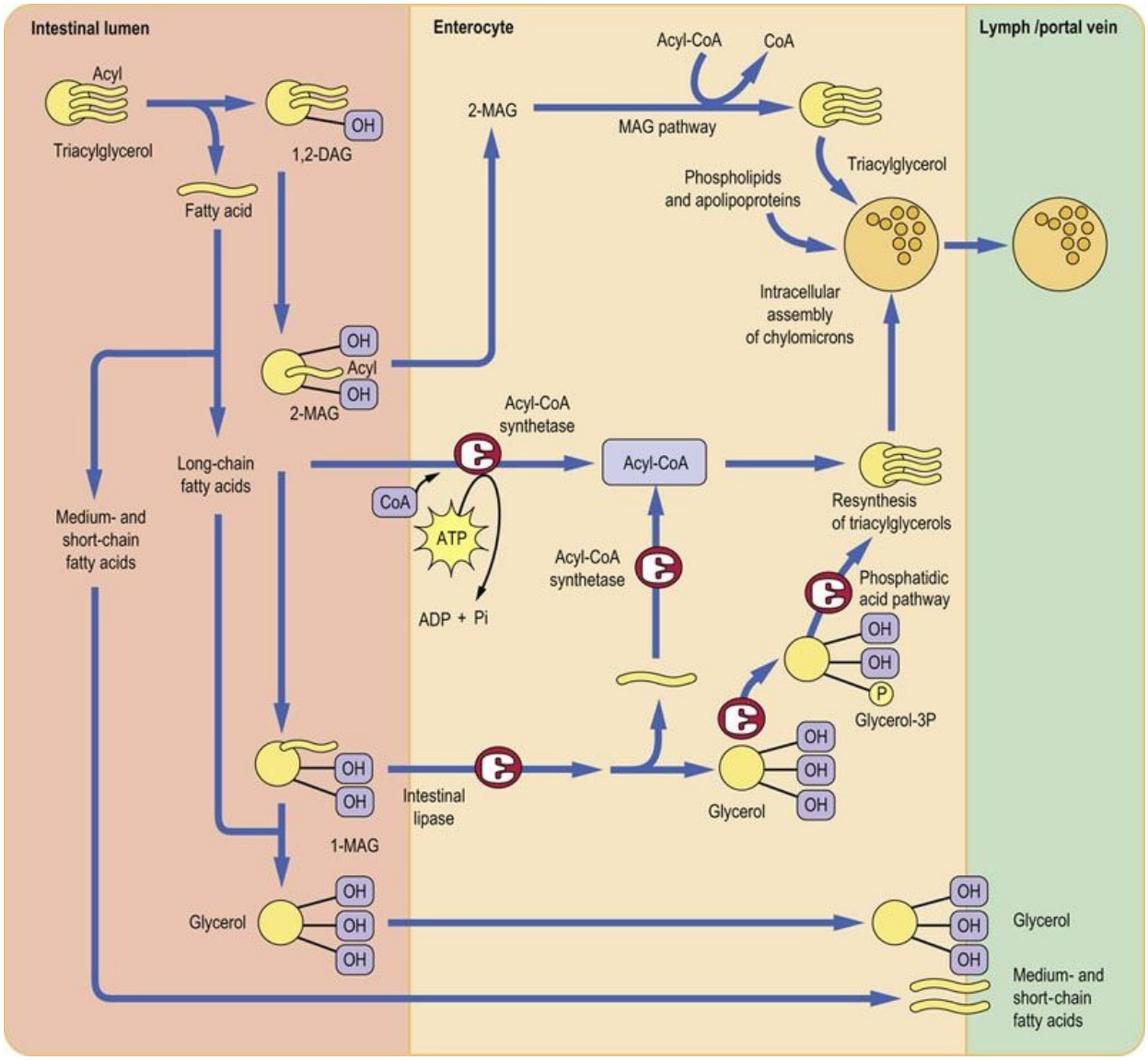


FIG. 10.8 Digestion and absorption of dietary lipids.
Dietary triglycerides undergo variable degrees of hydrolysis in the intestinal lumen. Subsequently, medium-and short-chain fatty acids are absorbed into the portal blood. However, long-chain fatty acids are resynthesized into triacylglycerols and then are incorporated into chylomicrons. The fatty acids are activated by acetyl-CoA before the synthesis of acylglycerols can take place. Note that enterocytes do not possess glycerol kinase: formation of glycerol phosphate requires the presence of glucose. TAG, triacylglycerol; DAG, diacylglycerol; MAG, monoacylglycerol; CoA, coenzyme A.

Bile salts are essential for solubilizing lipids during the digestive process

Bile acids (which are bile salts at the alkaline pH of the intestine) act as detergents and reversibly form lipid aggregates (micelles). Micelles are considerably smaller than lipid emulsion droplets. Their size depends on the bile acid concentration and the ratio of bile acids to lipids.

Thus, the lipid digest changes from fat emulsion droplets into micellar structures. The micelles transport the lipids to the brush border of the enterocyte.

The absorption of lipids into the epithelial cells lining the small intestine occurs by diffusion through the plasma membrane. Almost all the fatty acids and 2-MAGs are absorbed, as both are water soluble. Water-insoluble lipids are poorly absorbed: only 30–40% of dietary cholesterol is absorbed. The bile salts pass into the ileum, where they themselves are reabsorbed and transferred back to the liver. This is called the enterohepatic circulation (Chapter 17).

Clinical box An alcoholic man with central abdominal pain: pancreatitis

A 56-year-old man with a long history of alcohol abuse presented with chronic central abdominal pain, weight loss and diarrhea. He described his bowel motions as pale and greasy, and difficult to flush away. The abdominal radiograph revealed epigastric calcification in the area of the pancreas, and CT scanning revealed an atrophic calcified pancreas. The stool sample sent for fecal elastase quantification revealed this to be significantly reduced. Treatment was initiated with pancreatic enzyme supplements, resulting in resolution of his diarrhea and weight gain.

Comment.

Acute pancreatitis is a serious and life-threatening illness, caused by gallstones blocking the pancreatic duct, alcohol abuse, or more rarely by drugs such as azathioprine, viruses such as mumps, or hypertriglyceridemia. Patients present with severe abdominal pain,

nausea and vomiting. The most important biochemical marker of pancreatitis is **increased amylase activity in serum**. Increased activity of lipase and a decrease in serum calcium can also occur.

Chronic pancreatitis is a consequence of long-term inflammation and leads to malnutrition and **steatorrhea** (excessive fecal fat). It is also associated with failure of endocrine pancreatic function, leading to hyperglycemia.

Advanced concept box Exocrine function of the pancreas

The pancreas has two distinct functional roles: an **exocrine** function, i.e secretion of digestive enzymes via the pancreatic duct; and an **endocrine** function, i.e secretion of insulin, glucagon and other hormones by the islets of Langerhans. (Chapter 21) These hormones are responsible for glycemic control and aspects of gastrointestinal function.

Exocrine secretions flow into the pancreatic duct, which empties into the duodenum along with the common bile duct from the liver and the gall bladder. Food entering the duodenum stimulates the secretion of cholecystokinin, and this in turn stimulates pancreatic enzyme production and secretion. The acidity of the stomach contents entering the duodenum stimulates the release of another hormone, secretin, which triggers the secretion of bicarbonate-rich pancreatic fluid, which neutralizes the acidity in the duodenum.

The pancreas secretes enzymes which digest carbohydrates, lipids and proteins. Pancreatic **amylase** digests carbohydrates to oligo-and monosaccharides; **lipase** digests triacylglycerols while **cholesteryl esterase** yields free cholesterol and fatty acids; finally, **proteases and peptidases** break down proteins and peptides. To prevent the powerful proteases breaking down the pancreas itself (autodigestion), they are secreted as proenzymes (Chapter 6) and are activated in the intestinal lumen.

The fate of fatty acids depends on their chain length

Medium-and short-chain fatty acids (less than 10 carbon atoms) pass directly through the enterocytes into the hepatic portal system. In contrast, fatty acids containing more than 12 carbon atoms are bound to a fatty acid-binding protein within the cell, and are transferred to the rough endoplasmic reticulum for resynthesis into TAGs. The glycerol required for this process is obtained from the absorbed 2-MAGs (the MAG pathway; Fig. 10.8), from the hydrolysis of 1-MAG (which yields free glycerol), or from the glycerol-3-phosphate obtained from glycolysis (the phosphatidic acid pathway). Glycerol produced in the intestinal lumen is not used in the enterocyte for TAG synthesis, and passes directly to the portal vein.

Triacylglycerol synthesis requires activation of fatty acids

All absorbed long-chain fatty acids are reutilized to form TAG before being transferred to chylomicrons. Fatty acid activation is accomplished by the acyl-CoA synthase. Chylomicrons are assembled within the rough endoplasmic reticulum before being released by exocytosis into the intercellular space. They leave the intestine via lymph (Chapter 18).

Digestion and absorption of proteins

The gut receives 70–100 g dietary proteins per day and 35–200 g of endogenous proteins. The latter (mostly enzymes) are either secreted into the gut or shed from the epithelium as a result of cell turnover. The digestion and absorption of protein is extremely efficient: of this large load, only 1–2 g of nitrogen, equivalent to 6–12 g of protein, are lost in the feces daily.

Proteins are hydrolyzed by peptidases

Proteins are broken down by hydrolysis of peptide bonds by peptidases. These enzymes can either cleave internal peptide bonds (endopeptidases) or cleave off one amino acid at a time from either the –COOH (carboxypeptidases) or $–NH_2$ (aminopeptidases) terminal of the polypeptide. Endopeptidases break down large polypeptides to smaller oligopeptides, which can subsequently be acted upon by the exopeptidases to produce amino acids and di-and tripeptides, the final products of protein digestion which are absorbed by the enterocytes. Depending on the source of the peptidases, the protein digestion can be divided into gastric, pancreatic and intestinal phases (Fig. 10.9).

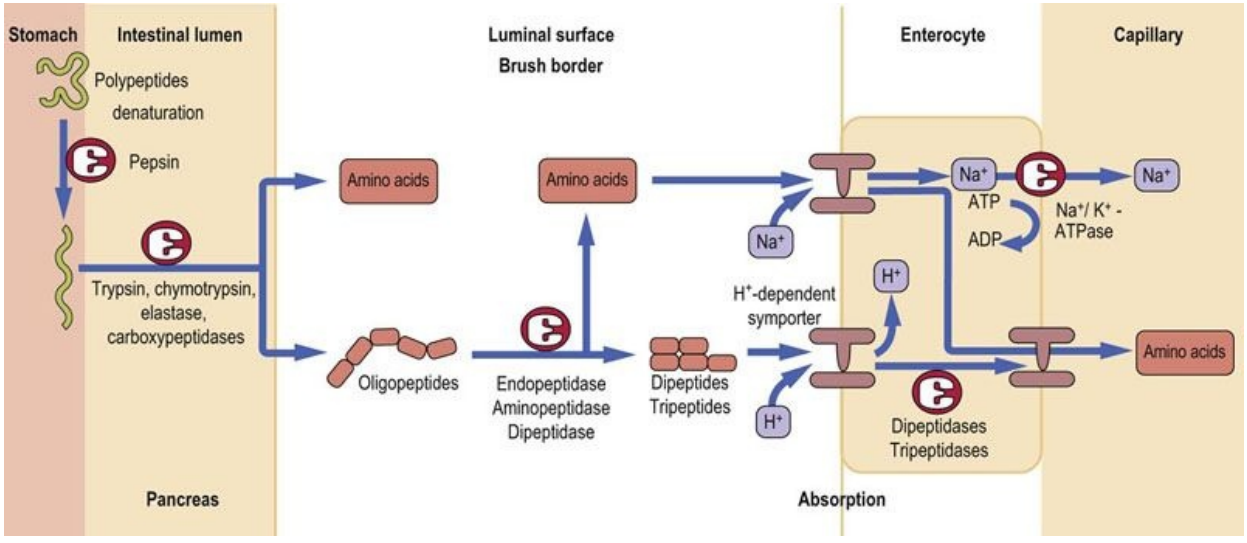


FIG. 10.9 Digestion and absorption of dietary proteins.
The preliminary stage is protein denaturation, which takes place in the stomach. The peptide bonds between amino acids are hydrolyzed by endo-and exopeptidases. Single amino acids and di-and tripeptides are absorbed using specific transport systems located

in the enterocyte membrane.

Protein digestion begins in the stomach

In the stomach, the secreted HCl reduces the pH to 1–2, with consequent denaturation of dietary proteins. Denaturation unfolds polypeptide chains, making proteins more accessible to proteases. In addition, the chief cells of the gastric mucosa secrete pepsin. It is released as inactive precursor-, pepsinogen, and is activated by either an intramolecular reaction (autoactivation) at pH below 5.0 or by active pepsin (autocatalysis). At pH above 2.0 the liberated peptide remains bound to pepsin and acts as an inhibitor of its activity. This inhibition is removed by either a drop in pH below 2.0 or by further pepsin action. The products of digestion of proteins by pepsin are large peptide fragments and some free amino acids. They stimulate cholecystokinin release in the duodenum, in turn triggering the release of the main digestive enzymes by the pancreas, as well as the contraction of the gallbladder to release bile.

Proteolytic enzymes are released from the pancreas as inactive zymogens, in a manner similar to pepsinogen

Duodenal enteropeptidase converts trypsinogen to the active trypsin. This enzyme is then capable of autoactivation. It also activates all other pancreatic zymogens (chymotrypsin, elastase and carboxypeptides A and B). Because of this prime role of trypsin in activating other pancreatic enzymes, its activity is controlled within the pancreas and pancreatic ducts by a low-molecular-weight inhibitory peptide.

Pancreatic proteases cleave peptide bonds in different locations in a protein

Trypsin cleaves proteins at arginine and lysine residues, **chymotrypsin** at aromatic amino acids, and **elastase** at hydrophobic amino acids. The combined effect is to produce an abundance of free amino acids and low-molecular-weight peptides of 2–8 amino acids in length. Alongside protease secretion, the pancreas also produces copious amounts of **sodium bicarbonate**. This neutralizes the stomach as they pass into the duodenum, thus promoting pancreatic protease activity.

Final digestion of peptides is dependent on small intestinal peptidases

The final digestion of di-and oligopeptides is carried out in the small intestine by membrane-bound endopeptidases, dipeptidases and aminopeptidases. The end-products of this are free amino acids, and di-and tripeptides, which are then absorbed across the enterocyte membrane by specific carrier-mediated transport. Within the enterocyte, di-and tripeptides are further hydrolyzed to their constituent amino acids. The final step is the transfer of free amino acids out of the enterocyte into the portal blood.

Clinical box Diagnostic approaches to malabsorption

Malabsorption can be caused by cystic fibrosis, and lactase or other specific digestive enzyme deficiencies. The most common cause of carbohydrate malabsorption is lactose deficiency. Pancreatic insufficiency is also an important cause, as is inadequate amount of bile. Malabsorption can also result from the damage to the intestinal wall by, for instance, lymphoma, inflammatory bowel disease or radiotherapy. Important causes are surgical interventions: gastrectomy, pancreatectomy and resection of large fragments of the small bowel.

Rare endocrine causes include Zollinger–Ellison syndrome and abetalipoproteinemia (a rare disorder of lipoprotein metabolism in which chylomicron assembly is impaired).

The signs of malabsorption are **chronic diarrhea, steatorrhea and loss of weight**, and-in children – **failure to thrive**. Its complications result from the inadequate intake of nutrients, vitamins or trace metals (Chapter 11).

Diagnosis of malabsorption syndromes involves the conventional hematology and biochemistry tests and testing for active inflammatory processes (C-reactive protein), as well as stool examination and stool culture. Specialist tests include testing for vitamin deficiencies. Imaging-based investigations such as abdominal ultrasound and CT scan can be performed, and portions of the GI tract can be visualized through esophago-gastro-duodenoscopy. Biopsies can be taken from the stomach, duodenum

and the small bowel.

The hydrogen breath tests are used in the diagnosis of carbohydrate malabsorption. Older tests for malabsorption include the xylose absorption test and lactose absorption test. If pancreatic insufficiency is suspected, the fecal excretion of enzymes such as elastase and lipase can be assessed, and endoscopic retrograde pancreatography (ERCP) performed.

Advanced concept box Active transport of amino acids into intestinal epithelial cells

Mechanisms of active transport of amino acids and di-or tripeptides into intestinal epithelial cells are similar to those described for glucose. At the brush border membrane, Na^+-dependent symporters mediating amino acid uptake are linked to ATP-dependent pumping out of Na^+ at the basolateral membrane. A similar H^+-dependent symporter is present on the brush border surface for di-and tripeptide transport into the cell. Na^+-independent transporters are present on the basolateral surface, allowing facilitated transport of amino acids into the portal vein.

At least six specific symporter systems have been identified for the uptake of L-amino acids from the intestinal lumen:

- Neutral amino acid symporter for amino acids with short or polar side chains (Ser, Thr, Ala).
- Neutral amino acid symporter for aromatic or hydrophobic side chains (Phe, Tyr, Met, Val, Leu, Ileu).
- Imino acid symporter (Pro, OH-Pro).
- Basic amino acid symporter (Lys, Arg, Cys).
- Acidic amino acid symporter (Asp, Glu).
- β-amino acid symporter (β-Ala, Tau).

These transport systems are also present in the renal tubules and defects in their molecular structure can lead to disease (e.g.

Hartnup disease, an inherited disorder with defects of intestinal amino acid absorption and urinary loss of neutral amino acids described in the box on p. 84).

Summary

■ Digestion is a series of processes which prepare food for absorption.
■ Digestion and absorption of foods make the metabolic fuels available to the organism.
■ Carbohydrates are digested to simple sugars.
■ Fats are hydrolyzed to di-and monoglycerides.
■ Proteins are hydrolyzed to di-and tripeptides and free amino acids.
■ Defects in these mechanisms result in a variety of malabsorption and food intolerance syndromes.

Active learning

1. Describe the process of digestion of starch.
2. Discuss the possible complications of persistent vomiting.
3. Which hormones aid digestion?
4. List the secretory products of the stomach.
5. Outline the mechanisms of sugar transport in the small intestine.
6. What is the role of micelles in the digestion of fat?

Further reading

Baumgart, DC, Sandborn, WJ. Crohn's disease. *Lancet*. 2012; 380:1590–1605.

Binder, HJ. Causes of chronic diarrhea. *N Engl J Med*. 2006; 355:236–239.

Broer, A, Cavanaugh, JA, Rasko, JEJ, *et al.* The molecular basis of neutral aminoacidurias. *Pflugers Arch Eur J Physiol*. 2006; 451:511–517.

Drozdowski, LA, Thomson, ABR. Intestinal sugar transport. *World J Gastroenterol*. 2006; 12:1657–1670.

Harris, JB, LaRocque, RC, Qadri, F, *et al.* Cholera. *Lancet*. 2012; 379:2466–2476.

Hou, W, Schubert, ML. Gastric secretion. *Curr Opin Gastroenterol*. 2006; 22:593–598.

Kunzelmann, K, Mall, M. Electrolyte transport in the mammalian colon: mechanisms and implications for disease. *Physiol Rev*. 2002; 82:245–289.

Websites and downloads

Diarrhoea: why children are still dying and what can be done, WHO 2009. www.who.int/maternal_child_adolescent/documents/9789241598415/en/index.html.

Johns Hopkins Children's Center. What is malabsorption. www.hopkinschildrens.org/tpl_rlinks_nav1up.aspx?id=5130.

Malabsorption. http://labtestsonline.org/understanding/conditions/malabsorption/.

MedlinePlus. Malabsorption. www.nlm.nih.gov/medlineplus/ency/article/000299.htm.

UNICEF rehydration solution 2006. www.unicef.org/media/media_31825.html.

CHAPTER 11

Vitamins and Minerals

Marek H. Dominiczak and John I. Broom

Learning objectives

After reading this chapter you should be able to:

- Describe fat-soluble and water-soluble vitamins.
- Discuss the actions and sources of vitamins.
- Discuss signs and symptoms of vitamin deficiencies.
- Describe the role of trace metals in metabolism.

Introduction

Vitamins and trace elements are micronutrients essential for metabolism

They form prosthetic groups of enzymes, or serve as their cofactors. They participate in metabolism of carbohydrates, fat and proteins. Some vitamins (A and D) act as hormones. Both vitamins and trace metals are also important for cell growth, proliferation and differentiation, and many of them affect immune phenomena. The requirement for vitamins depends, to some extent, on the macronutrient intake (Chapter 22).

Deficiency of a micronutrient may develop because of inadequate intake, poor absorption from the intestinal tract, inefficient utilization or increased loss, or increased demand. Such deficiencies of micronutrients lead to specific clinical syndromes. They may develop as a component of general malnutrition, may themselves be a cause of illness, or may develop during periods of increased demand such as pregnancy or the adolescent growth spurt. In old age, deficiencies may be associated with less efficient intestinal absorption (Chapter 10). They may also occur as complications of gastrointestinal surgery. Multiple micronutrient deficiencies are much more common than single ones.

Trace metals – and some vitamins, are toxic in excess

Trace elements such as cadmium, mercury, and aluminum find their way into the food chain and are cytotoxic. Essential trace elements, *e.g.* copper and manganese, may also be toxic in excess. For evaluation of trace element toxicity, tissues other than blood may need to be analyzed before a diagnosis of metal poisoning can be made.

Vitamins are divided into fat-soluble and water-soluble vitamins

Fat-soluble vitamins are A, D, E, and K, and water-soluble vitamins are B_1, B_2, B_3, B_5, B_6, B_{12}, folate, biotin and vitamin C.

Assessment of micronutrient status is difficult for several

reasons

Measurements of plasma concentrations of vitamins are inappropriate in the case of water-soluble ones, because these levels relate to the recent intake and do not reflect the overall body status. In such cases, the measurement of activities of enzymes associated with particular vitamins has been more appropriate. This is usually carried out as a stimulation test: the activity of an enzyme is measured in the absence and in the presence of a tested vitamin. A deficit is recognized if the enzyme activity is stimulated in the presence of added vitamin.

Finally, a decrease in a concentration of a nutrient in blood or plasma does not necessarily indicate a deficiency; it could be simply reflecting a metabolic response to stress or a change in physiologic state such as pregnancy.

Fat-soluble vitamins: A, D, E, K

Fat-soluble vitamins are stored in tissues

Fat-soluble vitamins are associated with body fat and are often stored in tissues, with circulating concentrations being kept relatively constant. For example, vitamin A is stored in the liver and is transported in plasma by specific binding proteins. Fat-soluble vitamins are not as readily absorbed from the diet as are water-soluble vitamins but, on the other hand, ample amounts are stored in tissues. With the exception of vitamin K they do not act as coenzymes: vitamins A and D behave like hormones. Vitamins A and D, but not vitamin E or K, can be toxic in excess.

Clinical box Causes of fat malabsorption

Fat malabsorption may lead to deficiencies of vitamins A, D, E and K. It may occur as a consequence of diseases of liver or galbladder, inflammatory bowel disease (Crohn's disease, celiac disease) and cystic fibrosis.

Vitamin A

'Vitamin A' is a generic term for **retinol, retinal and retinoic acid**, all of which are found in animals. Retinal and retinoic acid are the active forms of vitamin A. The term 'retinoids' has been used to define these substances as well as other synthetic compounds associated with vitamin A-like activity.

The provitamin of vitamin A is a plant pigment **β-carotene** and other carotenoids. β-Carotene is converted in the small bowel to *all-trans* retinal by the action of the β-carotene dioxygenase. Further metabolism in the enterocytes produces retinol and retinoic acid, which are transported to the liver for storage (Fig. 11.1). β-Carotene is water soluble and is found in plant food. Good sources of β-carotene are dark-green and yellow vegetables and tomatoes.

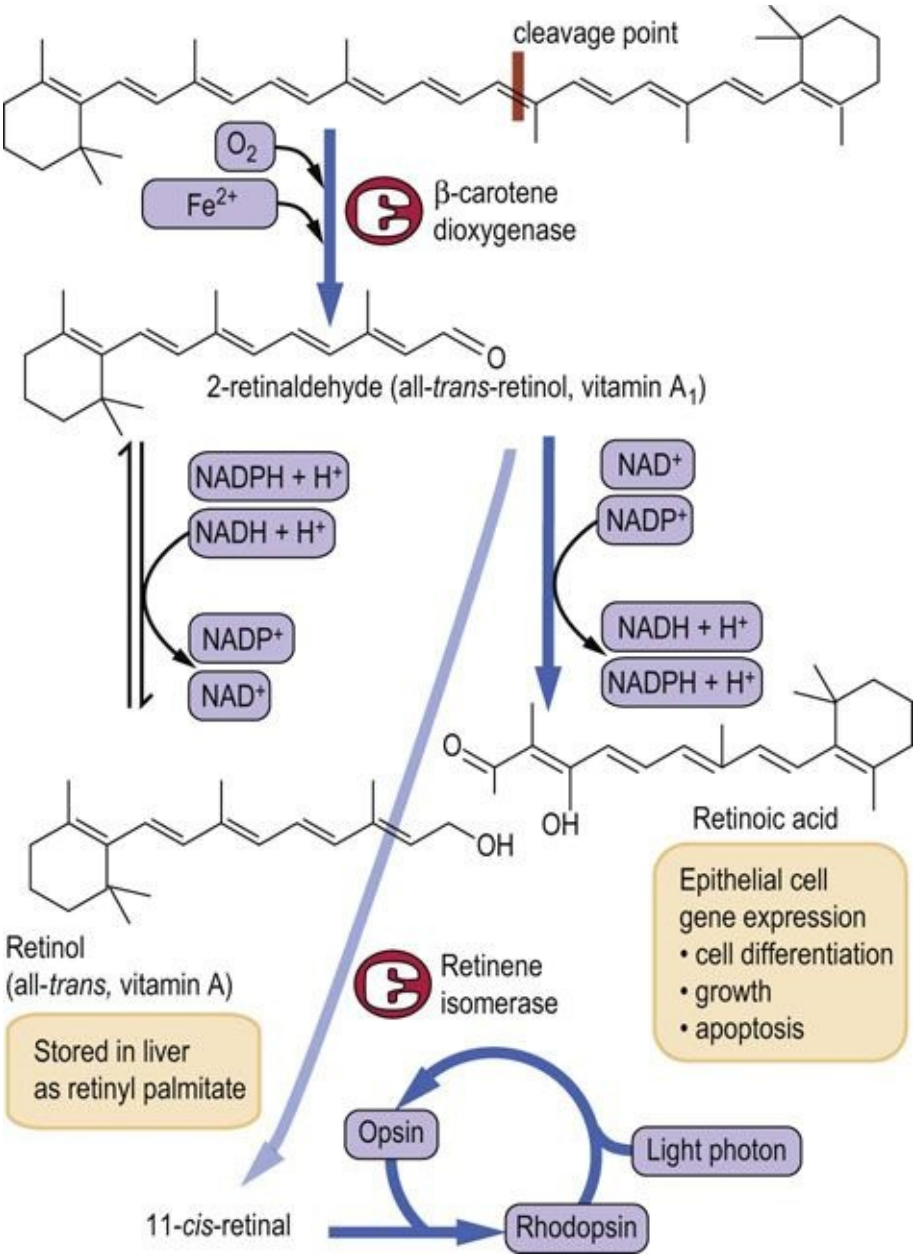


FIG. 11.1 Structure, metabolism and function of vitamin A.
Conversion of retinaldehyde to retinoic acid is irreversible. (See also Chapter 40.)
Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012, Card 39.

All dietary forms of vitamin A are converted to retinol

All dietary forms of vitamin A are converted to retinol. Conversion of carotenoids to vitamin A is rarely 100% efficient and the potency of foods is described in retinol activity equivalents (RAE); 1 µg of retinol is equivalent to 12 µg β-carotene, or 24 µg of other carotenes. Retinol in turn can be converted to

retinal and retinoic acid. Liver, fish oil, egg yolk, butter, and milk are good sources of preformed retinol and retinoic acid.

Vitamin A is stored in the liver and needs to be transported to its sites of action

Vitamin A is stored in the liver in the form of retinol and retinyl esters (retinol palmitate), bound to the cytosolic retinol-binding proteins (CRBP). The stores in the liver comprise approximately 1 year's supply. Retinol is secreted from the liver bound to serum retinol-binding protein (RBP). Retinoic acid is thought to be transported to cells bound to either albumin or to a specific retinoic acid-binding protein (RABP). Retinol is taken up by cells via a membrane receptor.

Retinoic acid is a signaling molecule. It interacts with ligand-activated transcription factors, known as the nuclear retinoid receptors. The retinoid acid receptors (RARs) bind all-*trans*- and 9-*cis*-retinoic acid, while the so-called rexinoid receptors (RXRs) bind the 9-*cis* isomer only. These receptors can form heterodimers. RXR-type receptors can also interact with other nuclear receptors such as those for vitamin D_3, thyroid hormones or peroxisome proliferator-activated receptors (PPARs). Retinoic acid also has a role in the growth and differentiation of cells, including the ones in the central nervous system.

Vitamin A deficiency presents as night blindness

Vitamin A is a component of the visual pigment rhodopsin, which is found in the rod cells of the retina and is formed by the binding of 11-*cis*-retinal to the apoprotein opsin. When rhodopsin is exposed to light, it is bleached; the retinal dissociates and is isomerized and reduced to all-*trans*-retinol (Fig. 11.1). This reaction is accompanied by a conformational change and elicits a nerve impulse perceived by the brain as light. The rod cells are responsible for vision in poor light.

Vitamin A deficiency presents as **defective night vision or night blindness** (it is the most common symptom of vitamin A deficiency in children and pregnant women). Since vitamin A affects growth and differentiation of epithelial cells, its deficiency produces defective epithelialization and keratomalacia – corneal softening and opacity. Severe vitamin A deficiency leads to permanent blindness. Vitamin A deficiency is the commonest cause of blindness in the world.

Subclinical deficiency may also lead to increased susceptibility to infections.

Severe vitamin A deficiency occurs mostly in the developing world but it is also fairly common in patients with severe liver disease or fat malabsorption (e.g. cystic fibrosis).

Pregnant and lactating women are also prone to vitamin A deficiency. The most vulnerable group are premature infants and, in the developing countries, breastfed children of mothers who themselves are vitamin A deficient.

Vitamin A is toxic in excess

Vitamin A is toxic in excess, with symptoms including increased intracranial pressure, headaches, double vision, dizziness, bone and joint pains, hair loss, dermatitis, hepatosplenomegaly, and diarrhea and vomiting. It is virtually impossible to develop vitamin A toxicity by ingesting normal foods; however, toxicity may result from the use of vitamin A supplements. Increased intake of vitamin A is also associated with teratogenicity and it should be avoided during pregnancy.

Vitamin D

Vitamin D (calciol) is a hormone. It is a group of closely related sterols produced by the action of ultraviolet light (wavelength 290–310 nm) on provitamins (ergosterol in plants and 7-dehydrocholesterol in animals). 7-Dehydrocholesterol is synthesized in the liver and is found in the skin. The products of the photolytic reaction are ergocalciferol (vitamin D_2) and cholecalciferol (vitamin D_3). They are equipotent and both are converted to a series of hydroxylated derivatives, in the liver to 25-hydroxycholecalciferol ($25(OH)D_3$; calcidiol) and then in the kidney, to the active compound 1α-,25-dihydroxycholecalciferol ($1,25(OH)_2D_3$; calcitriol). Vitamin D metabolism and action are described in Chapter 26.

Vitamin D also influences genes involved in cell proliferation, differentiation and apoptosis. It modulates growth, participates in immune function, and is anti-inflammatory.

Vitamin D is the only vitamin that is not usually required in the diet

It is only under conditions of inadequate exposure to sunlight that dietary intake of vitamin D is required. Most of the intake is via milk and other fortified

foodstuffs. Fish oil, beef, egg yolks and liver are also rich in vitamin D. The requirements are greater in winter due to lower exposure to sunlight.

Deficiency of vitamin D produces rickets in children and osteomalacia in adults

Vitamin D deficiency may be caused by insufficient exposure to sunlight or increased metabolism of vitamin D due to low calcium intake or absorption. Deficiency may also develop in kidney disease, fat malabsorption (in cystic fibrosis, Crohn's disease and after gastric bypass).

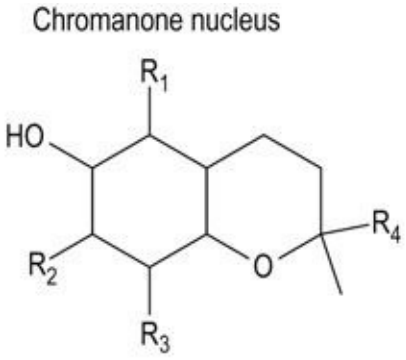
Rickets is characterized by soft pliable bones due to defective mineralization secondary to calcium deficiency. The characteristic bowing of the leg bones and the formation of the ‘rickety rosary’ around costochondral junctions result. In the adult, demineralization of preexisting bones takes place, increasing susceptibility to fractures. Vitamin D deficiency is also characterized by low circulating concentrations of calcium and an increased serum alkaline phosphatase activity (Chapter 26).

Vitamin D is toxic in excess

Vitamin D excess causes enhanced calcium absorption and bone reabsorption, leading to hypercalcemia and metastatic calcium deposition. The symptoms are anorexia, weight loss and polyuria, There is also a tendency to develop kidney stones because of the hypercalciuria secondary to hypercalcemia.

Vitamin E

Dietary vitamin E is a mixture of several compounds, called tocopherols. Ninety percent of vitamin E present in human tissues is in the form of the natural isomer, α-tocopherol (Fig. 11.2). It is involved in the immune function, and also in cellular signaling and gene expression. α-Tocopherol inhibits the activity of protein kinase C (PKC) and affects cell adhesion as well as arachidonic acid metabolism. In European folklore, vitamin E has been associated with fertility and sexual activity. This is certainly true in other animal species where vitamin E plays a role in sperm production and egg implantation, but it is not the case in man. It is absorbed from the diet in the small intestine with lipids, and there is no specific transport protein. It is packed into the chylomicrons and in the circulation it is associated with lipoproteins.



R_1–R_3			R_4
α-tocopherol	R_1,R_2,R_3,	Me	$-CH_2(CH_2-CH_2-\underset{\displaystyle CH_3}{\underset{\vert}{CH}}-CH_2)_3-$
β-tocopherol	R_1,R_3,	Me	
γ-tocopherol	R_2,R_3,	Me	
δ-tocopherol	R_2,R_3,	Me	

FIG. 11.2 Structure of vitamin E family (tocopherols).
R_1–R_3 can be methylated in a variety of combinations. R_4 is a polyisoprenoid chain. Me, methyl.

Vitamin E is a membrane antioxidant

Vitamin E is the most abundant natural antioxidant and, owing to its lipid solubility, it is associated with all lipid-containing structures: membranes, lipoproteins and fat deposits (Fig. 37.9). It protects lipids from oxidation by the reactive oxygen species.

The richest sources of naturally occurring vitamin E are vegetable oils, nuts and also green leafy vegetables.

Fat malabsorption reduces vitamin E absorption

Apart from fat malabsorption, abetalipoproteinemia may also cause vitamin E deficiency. Deficiency may develop as a result of low vitamin E intake in pregnancy and newborn infants (mostly in preterm infants fed with formula milk with low vitamin E content). Deficiency of vitamin E in premature infants causes **hemolytic anemia, thrombocytosis and edema and also peripheral neuropathy, myopathy and ataxia**. There is little evidence in support of vitamin E toxicity.

Vitamin K

Vitamin K is a group of compounds, varying in the number of isoprenoid units in their side chain. Vitamin K circulates as phylloquinone (vitamin K_1) and its hepatic stores are in the form of manaquinones (vitamin K_2). The structure, nomenclature and sources of the vitamins K are outlined in Figure 11.3. Absorption of vitamin K depends on the ability to absorb fat.

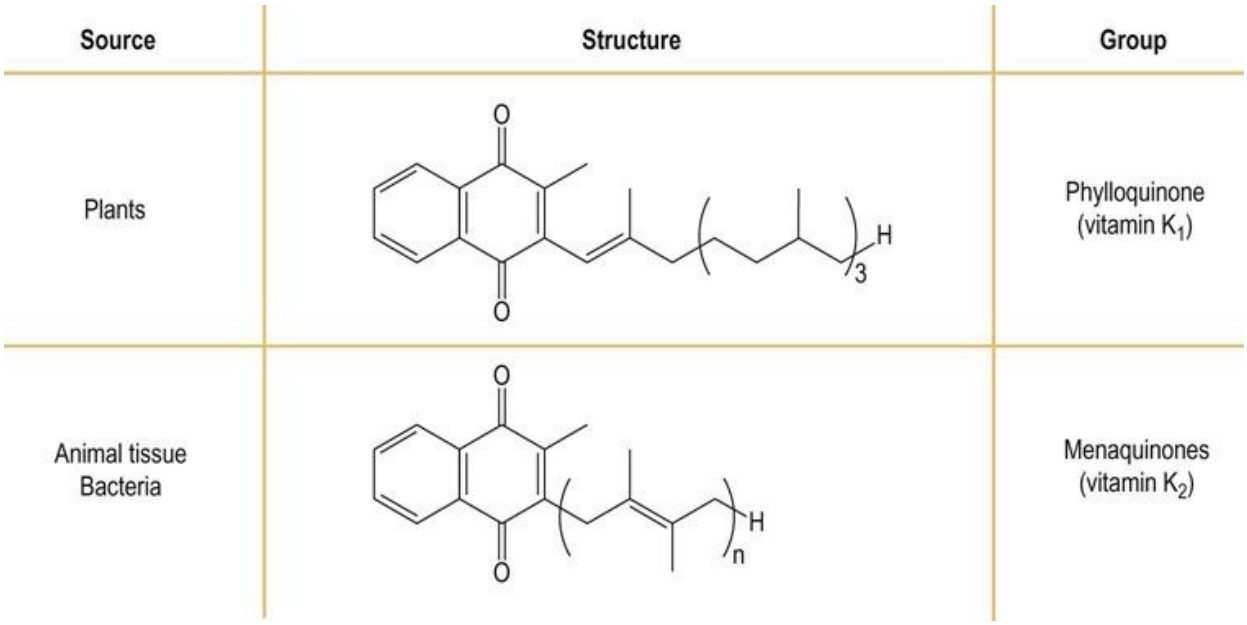


FIG. 11.3 Structures of the different forms of vitamin K.

Vitamin K is necessary for blood clotting

Vitamin K is required for post-translational modification of coagulation factors (factors II, VII, IX, and X; Fig. 7.3). All these proteins are synthesized by the liver as inactive precursors and are activated by the carboxylation of specific glutamic acid (Gla) residues by a vitamin K-dependent enzyme (Fig. 11.4). Prothrombin (factor II) contains 10 of these carboxylated residues and all are required for this protein's specific chelation of Ca^{2+} ions during its function in the coagulation process. Recently, other proteins containing vitamin K-dependent Gla residues, such as osteocalcin, have been identified in tissues.

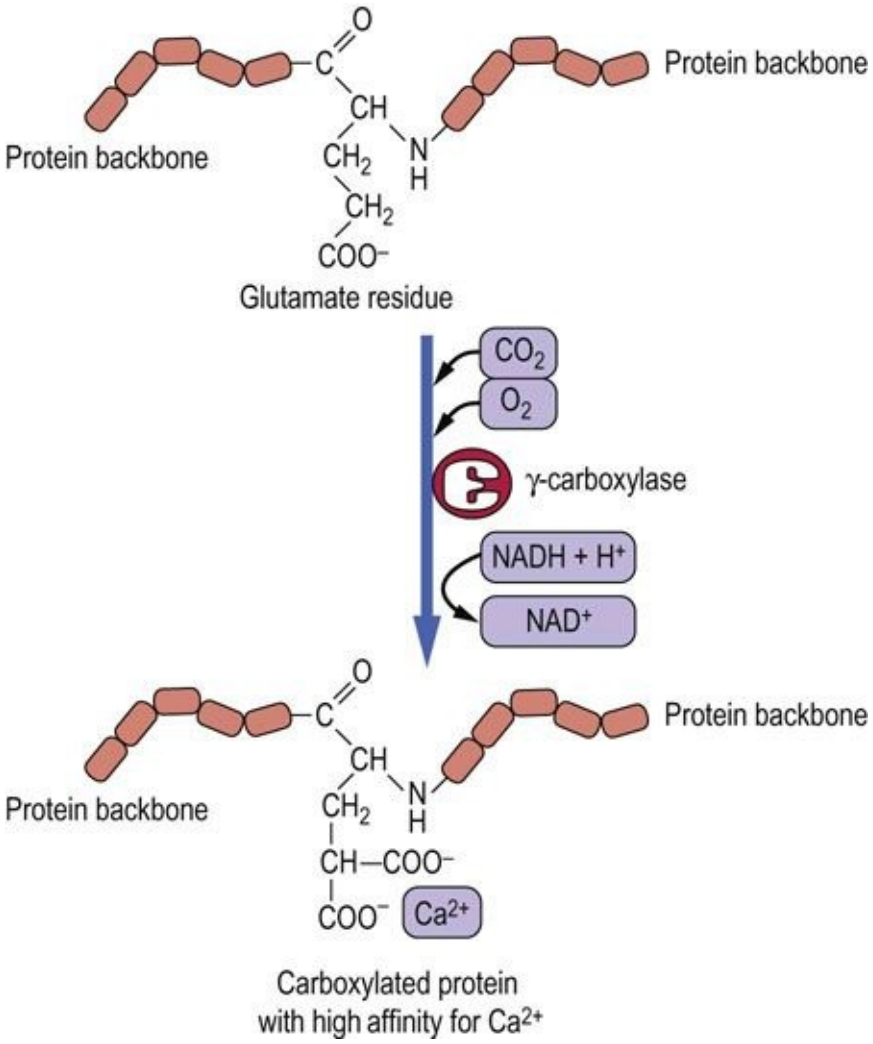


FIG. 11.4 Vitamin K-mediated carboxylation of glutamate residues. The carboxylated residues are required for Ca^{2+} chelation.

Vitamin K is widely distributed in nature: its dietary sources are green leafy vegetables, fruits, dairy products, vegetable oils and cereals. Vitamin K is also produced by the intestinal microflora.

Vitamin K deficiency causes bleeding disorders

Vitamin K supply by the intestinal microflora virtually ensures that dietary deficiency does not occur in man, except for the newborn infants. Rarely, deficiency may develop in those with liver disease or fat malabsorption. The deficiency is associated with bleeding disorders.

Premature infants are especially at risk of deficiency and

may develop the hemorrhagic disease of the newborn

Placental transfer of vitamin K to the fetus is inefficient. Immediately after birth the circulating concentration decreases. Normally, it recovers on absorption of food but this might be delayed in preterm infants. In addition, the gut of the newborn is sterile, and therefore for several days after birth there is no source of vitamin K.

Inhibitors of vitamin K action are valuable antithrombotic drugs

Specific inhibitors of vitamin K-dependent carboxylation are used in the treatment of thrombosis-related diseases, *e.g.* in patients with **deep vein thrombosis** and **pulmonary thromboembolism**, or those with **atrial fibrillation** who are at risk of thrombosis. These are the drugs of the dicoumarin group, *e.g.* **warfarin**, which inhibit the action of vitamin K (Chapter 7). Warfarin is also used as rat poison and vitamin K is the antidote for human poisoning by this agent.

Water-soluble vitamins B, C

With the exception of vitamin B_{12} the body has no storage capacity for water-soluble vitamins. As a consequence, they must be regularly supplied in the diet. Any excess is excreted in the urine. In contrast to the fat-soluble vitamins, there is no toxicity associated with B vitamins.

B-complex vitamins

B-complex vitamins are essential for normal metabolism and serve as coenzymes in many reactions in carbohydrate, fat and protein metabolism. The greater the caloric intake, the larger the requirement for B vitamins. Increased energy supply, in particular from carbohydrates, requires increased amounts of B vitamins. Diseases associated with high caloric requirement require greater intake of thiamine and other B vitamins. Therefore, beriberi (see below) might develop on a high-carbohydrate diet.

Thiamine (vitamin B_1)

Thiamine is essential for carboxylation reactions

Thiamine is important for carbohydrate metabolism. In its active form, thiamine pyrophosphate (TPP), it is a coenzyme of pyruvate dehydrogenase (the E1 enzyme in the PDH complex, Chapter 14). It participates in a similar reaction of oxidative decarboxylation of α-ketoglutarate and also in the metabolism of branched chain amino acids. It is also a coenzyme for transketolase in the pentose phosphate pathway (Chapter 12), and it is important in the production of hydrochloric acid in the stomach.

Beriberi was the first discovered ‘deficiency disease’

Severe thiamine deficiency results in **beriberi, either ‘dry’ (without fluid retention) or ‘wet’** (associated with cardiac failure with edema). Beriberi is characterized primarily by neuromuscular symptoms, and occurs in populations relying exclusively on polished rice for food. The signs and symptoms of deficiency may also be seen in the elderly or in low-income groups with poor diet.

Thiamine deficiency is associated with alcoholism

Thiamine depletion can occur quickly (within approximately 14 days), Early symptoms are loss of appetite, constipation and nausea. They may progress to depression, peripheral neuropathy and unsteadiness. Further deterioration results in mental confusion (loss of short-term memory), ataxia and loss of eye coordination. This combination, often seen in alcoholic patients, is the Wernicke–Korsakoff psychosis. Wet beriberi is particularly associated with alcoholism.

The tests used to assess the thiamine status include its direct measurement by high-pressure liquid chromatography and the 'classic' measurement of erythrocyte transketolase activity.

Riboflavin (vitamin B_2)

Riboflavin is associated with oxidoreductases

Riboflavin is attached to the sugar alcohol ribitol. The molecule is colored, fluorescent and decomposes in visible light but is heat stable. It is found in the oxidoreductases such as the flavin mononucleotide (FMN) and the flavin adenine dinucleotide (FAD), and is required for the energy metabolism of both carbohydrates and lipids (Chapter 9).

Lack of riboflavin in the diet causes a deficiency syndrome of inflammation of the corners of the mouth (angular stomatitis), the tongue (glossitis) and scaly dermatitis. Photophobia may also develop. Owing to its light sensitivity, riboflavin deficiency may occur in newborn infants with jaundice, who are treated by phototherapy. Hypothyroidism is also known to affect the conversion of riboflavin to FMN and FAD.

To determine the riboflavin status, erythrocyte glutathione reductase activity is measured.

Niacin (vitamin B_3)

Niacin is required for NAD^+ and $NADP^+$ synthesis

Niacin is a generic name for nicotinic acid or nicotinamide, both of which are essential nutrients. Niacin is active as part of the coenzyme nicotinamide adenine dinucleotide (NAD^+) and nicotinamide adenine dinucleotide phosphate ($NADP^+$), both of which participate in oxidoreductase-catalyzed reactions. The

active form of the vitamin required for the synthesis of NAD^+ and $NADP^+$ is nicotinate, and therefore nicotinamide must be deamidated before becoming available for synthesis of these coenzymes. Niacin can be synthesized from tryptophan and, hence, in the truest sense, is not a vitamin. The conversion is, however, very inefficient and cannot supply sufficient amounts of niacin. In addition, the conversion requires thiamine, pyridoxine, and riboflavin, and on marginal diets such synthesis would be problematic. The requirement for niacin is also related to energy expenditure.

Severe niacin deficiency produces dermatitis, diarrhea and dementia

Niacin deficiency initially produces a superficial glossitis but may progress to **pellagra**, which is characterized by dermatitis, sunburn-like skin lesions in areas of body exposed to sunlight and to pressure, and also by diarrhea, and dementia. Untreated pellagra is fatal. Certain drugs, *e.g.* the antituberculosis drug isoniazid, predispose to niacin deficiency. In the modern world pellagra is a medical curiosity.

On the other hand, very high doses of niacin can cause hepatotoxicity, which is reversible on withdrawal.

Pyridoxine (vitamin B_6)

Pyridoxine is important for amino acid metabolism

Vitamin B_6 is a mixture of pyridoxine (an alcohol), pyridoxal (an aldehyde), pyridoxamine, and their 5′-phosphate esters. Pyridoxine is the major form of vitamin B_6 in the diet, and pyridoxal phosphate is its active form. It is absorbed in the jejunum. During absorption some hydrolysis of the phosphates occurs. Most tissues, however, contain pyridoxal kinase, which resynthesizes the phosphorylated forms required for the synthesis, catabolism and interconversion of amino acids (Chapter 19).

Pyridoxine requirements increase with high protein intake

Pyridoxal phosphate and pyridoxamine are involved in over 100 reactions in carbohydrate (including the glycogen phosphorylase reaction), lipid, and particularly amino acid metabolism, and the metabolism of one-carbon units (see Fig. 19.2). Pyridoxine is required for the synthesis of the neurotransmitters

serotonin and noradrenaline (Chapter 41.1), and the synthesis of sphingosine, a component of sphingomyelin and sphingolipids (Chapter 28). It is also required for the synthesis of heme (Chapter 30) and it influences immune function. Because of its role in amino acid metabolism, vitamin B_6 requirements increase with protein intake.

Vitamin B_6 is present in a wide variety of foods such as fish, beef, liver, poultry, and also potatoes and fruits (but not citrus fruits).

Vitamin B_6 deficiency causes neurologic symptoms and anemia

Vitamin B_6 deficiency in its mild form causes irritability, nervousness and depression, progressing in severe deficiency to peripheral neuropathy, convulsions and coma. Severe deficiency is also associated with a sideroblastic anemia (anemia characterized by the presence of nucleated red blood cells with iron granules). Dermatitis, cheliosis and glossitis also occur. Decreased levels are observed in **alcoholism, obesity**, and in **malabsorption states** (Crohn's disease, celiac disease and ulcerative colitis), as well as in the **end-stage renal disease** and in **autoimmune conditions**.

The drug isoniazid, by binding to pyridoxine, and the oral contraceptive pill, by increasing the synthesis of enzymes requiring the vitamin, may precipitate deficiency. The debate concerning the contraceptive pill continues but it is generally accepted that there is an increased requirement for pyridoxine.

Assessment of pyridoxine status is based on the measurement of erythrocyte aspartate aminotransferase.

Biotin

Biotin is important for carboxylation reactions

Biotin serves as a coenzyme in multienzyme complexes involved in carboxylation reactions (Fig. 14.4). It is important in lipogenesis, gluconeogenesis, and in the catabolism of the branched-chain amino acids.

Biotin is normally synthesized by the intestinal flora and this meets most of the body requirements.

Symptoms of biotin deficiency include depression, hallucinations, muscle pain and dermatitis. Children with multiple decarboxylase deficiency also develop

immunodeficiency disease. Consumption of raw eggs can cause biotin deficiency because the egg-white protein avidin combines with biotin, preventing its absorption.

Vitamin B_{12}

Vitamin B_{12} is part of the heme structure

Vitamin B_{12} (cobalamin) has a complex ring structure similar to the porphyrin of heme (Chapter 30) but is hydrogenated to a greater extent. The iron at the center of the heme ring is replaced by a cobalt ion (Co^{3+}). This is the only known function of cobalt in the body. In addition, a dimethylbenzimidazole ring is also part of the active molecule (it is essential for the chelation of the cobalt ion; Fig. 11.5). and in methionine synthesis

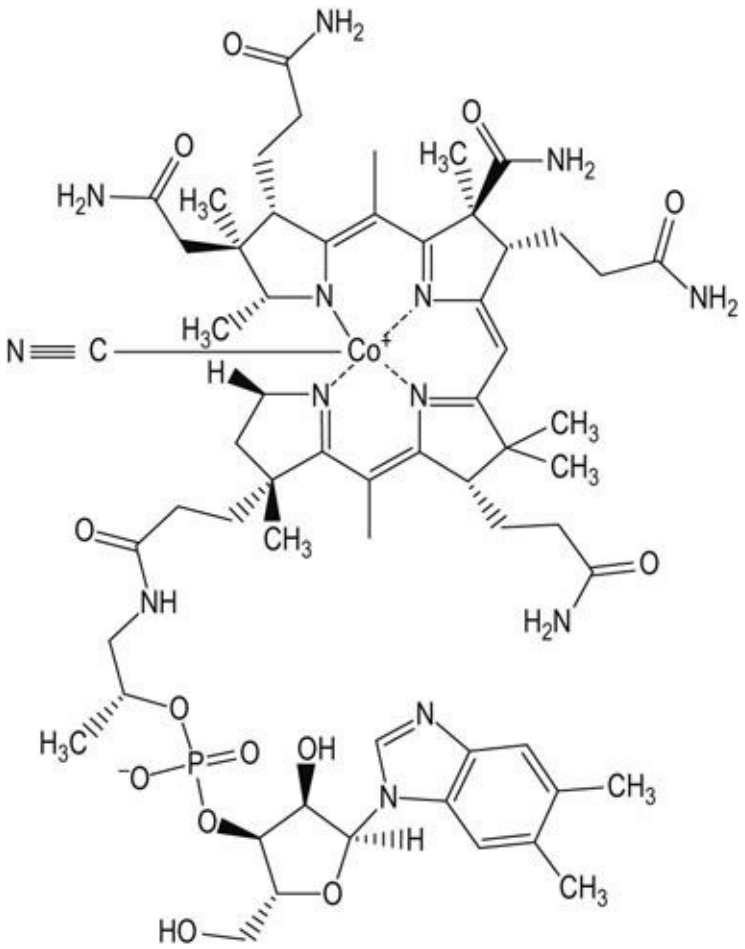


FIG. 11.5 Vitamin B_{12}.
There is a cyano-group (CN) attached to the cobalt: this is an artifact of extraction but it is also the most stable form of the vitamin and indeed is the commercially available product. The cyano group does require removal for conversion to the active form of the vitamin.

Vitamin B_{12} participates in the nucleic acid synthesis, in the production of erythrocytes, and also in the recycling of folates. Together with folate and vitamin B_6 it controls the homocysteine metabolism, where it is a cofactor for methionine synthase, which converts homocysteine to methionine. It participates in the synthesis of the methyl donor molecule, the S-adenosylmethionine. Vitamin B_{12} is required in only one further reaction, that of L-methylmalonyl-CoA mutase, which converts methylmalonyl-CoA to succinyl-CoA. The coenzyme form of the vitamin in this case is 5′-deoxyadenosyl cobalamin.

Vitamin B_{12} requires the intrinsic factor for its absorption

There is a specific mechanisms exist for absorption and transport of cobalamin (Fig. 11.6). Cobalamin is released from foods by a gastric protease and HCl. It is bound to the intrinsic factor secreted by the parietal cells in the stomach, and is absorbed in the distal ileum by receptor-mediated endocytosis. Vitamin B_{12} is excreted in the bile, and there is a marked enterohepatic circulation.

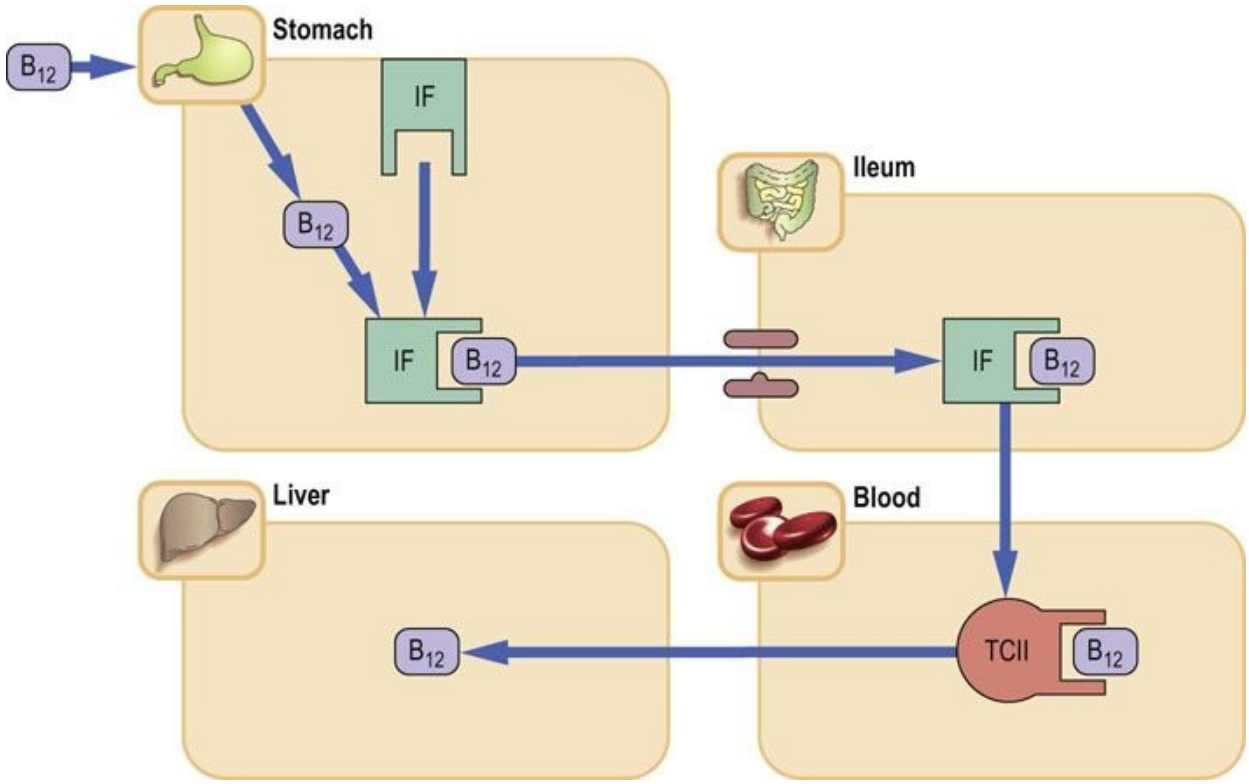


FIG. 11.6 Digestion, absorption and transport of vitamin B_{12}.
Simple diffusion of free vitamin B_{12} across the intestinal membrane accounts for 3% of transported vitamin, and complexing with intrinsic factor (IF) accounts for 97%. Vitamin B_{12} derivatives are released from food by peptic digestion in the stomach and bind to IF secreted by the parietal cells of the gastric mucosa. The IF–B_{12} complex binds to specific receptor sites on the ileal mucosa. The rate-limiting factor for vitamin B_{12} absorption is the number of ileal receptor sites. Other transport proteins, transcobalamin I, II and III (TC I, II and III) and R-proteins, are involved in the delivery or storage of the cobalamins. The latter are secreted by the salivary glands and gastric mucosa.

Vitamin B_{12} is only present in animal products

Vitamin B_{12} is synthesized solely by bacteria. It is absent from all plants but is concentrated in the livers of animals in three forms: methylcobalamin, adenosylcobalamin, and hydroxycobalamin. It is found only in animal products such as fish, dairy products, meats, and particularly organ meats such as liver and kidney. In the past, liver was used in the treatment of deficiency states. Fortified breakfast cereals also contain the vitamin. Vegans are therefore at risk of developing a dietary deficiency of vitamin B_{12}.

Vitamin B_{12} deficiency causes pernicious anemia

Vitamin B_{12} deficiency is characterized by anemia, fatigue, constipation, weight loss, diarrhea and neurological symptoms such as numbness and tingling, loss of balance, confusion, mood disturbances and dementia. Deficiency can occur through several mechanisms. The most common one is pernicious anemia, an autoimmune condition that results in gastric atrophy and the lack of the intrinsic factor, which prevents the vitamin absorption in the terminal ileum. Pernicious anemia affects 1–3% of older adults. The intrinsic factor deficiency can also be caused by gastric surgery and by bariatric (weight loss) surgery. A similar situation, albeit caused through a different mechanism, arises upon surgical removal of the ileum, for instance in Crohn's disease (Chapter 10).

Deficiency may also be caused by hypochlorhydria associated with age (Table 11.1).

Table 11.1

Causes of vitamin B_{12} deficiency

Mechanism	Time to develop clinical deficiency (years)
Vegan diet	10–12
Intrinsic factor failure	1–4
Ileal dysfunction	Rapid

Advanced concept box Vitamin B_{12} transport proteins

The intrinsic factor (IF) is a glycoprotein. Other cobalamin-binding proteins, R-proteins, secreted by the salivary glands and stomach, are also glycoproteins and along with transcobalamin (TC) I and III are termed cobalaphilins. The third type of protein, also a glycoprotein, is TCII. All three classes of B_{12}-transport proteins are single polypeptide chains (340–375 amino acid residues), and have a single binding site for cobalamin. They do not, however, immunologically cross-react with each other, and are coded for by different genes.

At an acid pH, R-proteins bind cobalamin more strongly than IF. In contrast to IF, they are normally degraded by pancreatic proteinases. Thus, in pancreatic disease where R-proteins are not

degraded, there is less cobalamin available to bind to IF, with loss of absorptive capacity for this vitamin.

■ In the final stages of the absorption process, the IF molecule binds to the ileal receptor in the presence of Ca^{2+} and at neutral pH. As the IF–B_{12} complex crosses the ileal mucosa, IF is released and the B_{12} is transferred to a plasma transport protein TCII. Other cobalamin-binding proteins, TCI and possibly TCIII, exist in the plasma and liver. In the liver, they provide storage forms of the vitamin, a situation that is unique for water-soluble vitamins.

■ Cobalamin bound to TCII and is taken up by the tissues. The major circulating form is methylcobalamin. In the liver, the main form is 5'-deoxyadenosyl cobalamin, which accounts for 70%, and methylcobalamin constitutes 3% of the total amount. The TCII–cobalamin complex delivers cobalamin to the tissues, where it binds to specific cell surface receptors. It enters the cell by endocytosis, ultimately releasing the cobalamin as hydroxycobalamin. Conversion of hydroxycobalamin to methylcobalamin occurs in the cytosol. TCII is also thought to be necessary for the delivery of vitamin B_{12} to the central nervous system.

The function of vitamin B_{12} needs to be considered together with folate

Functions of vitamin B_{12} and folate are interrelated, and deficiency of either produces the same signs and symptoms. The reaction, which involves both these vitamins is the conversion of homocysteine to methionine (Fig. 11.7).

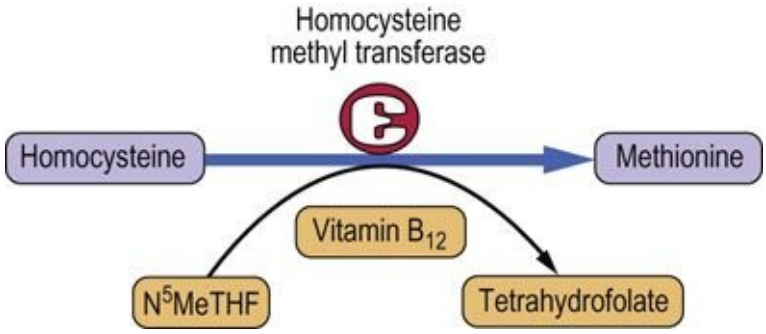


FIG. 11.7 The 'tetrahydrolate trap'.
Vitamin B_{12} and folate are involved in the conversion of homocysteine to methionine. An absence of vitamin B_{12} inhibits the reaction and leads to the build-up of N^5-methyltetrahydrofolate (N^5MeTHF).

Megaloblastic anemia characteristic of vitamin B_{12} deficiency is probably due to a secondary deficiency of reduced folate, and a consequence of the accumulation of N^5-methyltetrahydrofolate. A neurologic presentation can also develop in the absence of anemia. This is known as **subacute combined degeneration of the cord** and is probably secondary to a relative deficiency of methionine in the cord.

Deficiency of vitamin B_{12} results in the accumulation of methylmalonic acid and homocysteine, and consequent methylmalonic aciduria and homocystinuria.

Folic acid

Folic acid derivatives are important in single carbon transfer reactions. They are necessary for the synthesis of DNA

Folic acid (pteroyl-L-glutamic acid) exists in a number of derivatives collectively known as folates. It participates in single carbon transfer reactions such as methylation (important in both metabolism and regulation of gene expression), and in synthetic pathways of choline, serine, glycine and methionine. Folic acid is also necessary for the synthesis of purines and pyrimidine thymine, and thus for the synthesis of nucleic acids. Polymorphisms associated with variants of 5,10-methylenetetrahydrofolate reductase gene, the key enzyme in folate metabolism, are associated with conditions such as colon cancer, spina bifida and adult acute lymphocytic leukemia.

Folic acid is physiologically inactive until reduced to dihydrofolic acid. Its main forms are tetrahydrofolate, 5-methyl tetrahydrofolate (N^5MeTHF), and N^{10}-formyltetrahydrofolate-polyglutamate derived from N^5MeTHF predominant in fresh food. Before polyglutamates can be absorbed, they must be hydrolyzed by glutamyl hydrolase in the small intestine. The main circulating form of folate is the monoglutamate-N^5-THF.

Folic acid is present in liver, yeast, and green leafy vegetables (spinach) and fruits, including citrus fruits. Its sources are also folic acid enriched cereals and grains.

It can be measured by high-performance liquid chromatography (HPLC).

Structural analogues of folate are used as antibiotics and anticancer drugs

Not surprisingly, rapidly dividing cells have high requirements for folate since it is necessary for the synthesis of purines and pyrimidine thymine, all required for DNA synthesis (Chapter 31). Structural analogues of folate exhibit selective toxicity towards rapidly growing cells such as bacteria and cancer cells. This is the principle behind the development of drugs known as the **folic acid antagonists**, which are used as antibiotics (e.g. trimethoprim) and anticancer agents (methotrexate).

Folate deficiency is one of the commonest vitamin deficiencies

Causes of folate deficiency include inadequate intake, impaired absorption, impaired metabolism, and increased demand.

The most common examples of increased demand are pregnancy and lactation. Folic acid requirements increase greatly as the blood volume and number of erythrocytes increase in **pregnancy**. By the third trimester of pregnancy folic acid requirements double. Megaloblastic anemia in pregnancy (other than multiple pregnancy) is rare. However, folate deficiency increases the risk of **neural tube defects**. low **birth weight** and **premature birth**. In infants it results in a decreased growth rate. Other causes of folate deficiency are alcoholism, malabsorption, dialysis and liver disease. Folate deficiencies are seen in the elderly as a result of poor diet and poor absorption.

Folate deficiency in adults causes megaloblastic anemia

Failure to synthesize methionine and nucleic acids in folate deficiency states accounts for the signs and symptoms of megaloblastic anemia, *i.e.* the presence of enlarged blast cells in the bone marrow. Macrocytic erythrocytes have fragile membranes and a tendency to hemolyze: a macrocytic anemia exists in association with a megaloblastic bone marrow. The hematologic abnormalities cannot be distinguished from vitamin B_{12} deficiency (see below). The neurologic changes are also similar. Deficiency of folate also contributes to hyperhomocysteinemia. Many symptoms such as loss of appetite, diarrhea and weakness are nonspecific.

Adequate intake of folate around conception is essential

Common practice is to provide folate supplements during pregnancy. The supplementation during the periconception period (definitions of that period are variable: the one used in clinical studies is 4 weeks before and 8 weeks after conception) prevents spina bifida, as the closure of the neural tube occurs between 22 and 28 days after conception.

Vitamin B_{12} must be supplemented during folate treatment

Folate supplementation without B_{12} supplements can mask symptoms but lead to neurologic damage.

Giving folate alone in a case of vitamin B_{12} deficiency aggravates the neuropathy. Therefore, if supplementation is required during investigation of the cause of megaloblastic anemia, after blood and bone marrow specimens have been taken to confirm the diagnosis, folate needs to be given together with vitamin B_{12}.

Panthotenic acid

Panthotenic acid forms part of the molecule of coenzyme A (CoA)

It is widely distributed in animals and plants. There is no evidence of deficiency in man, except on experimental diets.

Vitamin C

Vitamin C serves as a reducing agent. Its active form is ascorbic acid, which is oxidized during the transfer of reducing equivalents, yielding dehydroascorbic acid. The synthetic pathway and structure of vitamin C are shown in Figure 11.8 and its antioxidant activity is illustrated in Figure 37.8. It participates in the regeneration of another antioxidant vitamin – α-tocopherol. Vitamin C takes part in the synthesis of collagen and epinephrine, in steroidogenesis, degradation of tyrosine, formation of bile acids and also in the synthesis of L-carnitine and neurotransmitters. It improves absorption of nonheme iron, and participates in bone mineral metabolism. Its prime function is to maintain metal cofactors in their lower valence state, *e.g.* Fe^{2+} and Cu^{2+}. In the synthesis of collagen it is required specifically for the hydroxylation of proline (Chapter 27).

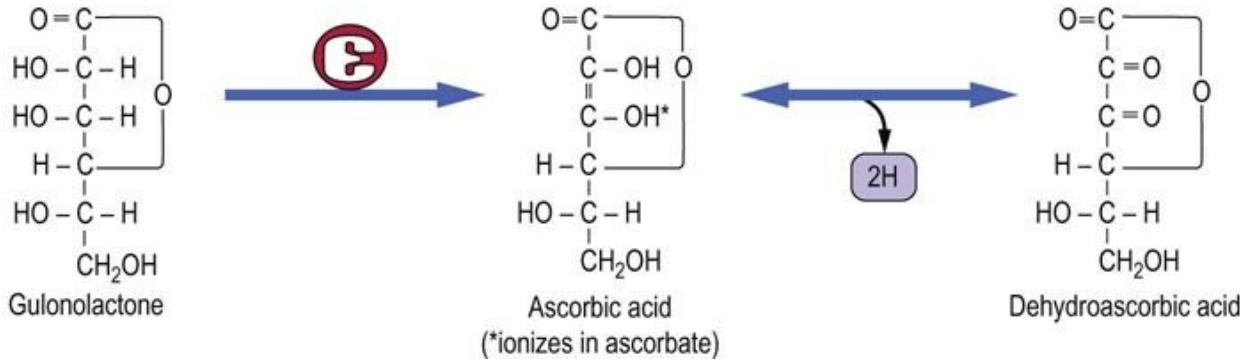


FIG. 11.8 Structure and synthesis of vitamin C (ascorbic acid).
Note that the enzyme that converts gulonolactone to ascorbic acid is absent in man and higher primate.

Vitamin C is absorbed in the intestine by a carrier-mediated, sodium-dependent transporter. It is reabsorbed in renal proximal tubules. Progressively more vitamin C is excreted in urine as intake increases.

Humans cannot synthesize ascorbic acid; therefore it is an essential nutrient

Vitamin C is labile: it is easily destroyed by oxygen, metal ions, increased pH, heat and light. Citrus, soft fruits, tomatoes and peppers are rich sources of vitamin C.

Vitamin C deficiency causes scurvy and compromises

immune function

Vitamin C deficiency causes defective collagen synthesis. **Scurvy** is characterized by capillary fragility causing subcutaneous and other hemorrhages, muscle weakness, soft, swollen, bleeding gums, loosening of teeth, poor wound healing and anemia. There is fatigue, malaise and depression. Inability to maintain bone matrix in association with demineralization results in osteoporosis.

Vitamin C deficiency resulting in the full clinical picture of scurvy is now rare, except in older individuals. Milder forms of vitamin C deficiency are more common, and their manifestation includes easy bruising and the formation of petechiae (small, pinpoint hemorrhages under the skin). Immune function is also compromised. This reduction in immunocompetence has been the basis for providing megadoses of the vitamin to prevent the common cold and also for its role in cancer prevention. No clear evidence exists, however, to substantiate these claims first made by Linus Pauling in the 1970s. Vitamin C is certainly required for normal leukocyte function, and leukocyte vitamin C levels drop precipitously during stress caused by either trauma or infection. Elderly individuals are at increased risk of deficiency as are smokers and infants fed evaporated or boiled milk.

There is no evidence that vitamin C taken in excess is toxic. Theoretically, since it is metabolized to oxalate, there is a risk of the development of renal oxalate stones in susceptible individuals. However, this has not been substantiated in practice.

Dietary supplementation of vitamins

Supplementation of some vitamins results in a clear health benefit. This includes supplementation of folic acid to women who are pregnant or are planning pregnancy, to prevent neural tube defects. Vitamin D provision to people living in areas of low sunlight has also been beneficial.

Benefits of vitamin supplementation in cancer and cardiovascular disease are uncertain

Because supplementation of folic acid and vitamin B_6 and B_{12} lowers plasma homocysteine concentration, it has been suggested that it could be beneficial for the prevention of cardiovascular disease. There also were suggestions that

supplementation of vitamins A, C and E is protective against cancer. Some observational studies suggested that the supplementation of vitamins C and E could also be useful in the prevention of cardiovascular disease. However, prospective studies of this yielded controversial results. The recommendations of the US Preventive Services Task Force published in 2003 say that 'current evidence is insufficient to recommend for or against the use of supplements of vitamins A, C, or E, multivitamins with folic acid, or antioxidant combinations for the prevention of cancer or cardiovascular disease'. Note that these recommendations do not apply to people with nutritional deficiencies, pregnant and lactating women, children, elderly persons, and people with chronic illnesses.

Vitamin supplementation can be harmful

As mentioned above, high-dose vitamin supplementation may be harmful: the example is the reduction of bone mineral density, hepatotoxicity, and teratogenicity associated with high doses of vitamin A. Supplementation of β-carotene to smokers was also harmful, resulting in an increase in lung cancer mortality.

Fruit and vegetables are the best sources of vitamins

In clinical studies, vitamins have been supplemented in a pure form, rather than as complete foodstuffs, and it might be that this is why the benefit of supplementation was not evident. Clearly, there are benefits of eating diets rich in vegetables and fruit, which are the most important sources of vitamins. There is no reason to discourage people from taking vitamin supplements apart from proven instances of toxicity.

The deficiency states of vitamins B and folate are summarized in Figure 11.9. Patients often present with multiple deficiencies; a deficiency of a single B vitamin is rare.

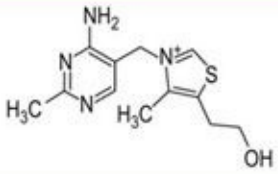
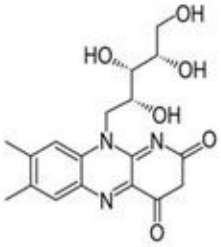
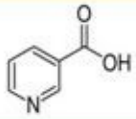
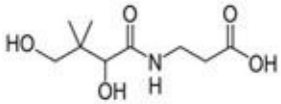
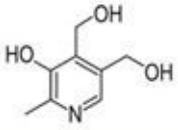
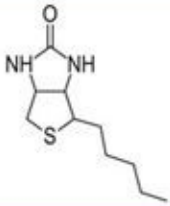
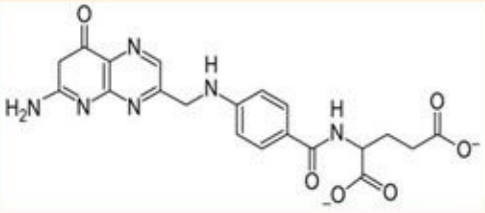
Vitamin	Structure	Deficiency disease	Food source
Thiamine (vit B_1)		Beriberi	Seeds, nuts, wheatgerms, legumes, lean meat
Riboflavin (vit B_2)		Pellagra	Meats, nuts, legumes
Niacin (vit B_3)		Pellagra	Meats, nuts, legumes
Panthothenic acid (vit B_5)			Yeast, grains, egg yolk, liver
Pyridoxine (vit B_6)		Neurologic disease	Yeast, liver, wheatgerm, nuts beans, bananas
Biotin		Widespread injury	Corn, soy, egg yolk, liver, kidney, tomatoes
Folate		Anemia	Yeast, liver, leafy vegetables
Cobalamin (vit B_{12})	Complex	Pernicious anemia	Liver, kidney, egg, cheese

FIG. 11.9 Structure, sources and deficiency diseases of B vitamins and folate.

Minerals

Major minerals present in the human body are sodium, potassium, chloride, calcium, phosphate, and magnesium

The daily body requirements for minerals range from gram (sodium, calcium, chloride, phosphorus) through milligram (iron, iodine, magnesium, manganese, molybdenum) to microgram (zinc, copper, selenium, other trace elements) amounts. Many are essential for normal biological function.

Sodium and chloride are important for the maintenance of osmolality of the extracellular fluid and cell volume (Chapter 24). Sodium participates in electrophysiologic phenomena and, together with potassium, is essential for maintaining transmembrane potential and impulse transmission (Chapter 8). **Potassium** is the main intracellular cation. Potassium is contained in vegetables and fruit, particularly bananas, and in fruit juices. Dietary potassium intake needs to be limited in renal disease because of its impaired excretion and a consequent tendency to hyperkalemia (Chapter 24). Importantly, both hyperkalemia and hypokalemia may lead to life-threatening arrhythmias.

- **Magnesium** functions as cofactor for many enzymes and is also important in the maintenance of membrane electrical potential. Its role is linked to that of potassium and calcium. It is important for skeletal development and for the maintenance of electrical potential in nerve and muscle membranes. It is also a cofactor for ATP-requiring enzymes, and is important for the replication of DNA, and for RNA synthesis. Magnesium deficiency develops in starvation and malabsorption, may be due to the loss from the gastrointestinal tract in diarrhea and vomiting, and sometimes occurs as a result of diuretic treatment and surgical procedures on the gastrointestinal tract. It is also associated with acute pancreatitis and alcoholism. Hypomagnesemia is often accompanied by hypocalcemia and hypokalemia. Magnesium deficiency leads to muscle weakness and cardiac arrhythmias.
- **Calcium and phosphate** are essential for bone metabolism, and for secretory processes and cellular signaling (Chapter 26). Calcium is present in milk and milk products, and in some vegetables. Phosphates are abundant in plant and animal cells.
- **Iodine** is essential for the synthesis of thyroid hormones (Chapter 39). The iodine content of food depends on the composition of the soil where it is grown.

Marine fish and shellfish have the highest content. It is also present in freshwater fish, meat and dairy products, as well as in legumes, vegetables and fruit.

■ **Fluoride** influences the structure of the bone and teeth enamel. In many areas fluoride is added to water to prevent teeth decay. Excess leads to teeth discoloration and fragility of bones.

Iron

Iron is important in the transfer of molecular oxygen

Iron is a component of heme in hemoglobin and myoglobin (Chapter 5). Cytochromes *a*, *b*, and *c* also contain iron (Chapter 9). Altogether, there are 3–4 g of iron in the body. Seventy-five percent of body iron is in hemoglobin and myoglobin, and 25% is stored in tissues such as bone marrow, liver, and reticuloendothelial system.

Dietary sources of iron include organ meats, poultry and fish and oysters, and also egg yolks, dried beans, dried figs and dates, and some green vegetables.

Iron is transported in plasma bound to transferrin

Iron is absorbed in the upper small intestine. Meat and ascorbic acid increase its absorption, and vegetable fiber inhibits it. It is transported in blood bound to transferrin and is stored as ferritin and hemosiderin. Transferrin is normally about 30% saturated with iron. Iron is lost through the skin and through the gastrointestinal tract.

Dietary iron is in the ferric (Fe^{3+}) form. It is reduced in the gastrointestinal tract to divalent Fe^{2+} by ascorbate and a ferrireductase enzyme located in the intestinal brush border. Fe^{2+} is transported into the cells by a divalent metal transporter (which also transports most trace metals). The iron pool within the enterocyte is controlled by the iron regulatory proteins.

Erythrocyte content of iron affects its absorption from the intestine

If erythrocytes are iron rich, the iron is stored in the enterocytes incorporated into ferritin. Otherwise, it is transported through the basolateral membrane, where one of the transport-facilitating proteins, ferroxidase, also called hephaestin, oxidizes Fe^{2+} to Fe^{3+}, which is then bound to transferrin in plasma. Transferrin is taken up in the bone marrow by erythrocyte precursors cells in a receptor-dependent manner. Within the cells, iron is released, again reduced to

Fe^{2+}, and transported to the mitochondria for incorporation into heme in the Fe^{2+} form. After destruction of the old erythrocytes by macrophages in the reticuloendothelial system, the iron is released as Fe^{2+}, re-oxidized to Fe^{3+}, and loaded back onto transferrin. The outline of iron metabolism is shown in Figure 11.10.

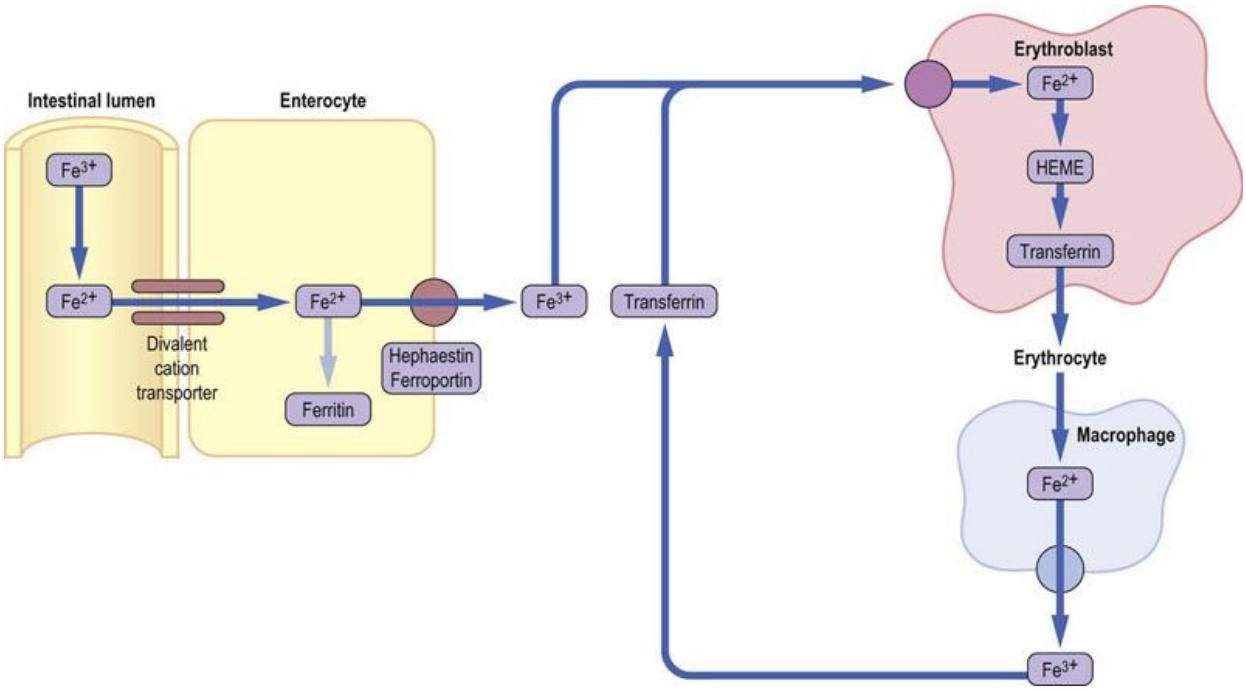


FIG.11.10 Iron metabolism.
Dietary iron is absorbed in the intestine and either stored in enterocytes as ferritin or transported out to plasma. Hephaestin and ferroportin are iron transporters located in the basolateral membrane of ethe enterocyte. In plasma, iron remains bound to transferrin. It is taken up by cells such as erythroblasts, through the mediation of the membrane transferrin receptor. In erythroblasts, iron is incorporated into heme, and then hemoglobin. Old erythrocytes are degraded by macrophages in the reticuloendothelial system. Liberated iron is released from cells and recycled bound to transferrin. Note that the dietary iron is in the ferric (Fe^{3+}) form. This is reduced to ferrous ion (Fe^{2+}) at the intestinal brush border. The transported form of iron is ferric again, but the form incorporated into heme is ferrous.

Iron deficiency causes anemia

Requirement for iron increases during growth and pregnancy. Iron deficiency results in defective erythropoiesis and in normocytic or microcytic (small

erythrocytes) hypochromic anemia. This is most likely to develop in infants and adolescents, in pregnant and menstruating women, and also in the elderly. Iron deficiency most often develops as a result of abnormal blood loss, and therefore persons who present with iron deficiency anemia always need to be investigated for causes of bleeding, particularly from the gastrointestinal tract.

Assessment of iron status includes the measurements of transferrin and ferritin in plasma, hematologic variables, and the bone marrow smear.

Humans do not have a mechanism to excrete iron and free iron is toxic.

Clinical box Hemochromatosis

Hemochromatosis is an autosomal recessive disorder resulting from the increased absorption of iron. It is the most common inherited disorder in persons of North European ancestry.

Iron accumulates in heart, liver and pancreas and can cause liver cirrhosis, hepatocellular carcinoma, diabetes, arthritis and heart failure. In the classic form of hemochromatosis, the mutated gene encodes the protein known as hereditary hemochromatosis protein (HFE), which is structurally similar to class I major histocompatibility antigens (Chapter 38). It is now known that mutations of other proteins can lead to a very similar clinical picture.

Trace elements

Zinc

Zinc is a component of numerous (approximately 100) enzymes associated with carbohydrate and energy metabolism, protein synthesis and degradation, and nucleic acid synthesis

It plays a role in cellular transport and protection from oxidative damage, as well as immune function, cell division and growth. Spermatogenesis is also zinc-dependent. Zinc plays a role in maintaining exocrine and endocrine pancreatic function. Its effects are most obviously seen in the maintenance of skin integrity and in wound healing.

Zinc from shares transport mechanisms with copper and iron the gut

On absorption, zinc is bound to metallothioneins, a family of cysteine-rich proteins, which also bind other divalent metal ions such as copper. Synthesis of metallothioneins is dependent on the amount of trace metals present in the diet. Increased synthesis is part of the metabolic response to trauma and results in a reduction of serum zinc concentration.

Zinc deficiency is common

Increased losses of zinc occur in patients with major burns and in those with renal damage. Zinc loss in renal disease is due to its association with plasma albumin, and it accompanies urinary protein loss. Substantial amounts of zinc may also be lost during dialysis. During intravenous feeding, failure to replace it may produce symptomatic deficiency.

The sources of zinc include **oysters** (highest content), **red meat, poultry, beans and nuts**. Note that phytates bind zinc. Zinc is not stored in the body.

Zinc deficiency might be a result of malabsorption associated with gastrointestinal GI surgery, short bowel syndrome, Crohn's disease, and ulcerative colitis, and may occur in liver and kidney disease. Chronic illnesses such as diabetes, malignancy and chronic diarrhea also lead to deficiency.

Pregnant women and alcoholics are prone to deficiency.

Zinc deficiency affects growth, skin integrity and wound healing

In children zinc deficiency is characterized by growth retardation, skin lesions, and impairment of immune function and sexual development. A specific inherited defect in the absorption of zinc from the gut was identified in the 1970s; it was termed **acrodermatitis enteropathica** and presented with severe skin lesions, diarrhea and loss of hair (alopecia). Zinc deficiency also leads to impairment in taste and smell and to **delayed wound healing**.

Zinc is probably the least toxic of the trace metals but increased oral intake interferes with copper absorption, and may lead to copper deficiency and anemia.

Zinc supplements are used in the treatment of diarrhea in children

Zinc supplementation was shown to reduce the severity and duration of diarrhea in children in developing countries, and prevent further episodes of diarrhea. Therefore, zinc supplements are now recommended by WHO/UNICEF along with the oral rehydration treatment.

Measurement of serum zinc concentration is the usual method of assessing zinc status. However, many conditions and environmental factors affect its concentration in plasma, including inflammation, stress, cancer, smoking, steroid administration and hemolysis.

Clinical box A man treated with parenteral nutrition who developed a generalized rash: zinc deficiency

A 34-year-old man who required total intravenous feeding had been receiving the same prescription for some 4 months, with no assessment of his trace metal status. During this time, he continued to have major gastrointestinal losses and had intermittent pyrexia. He developed a rash across his face, head, and neck, with accompanying hair loss. He was clearly zinc deficient, with serum zinc concentration less than 1 µmol/L (range: 9–20 µmol/L; 60–

130 µg/dL).

Comment.

Patients with major catabolic illness and increased gastrointestinal losses have markedly increased zinc requirements. The zinc-depleted state this patient developed could aggravate his illness by preventing healing of his gastrointestinal lesions and by making him more susceptible to infection due to defects in his immune competence. Patients receiving intravenous feeding need to have their micronutrient status checked regularly.

Copper

Copper scavenges superoxide and other reactive oxygen species

Copper is associated with oxygenase enzymes including cytochrome oxidase and superoxide dismutase (the latter also requires zinc). One of the main roles of copper, especially in superoxide dismutase but also in association with the plasma copper-carrying protein ceruloplasmin, is the scavenging of superoxide and other reactive oxygen species. Copper is also required for the crosslinking of collagen, being an essential component of lysyl oxidase.

Pathways of copper metabolism are shared with other metals

Absorption of copper from the gut is, similarly to zinc, associated with metallothionein. Copper availability in the diet is less affected by dietary constituents than zinc, although high fiber intake reduces availability by complexing copper. In plasma, the absorbed copper is bound to albumin. Copper–albumin complex is quickly taken up by the liver. Within the hepatocyte, copper is associated with intracellular metallothioneins which are also capable of binding zinc and cadmium. Copper is transported within the hepatocyte to sites of protein synthesis by a chaperone protein and it is incorporated into apoceruloplasmin. The incorporation is catalyzed by an

ATPase called ATP7B. Ceruloplasmin is then released into circulation. The only mechanism of copper excretion is elimination in bile (Fig. 11.11).

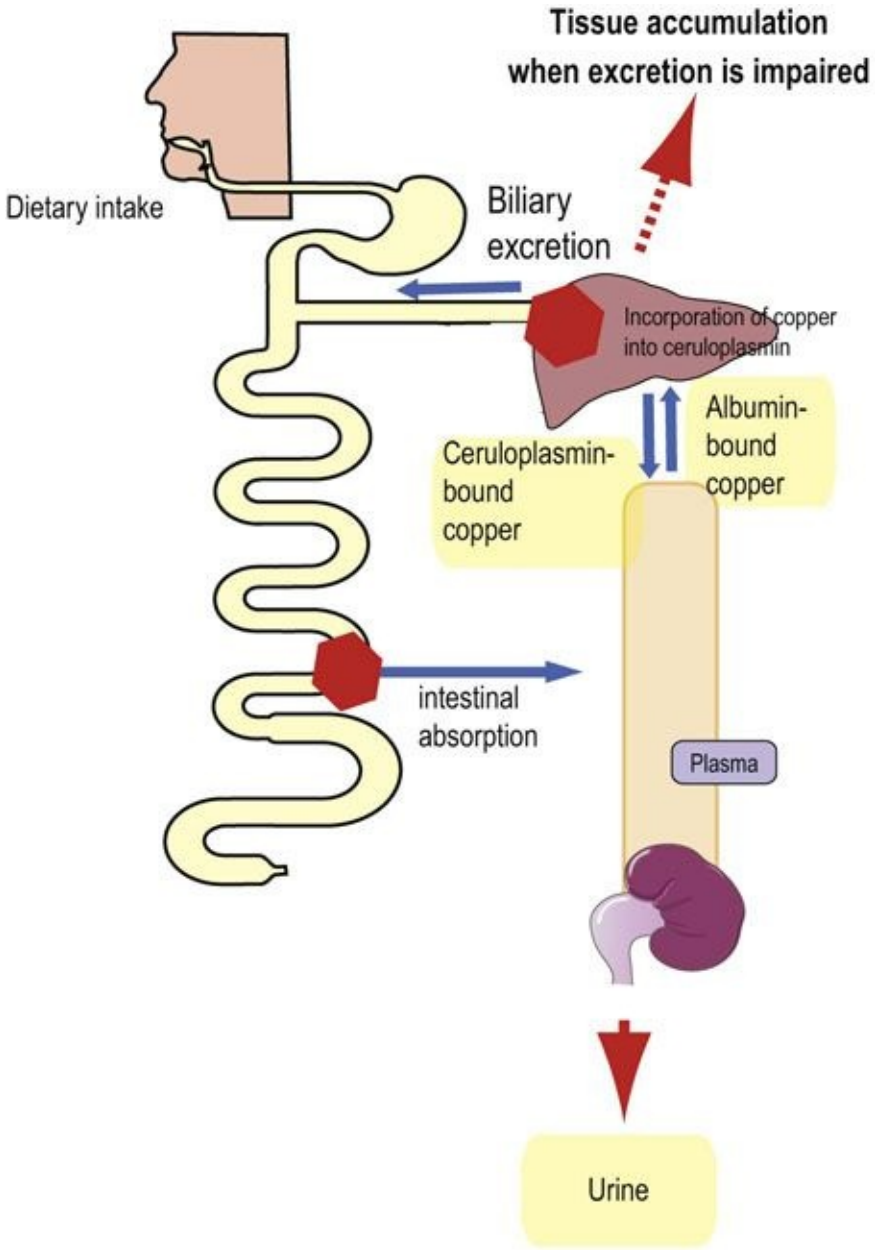


FIG 11.11 Copper metabolism. Reproduced from Dominiczak MH. Medical Biochemistry Flash Cards. London: Elsevier, 2012, Card 42.

Rare copper deficiency leads to an anemia; skin and hair

may also be affected

Rare copper deficiency is most likely to occur from reduced intake, or excess loss, *e.g.* during renal dialysis. Deficiency manifests itself as a **microcytic hypochromic anemia** (small pale erythrocytes) resistant to iron therapy. There is also a reduction in the number of leukocytes in the blood (neutropenia) and degeneration of vascular tissue with bleeding, due to defects in the synthesis of elastin and collagen. In severe deficiency, skin depigmentation and alteration in hair structure also occur. A very rare *Menkes' syndrome* results from copper depletion caused by a deficiency of the intestinal ATP7B ATPase.

Copper excess causes liver cirrhosis

When taken orally, copper is generally nontoxic. However, in large doses, it accumulates in tissues. Chronic excessive intake results in liver cirrhosis. Acute toxicity is manifested by marked hemolysis and damage to both liver and brain cells. The latter is seen in the autosomal dominant inherited metabolic defect, **Wilson's disease**, where the liver's capacity to synthesize ceruloplasmin is compromised. The cause is mutations in the gene coding for the ATP7B ATPase. This results in a reduced incorporation of copper into ceruloplasmin, and in its cellular accumulation. Excess of apoceruloplasmin is degraded. Copper accumulates in tissues such as the brain and cornea. Patients present with neurologic symptoms or liver cirrhosis and have typical **Kaiser–Fleischer rings** in the cornea. There is also a low concentration of ceruloplasmin and high urinary copper excretion (see box on p. 88).

Selenium

Selenium is present in all cells as amino acids selenomethionine and selenocysteine

Selenium is a component of selenoproteins, which contain the amino acid selenocysteine. The antioxidant enzyme glutathione peroxidase is a selenoprotein, as are the iodothyronine deiodinases, enzymes that produces triiodothyronine (T_3) and reverse T_3 (rT_3). Thioredoxin reductases that participate in cell proliferation apoptosis and DNA synthesis also contain selenocysteine.

Selenium affects functions of the immune system, including stimulation of

differentiation of T cells and proliferation of activated T lymphocytes, as well as increase in natural killer cell activity. It also plays a role in spermatogenesis.

Selenium is absorbed in the small intestine. It remains protein bound in circulation, and is excreted in urine. Selenoprotein P possesses 10 selenocysteine residues and it transports selenium in plasma from the liver to, primarily, the brain, testis and kidney. In the brain it binds to a membrane receptor apoER2, which belongs to the family of lipoprotein receptors.

Selenium is present in diet as selenomethionine and selenocysteine. Brazil nuts are its richest source. Its dietary sources also include organ meats, fish (tuna) and shellfish, and cereals. Its content in plant-derived food depends on the content of the soil.

Selenium status may influence the risk of many chronic conditions

Low selenium is associated with decline in immune function and with cognitive problems. Low concentration has been observed in individuals with epileptic seizures and also in preeclampsia. Deficiency of selenium can also develop during **total parenteral nutrition**. There is a rare selenium-responsive cardiomyopathy (**Keshan disease**), which is endemic in China in areas of very low selenium intake. Selenium deficiency may result in chronic muscle pain, abnormal nail beds, and cardiomyopathy. Excess of selenium, on the other hand, leads to liver cirrhosis, splenomegaly, gastrointestinal bleeding and depression.

Increased intake of selenium might be required during lactation. Several studies indicate beneficial effect of selenium on the risk of lung, prostate, bladder and other cancers. Single nucleotide polymorphisms in selenoprotein genes were shown to be important in determining risk of conditions such as various cancers, preeclampsia and possibly cardiovascular disease.

Currently it seems that while people with low selenium concentration may benefit from supplementation, supplementing it to those with normal or high values can actually be harmful.

Other metals

Numerous other trace metals are required for normal biological function: for example, manganese, molybdenum, vanadium, nickel and cadmium. Some, similar to zinc and copper, form prosthetic groups of enzymes. These include **molybdenum** (xanthine oxidase) and **manganese** (superoxide dismutase and

pyruvate carboxylase) **Chromium** has been associated with glucose tolerance.

Many of these metals were previously thought to be toxic; indeed, their environmental excess does result in toxicity such as the renal toxicity observed in shipyard workers exposed to **cadmium** over long periods of time. As techniques for separation and analysis develop, other metals and other functions of known essential minerals will become known. This will lead to a better understanding of the epidemiology of certain diseases that may have, at least in part, an environmental etiology.

Summary

■ Vitamins function mostly as cofactors to enzymes.
■ Fat-soluble vitamins can be stored in the adipose tissue, but there usually is only a short-term supply of the water-soluble vitamins.
■ Dietary micronutrient deficiencies are most likely to occur in susceptible groups with increased demand, or in people unable to maintain sufficient intake. Children, pregnant women, the elderly, alcoholics and low-income groups are particularly vulnerable.
■ Gastrointestinal disease and gastrointestinal surgery are potential causes of micronutrient deficiencies.
■ Vitamin and trace metal supplements are particularly important in patients who remain on artificial diets and on parenteral nutrition.
■ While there are controversies regarding some vitamin supplementation, the intake of fruit and vegetables as sources of micronutrients is unequivocally recommended.

Active learning

1. Compare and contrast the deficiencies of vitamin B_{12} and folic acid.
2. When may an increased intake of a nutrient or energy precipitate vitamin deficiencies?
3. Is vitamin A supplementation safe?
4. Describe the clinical importance of copper.
5. Which vitamins play a role in the development of hyperhomocysteinemia?

Further reading

Asplund, K. Antioxidant vitamins in the prevention of cardiovascular disease: a systematic review. *J Int Med*. 2002; 251:372–392.

El-Youssef, M. Wilson disease. *Mayo Clin Proc*. 2003; 78:1126–1136.

Fairfield, KM, Fletcher, RH. Vitamins for chronic disease prevention in adults: scientific review. *JAMA*. 2002; 287:3116–3126.

Fletcher, RH, Fairfield, KM. Vitamins for chronic disease prevention in adults: clinical applications. *JAMA*.

2002; 287:3127–3129.

Fisher Walker, CL, Black, RE. Zinc treatment for serious infections in young infants. *Lancet*. 2012; 379:2031–2033.

Lucock, M. Is folic acid the ultimate functional food component for disease prevention? *BMJ*. 2004; 328:211–214.

Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy. *Dietary reference values for food energy and nutrients for the United Kingdom*. London: TSO; 2003.

Rayman, M. Selenium and human health. *Lancet*. 2012; 379:1256–1268.

Schneider, BD, Leibold, EA. Regulation of mammalian iron homeostasis. *Curr Opin Clin Nutr Metab Care*. 2000; 3:267–273.

Websites and downloads

Bender, DA. *Introduction to nutrition and metabolism*, 4th ed. Boca Raton, FL: CRC Press; 2008.

FAO Corporate Document Repository. Human vitamin and mineral requirements. http://www.fao.org/docrep/004/Y2809E/y2809e01.htm#TopOfPage.

Fact sheets concerning vitamins and trace elements are available from the Office of Dietary Supplements, National Institutes of Health, USA. http://ods.od.nih.gov/factsheets/list-all/.

UNICEF. Nutrition in emergencies. www.unicef.org/nutrition/training/2.1/5.html.

CHAPTER 12

Anaerobic Metabolism of Glucose in the Red Blood Cell

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

■ Outline the sequence of reactions in anaerobic glycolysis, the central pathway of carbohydrate metabolism in all cells.

■ Summarize the energetics of anaerobic glycolysis, including the reactions involved in the utilization and formation of ATP, and the net yield of ATP during glycolysis.

■ Identify the primary site of allosteric regulation of glycolysis and the mechanism of regulation of this enzyme.

■ Identify steps in glycolysis that illustrate the use of coupled reactions to drive thermodynamically unfavored processes, including substrate-level phosphorylation.

■ Describe the major roles of the pentose phosphate pathway in erythrocytes and nucleated cells.

■ Describe the role of anaerobic glycolysis in development of dental caries.

■ Explain why glycolysis is essential for normal red cell functions, including consequences of deficiencies in glycolytic enzymes and the role of glycolysis in adaptation to high altitude.

■ Explain the origin of drug-induced hemolytic anemia in persons with G6PD deficiency.

Introduction

Glycolysis is the central pathway of glucose metabolism in all cells

Glucose is the major carbohydrate on Earth, the backbone and monomer unit of cellulose and starch. It is also the only fuel that is used by all cells in our body. All of these cells, even the microbes in our intestines, begin the metabolism of glucose by a pathway termed glycolysis, *i.e.* carbohydrate (glyco) splitting (lysis). Glycolysis is catalyzed by soluble cytosolic enzymes and is the ubiquitous, central metabolic pathway for glucose metabolism. The erythrocyte, commonly known as the red blood cell (RBC), is unique among all cells in the body – it uses glucose and glycolysis as its sole source of energy. Thus, the RBC is a useful model for an introduction to glycolysis.

Pyruvate, a three-carbon carboxylic acid, is the end product of glycolysis; 2 moles of pyruvate are formed per mole of glucose

In cells with mitochondria and oxidative metabolism, pyruvate is converted completely into CO_2 and H_2O – glycolysis in this setting is termed **aerobic glycolysis**. In RBCs, which lack mitochondria and oxidative metabolism, pyruvate is reduced to lactic acid, a three-carbon hydroxyacid, the product of **anaerobic glycolysis**. Each mole of glucose yields 2 moles of lactate, which are then excreted into blood. Two molecules of lactic acid contain exactly the same number of carbons, hydrogens, and oxygens as one molecule of glucose (Fig. 12.1); however, there is sufficient free energy available from the cleavage and rearrangement of the glucose molecule to produce 2 moles of ATP per mole of glucose converted into lactate. The RBC uses most of this ATP to maintain electrochemical and ion gradients across its plasma membrane.

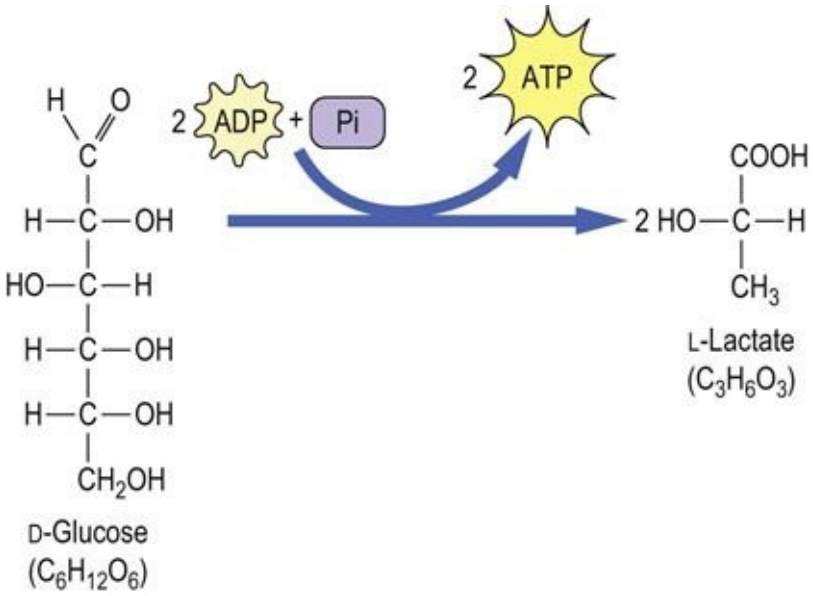


FIG. 12.1 Conversion of glucose to lactate during anaerobic glycolysis.
One mole of glucose is converted to 2 moles of lactate during anaerobic glycolysis. No oxygen is consumed, nor is CO_2 produced in this pathway. There is a net yield of 2 mole of ATP per mole of glucose converted to lactate.

In the RBC, 10–20% of the glycolytic intermediate, 1,3-bisphosphoglycerate, is diverted to the synthesis of 2,3-bisphosphoglycerate (2,3-BPG), an allosteric regulator of the O_2 affinity of Hb (Chapter 5). The **pentose phosphate pathway**, a shunt from glycolysis, accounts for about 10% of glucose metabolism in the red cell. In the red cell, this pathway has a special role in protection against oxidative stress, while in nucleated cells it also serves as a source of NADPH for biosynthetic reactions and pentoses for nucleic acid synthesis.

The erythrocyte

The erythocyte or red blood cell relies exclusively on blood glucose as a metabolic fuel

The erythrocyte, or red blood cell (RBC), represents 40–45% of blood volume and over 90% of the formed elements (erythrocytes, leukocytes, and platelets) in blood. The RBC is, both structurally and metabolically, the simplest cell in the body – the end product of the maturation of bone marrow reticulocytes. During its maturation, the RBC loses all its subcellular organelles. Without nuclei, it lacks the ability to synthesize DNA or RNA. Without ribosomes or an endoplasmic reticulum, it cannot synthesize or secrete protein. Because it cannot oxidize fats, a process requiring mitochondrial activity, the RBC relies exclusively on blood glucose as a fuel. Metabolism of glucose in the RBC is entirely anaerobic, consistent with the primary role of the RBC in oxygen transport and delivery, rather than its utilization.

Advanced concept box Glucose utilization in the red cell

In a 70-kg person, there are about 5 L of blood and a little over 2 kg (2 L) of RBCs. These cells constitute about 3% of total body mass and consume about 20 g (0.1 mole) of glucose per day, representing about 10% of total body glucose metabolism. The RBC has the highest specific rate of glucose utilization of any cell in the body, approximately 10 g of glucose/kg of tissue/day, compared with ~2.5 g of glucose/kg of tissue/day for the whole body.

In the RBC, about 90% of glucose is metabolized via glycolysis, yielding lactate, which is excreted into blood. Despite its high rate of glucose consumption, the RBC has one of the lowest rates of ATP synthesis of any cell in the body, ~0.1 mole of ATP/kg tissue/day, reflecting the fact that anaerobic glycolysis recovers only a fraction of the energy available from complete combustion of glucose to CO_2 and H_2O.

Glycolysis

Overview

Pyruvate is the endproduct of anaerobic glycolysis

Glucose enters the RBC by facilitated diffusion, via the insulin-independent glucose transporter, GLUT-1. Glycolysis then proceeds through a series of phosphorylated intermediates, starting with the synthesis of glucose-6-phosphate (Glc-6-P). During this process, which involves 10 enzymatically catalyzed steps, two molecules of ATP are expended (**investment** stage) to build up a nearly symmetric intermediate, fructose-1,6-bisphosphate (Fru-1,6-BP), which is then cleaved (**splitting** stage) to two three-carbon triose phosphates. These are eventually converted into lactate, with production of ATP, during the **yield** stage of glycolysis. The yield stage includes both redox and phosphorylation reactions, leading to formation of four molecules of ATP during the conversion of the two triose phosphates into lactate. The outcome is a net 2 moles of ATP per mole of glucose converted into lactate.

Glycolysis is a relatively inefficient pathway for extracting energy from glucose: the yield of 2 moles of ATP per mole of glucose is only about 5% of the 30–32 ATP that are available by complete oxidation of glucose to CO_2 and H_2O by mitochondria in other tissues (Chapter 14).

One might ask why a 10-step pathway is required to convert glucose to lactate; couldn't it have been done in fewer steps or by cleavage of one carbon at a time? The answer, from a metabolic point of view, is that glycolysis is not an isolated pathway; most glycolytic intermediates serve as branch points to other metabolic pathways. In this way, the metabolism of glucose intersects with the metabolism of fats, proteins and nucleic acids, as well as other pathways of carbohydrate metabolism. Some of these metabolic interactions are shown in Figure 12.2.

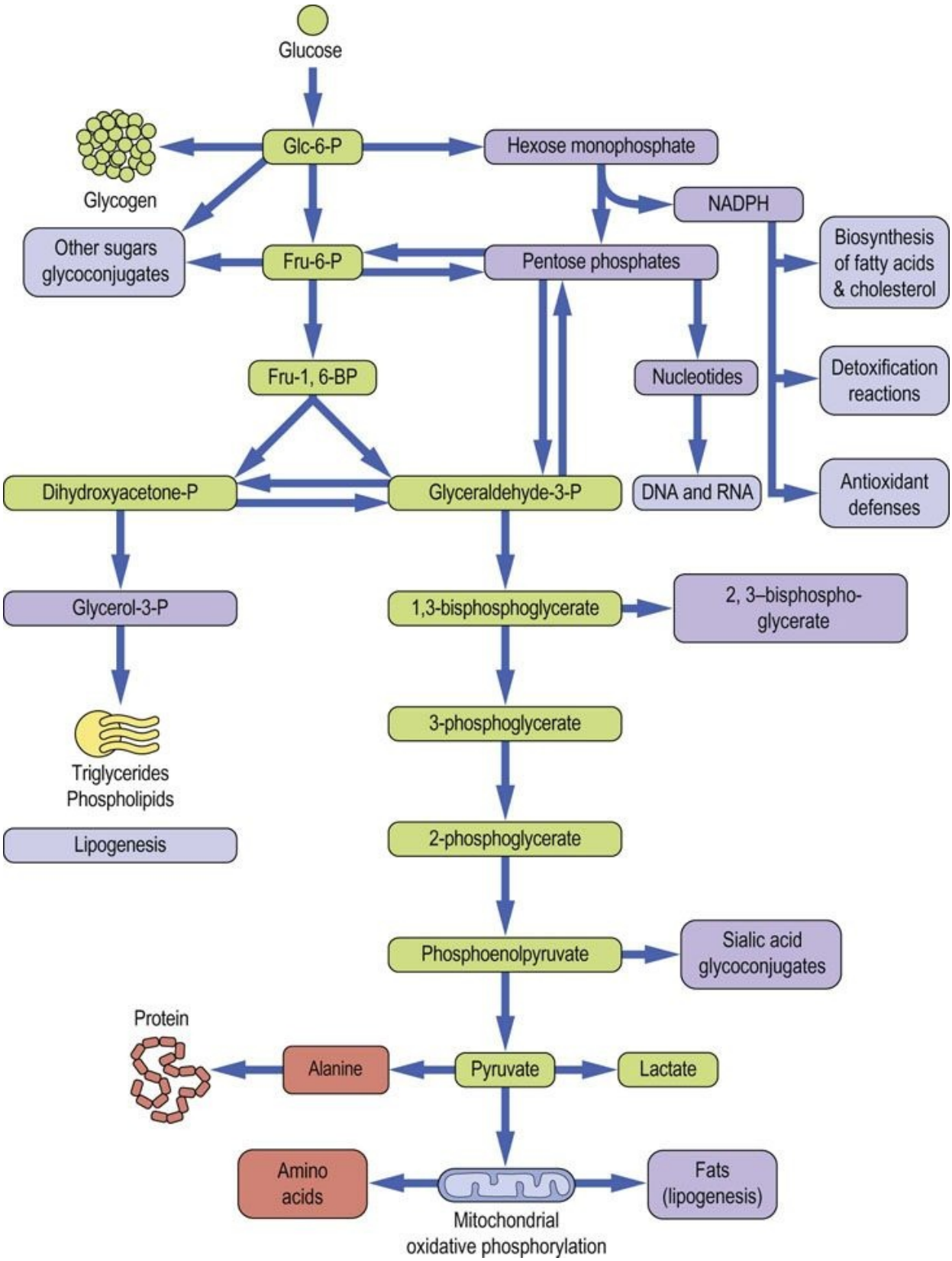


FIG. 12.2 Interactions between glycolysis and other metabolic pathways.

The green-colored boxes indicate intermediates involved in the pathway of glycolysis. Other boxes illustrate some of the metabolic interactions between glycolysis and other metabolic pathways in the cell. Not all of these pathways are active in the red cell, which has limited biosynthetic capacity and lacks mitochondria. Glc-6-P, glucose-6-phosphate; Fru-6-P, fructose-6-phosphate; Fru-1,6-BP, fructose-1,6-bisphosphate.

The investment stage of glycolysis

2 ATP are invested to prime the metabolism of glucose by glycolysis

Glucose-6-phosphate

Glucose is taken up into the red cell via the facilitated transporter GLUT-1 (Chapter 8); this protein accounts for about 5% of total red cell membrane protein, so that glucose transport is not rate limiting for glycolysis. Thus, the steady state concentration of glucose in the RBC is only ~20% lower than that in plasma. The first step in the commitment of glucose to glycolysis is the phosphorylation of glucose to Glc-6-P, catalyzed by the enzyme hexokinase (Fig. 12.3, top). The formation of Glc-6-P from free glucose and inorganic phosphate is energetically unfavorable, so that a molecule of ATP must be expended or *invested* in the phosphorylation reaction; the hydrolysis of ATP is coupled to the synthesis of Glc-6-P. Glc-6-P is trapped in the RBC, along with other phosphorylated intermediates in glycolysis, because there are no transport systems for sugar phosphates in the plasma membranes of mammalian cells.

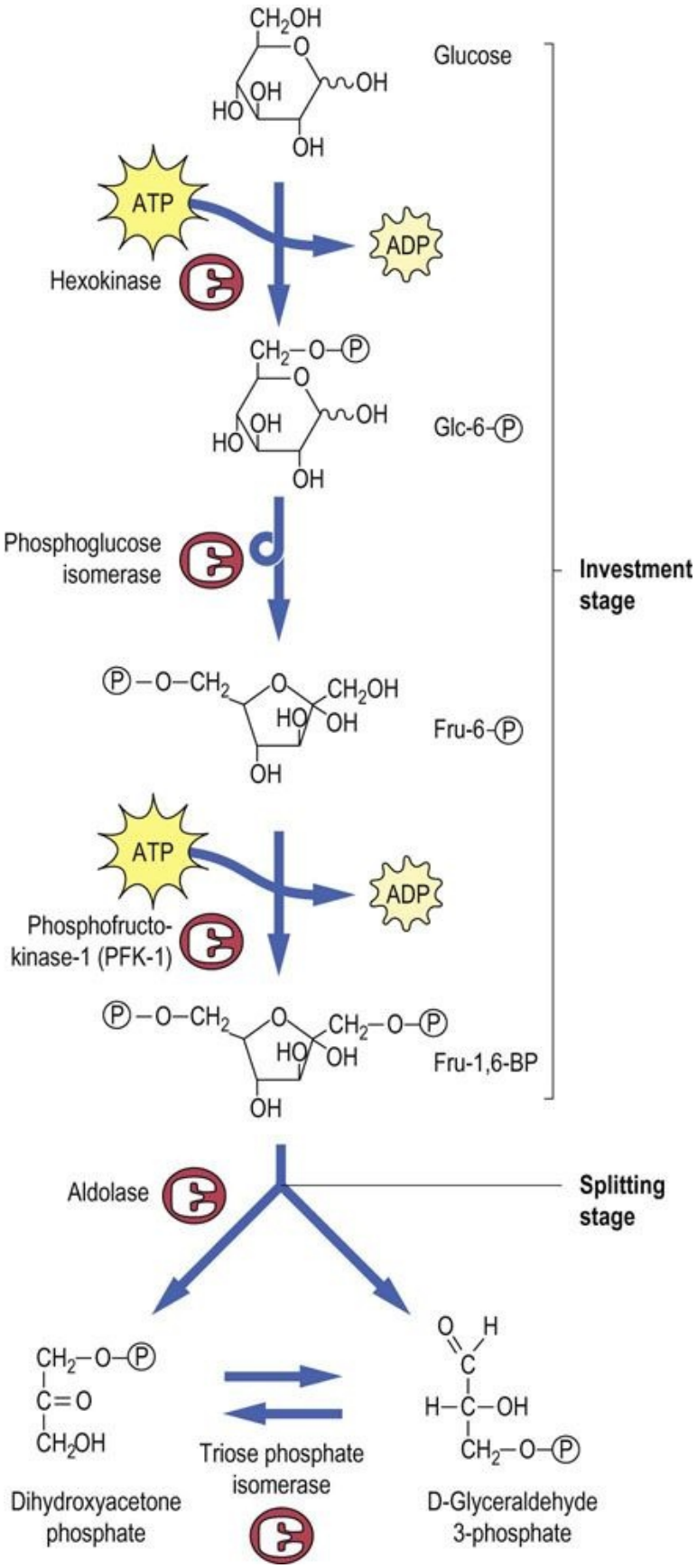

Glucose
ATP
ADP
Hexokinase
Glc-6-Ⓟ
Phosphoglucose isomerase
Investment stage
Fru-6-Ⓟ
ATP
ADP
Phosphofructo-kinase-1 (PFK-1)
Fru-1,6-BP
Aldolase
Splitting stage
Dihydroxyacetone phosphate
Triose phosphate isomerase
D-Glyceraldehyde 3-phosphate

FIG. 12.3 The investment and splitting stages of glycolysis.
Note the consumption of ATP at the hexokinase and phosphofructokinase-1 reactions.

Fructose-6-phosphate

The second step in glycolysis is the conversion of Glc-6-P into Fru-6-P by phosphoglucose isomerase (Fig. 12.3, middle). Isomerases catalyze freely reversible equilibrium reactions, in this case an aldose–ketose interconversion. A second molecule of ATP is invested to phosphorylate Fru-6-P at the C-1 position; the reaction is catalyzed by phosphofructokinase-1 (PFK-1). The product, fructose 1,6-bisphosphate (Fru-1,6-BP), is a pseudosymmetric intermediate, with a phosphate ester on each end of the molecule. Like hexokinase, PFK-1 requires ATP as a substrate and catalyzes an essentially irreversible reaction. Both hexokinase and PFK-1 are important regulatory enzymes in glycolysis, but PFK-1 is the critical, commitment step. This reaction directs glucose to glycolysis, the only pathway for metabolism of Fru-1,6-BP.

The splitting stage of glycolysis

Fructose-1,6-BP is cleaved in the middle by a reverse aldol (aldolase) reaction

The aldolase reaction (Fig. 12.3, bottom) is a freely reversible equilibrium reaction, yielding two triose phosphates, dihydroxyacetone phosphate and glyceraldehyde-3-phosphate, from the top and bottom halves of the Fru-1,6-BP molecule, respectively. Only the glyceraldehyde-3-phosphate continues through the yield stage of glycolysis, but triose phosphate isomerase catalyzes the interconversion of dihydroxyacetone phosphate and glyceraldehyde-3-phosphate, so that both halves of the glucose molecule are eventually metabolized to lactate.

The yield stage of glycolysis – synthesis of ATP by substrate-level phosphorylation

The yield stage of glycolysis produces 4 moles of ATP, yielding a net of 2 moles of ATP per mole of glucose converted into lactate

The synthesis of ATP during glycolysis is accomplished by kinases that catalyze **substrate-level phosphorylation**, a process in which a high-energy phosphate compound transfers its phosphate to ADP, yielding ATP.

$$\text{Substrate-level phosphorylation}: \quad X \sim P + ADP \rightarrow X + ATP$$

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)

GAPDH catalyzes a redox reaction, forming a high energy acyl phosphate compound

To set the stage for substrate-level phosphorylation, the aldehyde group of glyceraldehyde-3-phosphate is oxidized to a carboxylic acid and the energy available from the oxidation reaction is used, in part, to trap a phosphate from the cytoplasmic pool as an acyl phosphate. This reaction is catalyzed by glyceraldehyde-3-phosphate dehydrogenase (GAPDH), yielding the high-energy compound (X~P), 1,3-bisphosphoglycerate (1,3-BPG). The coenzyme NAD^+ is simultaneously reduced to NADH (Figs 12.4, 12.5).

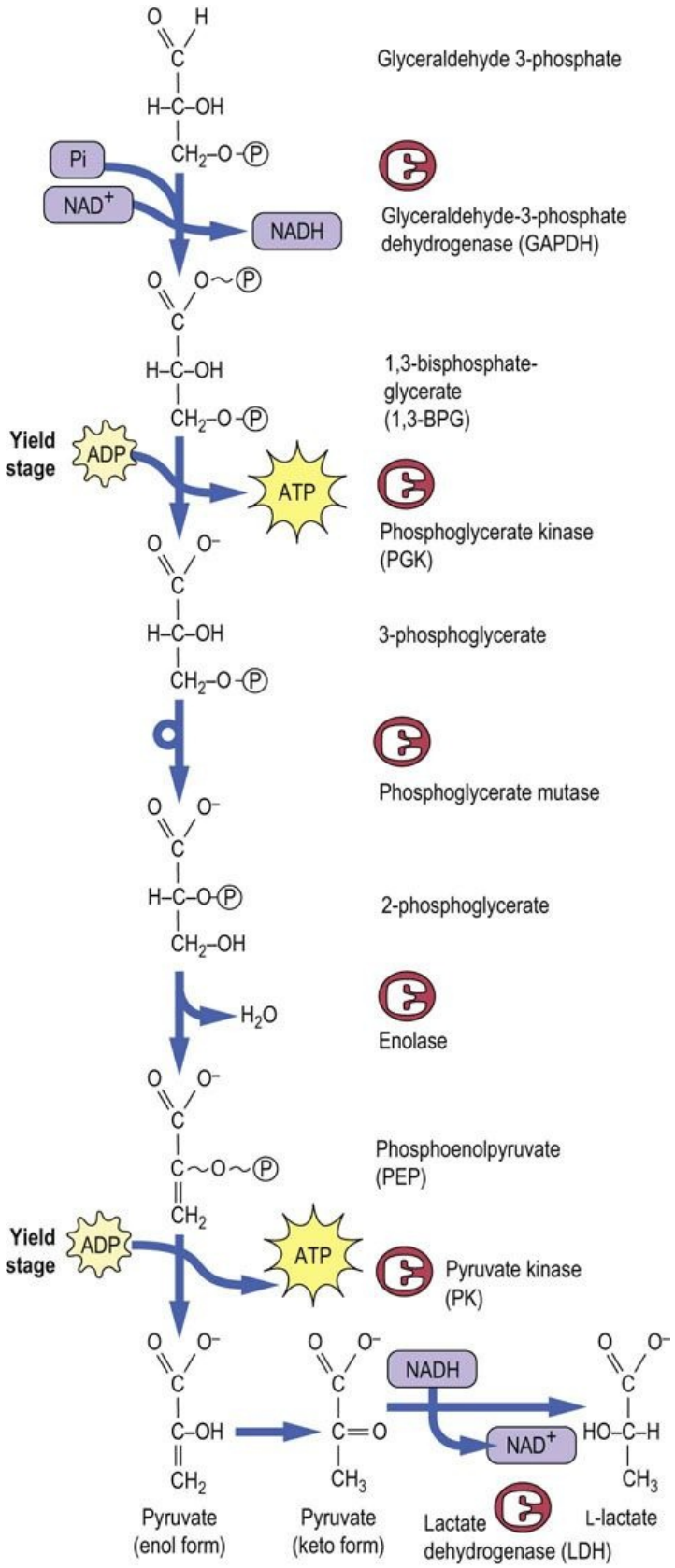

Glyceraldehyde 3-phosphate
Pi
NAD+
NADH
Glyceraldehyde-3-phosphate dehydrogenase (GAPDH)
1,3-bisphosphate-glycerate (1,3-BPG)
Yield stage
ADP
ATP
Phosphoglycerate kinase (PGK)
3-phosphoglycerate
Phosphoglycerate mutase
2-phosphoglycerate
H2O
Enolase
Phosphoenolpyruvate (PEP)
Yield stage
ADP
ATP
Pyruvate kinase (PK)
NADH
NAD+
Pyruvate (enol form)
Pyruvate (keto form)
Lactate dehydrogenase (LDH)
L-lactate

FIG. 12.4 The yield stage of glycolysis.
Substrate-level phosphorylation reactions catalyzed by phosphoglycerate kinase and pyruvate kinase produce ATP, using the high-energy compounds, 1,3-bisphosphoglycerate and phosphoenolpyruvate, respectively. Note that NADH produced during the glyceraldehyde-3-phosphate dehydrogenase reaction is recycled back to NAD^+ during the lactate dehydrogenase reaction, permitting continued glycolysis in the presence of only catalytic amounts of NAD^+.

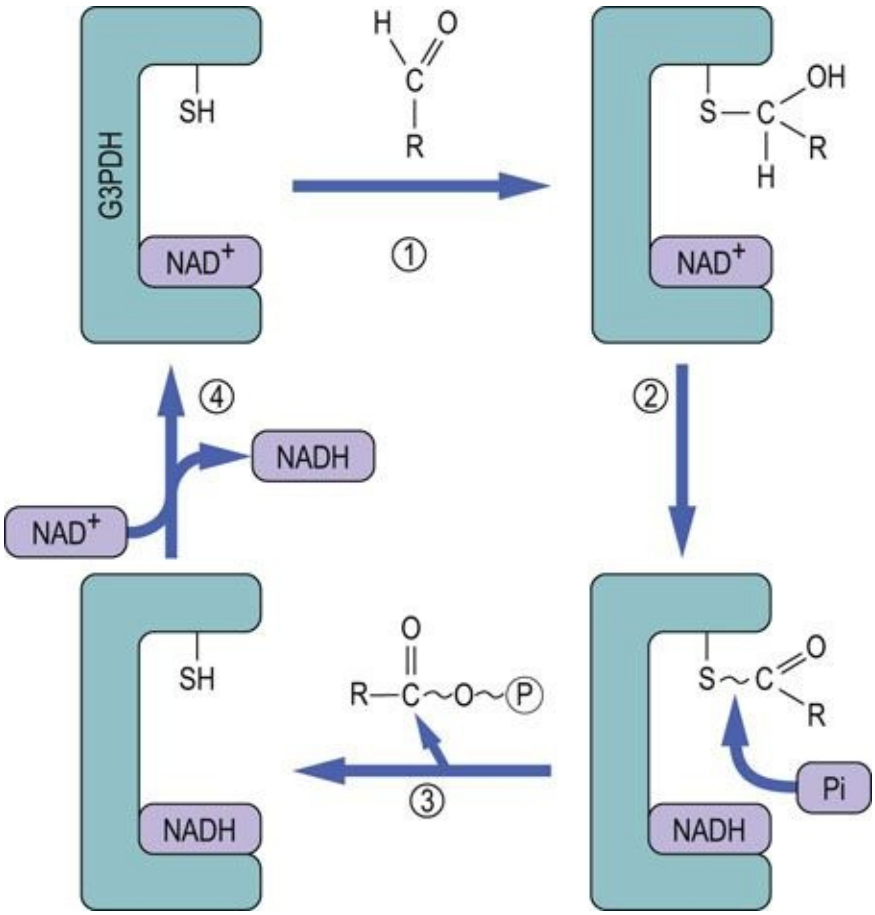


FIG. 12.5 Mechanism of the glyceraldehyde-3-phosphate dehydrogenase (GAPDH) reaction.
In Step 1, glyceraldehyde-3-P (RCHO) reacts with the active-site sulfhydryl group of GAPDH to form a thiohemiacetal adduct. In Step 2, the thiohemiacetal is oxidized to a thioester by NAD^+, which is bound in the active site of the enzyme and is reduced to NADH. In Step 3, phosphate enters the active site and, in a phosphorylase reaction, cleaves the carbon–sulfur bond, displacing the 3-phosphoglycerate group, producing 1,3-bisphosphoglycerate and regenerating the sulfhydryl group. In Step 4, the enzyme exchanges NADH for NAD^+, completing the catalytic cycle.

The GAPDH reaction provides an interesting illustration of the role of enzyme-bound intermediates in the formation of high-energy phosphates. How does the oxidation of an aldehyde and the reduction of NAD^+ lead to the formation of an acyl phosphate bond in 1,3-BPG? How does the phosphate enter the picture, and become activated to a high-energy state? The inhibition of GAPDH by thiol reagents such as iodoacetamide, *p*-chloromecuribenzoate and *N*-ethylmaleimide pointed to involvement of an active-site sulfhydryl residue. The mechanism of action of this enzyme is described in Figure 12.5.

Substrate-level phosphorylation

Substrate-level phosphorylation produces ATP from another high energy phosphate compound

Phosphoglycerate kinase (PGK) catalyzes transfer of the phosphate group from the high-energy acyl phosphate of 1,3-BPG to ADP, forming ATP. This substrate-level phosphorylation reaction yields the first ATP produced in glycolysis. The remaining phosphate group in 3-phosphoglycerate is an ester phosphate and does not have enough energy to phosphorylate ADP, so a series of isomerization and dehydration reactions is enlisted to convert the ester phosphate into a high-energy enol phosphate. The first step is to shift the phosphate to C-2 of glycerate, converting 3-phosphoglycerate into 2-phosphoglycerate, catalyzed by the enzyme phosphoglycerate mutase (see Fig. 12.4). **Mutases catalyze the transfer of functional groups within a molecule.** Phosphoglycerate mutase has an active-site histidine residue, and a phospho-histidine adduct is formed as an enzyme-bound intermediate during the phosphate transfer reaction.

2-Phosphoglycerate then undergoes a dehydration reaction, catalyzed by enolase, a hydratase, to yield the high-energy phosphate compound, phosphoenolpyruvate (PEP). PEP is used by pyruvate kinase to phosphorylate ADP, yielding pyruvate and the second ATP, again by substrate-level phosphorylation. It seems strange that the high-energy phosphate bond in PEP can be formed from the low-energy phosphate compound 2-phosphoglycerate by a simple sequence of isomerization and dehydration reactions. However, the thermodynamic driving force for these reactions is probably derived from charge–charge repulsion between the phosphate and carboxylate groups of 2-phosphoglycerate and the isomerization of enolpyruvate to pyruvate following

the phosphorylation reaction.

Phosphoglycerate kinase and pyruvate kinase catalyze substrate-level phosphorylation reactions

The ATP-generating reactions of glycolysis produce 2 moles of ATP per mole of triose phosphate, or a total of 4 moles of ATP per mole of Fru-1,6-BP. After adjustment for the ATP invested in the hexokinase and PFK-1 reactions, the net energy yield is 2 moles of ATP per mole of glucose converted into pyruvate.

Advanced concept box Inhibition of substrate-level phosphorylation by arsenate

Arsenic is just below phosphorus in the periodic table of the elements, and it might be expected to share some of the properties and reactivity of phosphate. In fact, arsenate has pK_a values similar to those of phosphate and can actually be used by GAPDH, producing 1-arsenato-3-phosphoglycerate. However, the acyl–arsenate bond is unstable and hydrolyzes rapidly, and ATP is not generated by substrate-level phosphorylation. While arsenate does not inhibit any of the enzymes of glycolysis, it dissipates the redox energy available from the GAPDH reaction and prevents the formation of ATP by substrate-level phosphorylation at the PGK reaction. In effect, arsenate *uncouples* the GAPDH and PGK reactions. Note that arsenic and arsenite are also toxic, but have a different mechanism of action: they react with thiol groups in sulfhydryl enzymes, such as GAPDH (Fig. 12.5), irreversibly inhibiting their activity.

Clinical TEST box Inhibition of enolase by fluoride

Measurements of blood glucose concentration are used for the

diagnosis and management of diabetes. Frequently these measurements are made in the clinical laboratory more than 1 h after the collection of the blood sample. Because RBCs can metabolize glucose to lactate, even in a sealed, anoxic container, glucose in blood will be consumed and lactate will be produced, which will lead to acidification of the blood sample. These reactions proceed in RBCs, even at room temperature, so that both blood glucose concentration and pH will decrease during standing – possibly leading to false diagnosis of hypoglycemia and acidemia.

Anaerobic metabolism of glucose can be prevented by adding an inhibitor of glycolysis to the blood collection tube. Sulfhydryl reagents would work – they are inhibitors of GAPDH; however, most blood samples are collected with a small amount of a much cheaper reagent, sodium fluoride, in the sample-collection vial. Fluoride is a strong competitive inhibitor of enolase, blocking glycolysis and lactate production in the RBC. It is an unusual competitive inhibitor, since fluoride bears little resemblance to 2-phosphoglycerate. In this case, fluoride forms a complex with phosphate and Mg^{2+} in the active site of the enzyme, blocking access of substrate.

Clinical box Pyruvate kinase deficiency

A child presented with jaundice and abdominal tenderness, which developed following a severe cold. Laboratory tests revealed a low hematocrit and hemoglobin concentration, normochromatic erythrocytes with normal morphology, and mild reticulocytosis. Serum bilirubin was increased.

Comment.

Pyruvate kinase deficiency is the most common of the hemolytic anemias that result from a deficiency in a glycolytic enzyme. It is

an autosomal recessive disorder that occurs with a frequency of 1/10,000 (~1% gene frequency) in the world population. It is second only to G6PDH deficiency (see below) as an enzymatic cause of hemolytic anemia. These diseases are diagnosed by measurement of erythrocyte levels of enzymes or metabolites, by demonstrating abnormalities in enzymatic activities, or by genetic analysis. Enzymatic defects in pyruvate kinase that have been characterized include thermal lability, increased K_m for PEP, and decreased activation by Fru-1,6-BP.

Pyruvate kinase deficiency varies significantly in severity, from a mild, compensated condition requiring little intervention to a severe disease requiring transfusions. The anemia results from inability to synthesize ATP, required for maintenance of RBC metabolism, ion gradients and cell shape. Interestingly, patients may tolerate the anemia quite well. Even with mild anemia, the accumulation of 2,3-bisphosphoglycerate in their RBCs decreases the oxygen affinity of hemoglobin, promoting oxygen delivery to muscle during exercise and even to the fetus during pregnancy.

Lactate dehydrogenase (LDH)

LDH regenerates NAD_+ consumed in the GAPDH reaction, producing lactate, the endproduct of anaerobic glycolysis

Two molecules of pyruvate have exactly the same number of carbons and oxygens as one molecule of glucose; however, there is a deficit of four hydrogens – each pyruvate has four hydrogens, a total of eight hydrogens for two pyruvates, compared with 12 in a molecule of glucose. The 'missing' four hydrogens remain in the form of the 2NADH and 2 H^+ formed in the GAPDH reaction. Since NAD^+ is present in only catalytic amounts in the cell and is an essential cofactor for glycolysis (and other reactions), there must be a mechanism for regeneration of NAD^+ if glycolysis is to continue.

The oxidation of NADH is accomplished under anaerobic conditions by lactate dehydrogenase (LDH), which catalyzes reduction of pyruvate to lactate by NADH + H^+ and regenerates NAD^+. In mammals, all cells have LDH, and

lactate is the end product of glycolysis under anaerobic conditions. Under aerobic conditions, mitochondria oxidize NADH to NAD^+ and convert pyruvate to CO_2 and H_2O, so that lactate is not formed. Despite their capacity for oxidative metabolism, however, some cells may at times ‘go glycolytic’, forming lactate, *e.g.* in muscle during oxygen debt and in phagocytes in pus or in poorly perfused tissues. Most of the lactate excreted into blood is retrieved by the liver for use as a substrate for gluconeogenesis (Chapter 13).

Fermentation

Fermentation is a general term for anaerobic metabolism of glucose, usually applied to monocellular organisms

Some anaerobic bacteria, such as lactobacilli, produce lactate, while others have alternative pathways for anaerobic oxidation of NADH formed during glycolysis. During fermentation in yeast, the pathway of glycolysis is identical with that in the RBC, except that pyruvate is converted into ethanol (Fig. 12.6). The pyruvate is first decarboxylated by pyruvate decarboxylase to acetaldehyde, releasing CO_2. The NADH produced in the GAPDH reaction is then reoxidized by alcohol dehydrogenase, regenerating NAD^+ and producing ethanol. Ethanol is a toxic compound and most yeast die when the ethanol concentration in their medium reaches about 12%, which is the approximate concentration of alcohol in natural wines. Alcoholic beverages are a rich source of energy; alcohol yields ~7 kcal/g (29 kJ/g) by aerobic metabolism (Table 9.1), intermediate between carbohydrates and lipids. As a food, alcoholic beverages are more stable to long-term storage, compared to the fruits and vegetables from which they are produced. Beer, wine, cider and mead also provide varying amounts of vitamins, minerals, phytochemicals and xenobiotics.

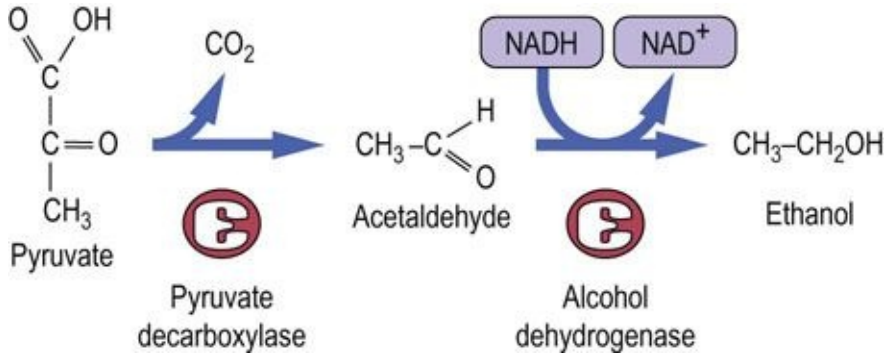


FIG. 12.6 Anaerobic glycolysis in yeast.
Formation of ethanol by anaerobic glycolysis during fermentation. Pyruvate is decarboxylated by pyruvate decarboxylase, yielding acetaldehyde and CO_2. Alcohol dehydrogenase uses NADH to reduce acetaldehyde to ethanol, regenerating NAD^+ for

glycolysis.

Other fermented food products, which are estimated to account for a third of all foods eaten by humans worldwide, include pickles, sauerkraut, buttermilk, yogurt, sausage, some fish and meats, bread, cheese, and various sauces and condiments – even coffee and chocolate. The acidic environment produced during fermentation limits spoilage and growth of pathogenic microorganisms. Fermentation also pre-digests and increases the digestibility of foods, sometimes improves its nutritional value by adding bacteria-derived vitamins, and invariably contributes to the flavor and aroma of foodstuffs.

There are as many as 1000 species of anaerobic bacteria in our intestines. These enterobacteria thrive in a symbiotic relationship with man. They assist significantly in the digestion and extraction of energy from foodstuffs, are a source of biotin and vitamin K, provide protection against infection by pathogens, and promote gastrointestinal peristalsis. The species distribution also changes in response to the carbohydrate, fat and protein content of our diet.

Clinical box Glycolysis and dental caries

Streptococcus mutans and *Lactobacillus* are anaerobic bacteria that colonize the oral cavity and contribute to the development of dental caries. These bacteria grow optimally on refined, fermentable carbohydrates in the diet, and excrete organic acids, such as lactate. They thrive in acidic, anaerobic microenvironments in fissures in the teeth and in gingival pockets. The organic acids gradually erode tooth enamel and dentin, and the chronic dissolution of the calcium phosphate (hydroxyapatite) matrix of the teeth sets the stage for cavity formation. Fluoride, provided either topically or in toothpaste, at levels too low to inhibit enolase, integrates into the tooth surface, forming fluoroapatite, which is more resistant to demineralization.

Regulation of glycolysis in erythrocytes

Glycolysis is regulated allosterically at three kinase reactions

Hexokinase

RBCs consume glucose at a fairly steady rate. They are not physically active like muscle, and do not require energy for transport of O_2 or CO_2. Glycolysis in red cells appears to be regulated simply by the energy needs of the cell, primarily for maintenance of ion gradients. The balance between ATP consumption and production is controlled allosterically at three sites: the **hexokinase, phosphofructokinase-1, and pyruvate kinase** reactions (see Fig. 12.2). Based on measurements of the V_{max} of the various enzymes in RBC lysates in vitro, hexokinase is present at the lowest activity of all glycolytic enzymes. Its maximal activity is about five times the rate of glucose consumption by the RBC, but it is subject to feedback (allosteric) inhibition by its product Glc-6-P. Hexokinase has 30% homology between its *N*- and *C*-terminal domains, the result of duplication and fusion of a primordial gene; binding of Glc-6-P to the *N*-terminal domain inhibits the activity of the enzyme and production of Glc-6-P at the active site in the C-terminal domain.

Phosphofructokinase-1 (PFK-1)

PFK-1 is the primary site of regulation of glycolysis

PFK-1 is the primary site of regulation of glycolysis,is the primary site of regulation of glycolysis,is the primary site of regulation of glycolysis, is the primary site of regulation of glycolysis, the flux of Fru-6-P to Fru-1,6-BP and, indirectly through the phosphoglucose isomerase reaction, the level of Glc-6-P and inhibition of hexokinase. Although present at 20 times higher activity than hexokinase, PFK-1 is strongly inhibited by ambient ATP, so that its activity varies with the energy status of the cell. Amazingly, ATP is both a substrate (see Fig. 12.3) and an allosteric inhibitor (Fig. 12.7) of PFK-1, a dual function that permits fine control over the activity of the enzyme.

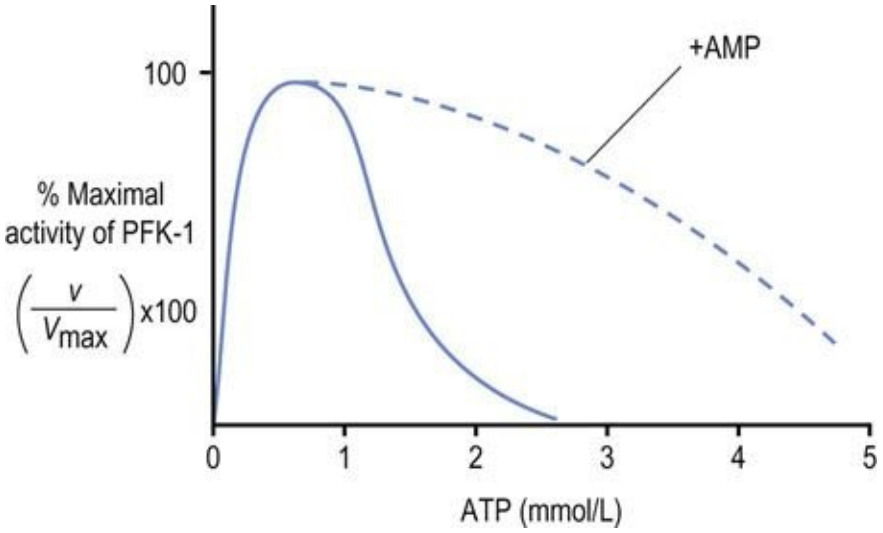


FIG. 12.7 Allosteric regulation of phosphofructokinase-1 (PFK-1) by ATP. AMP is a potent activator of PFK-1 in the presence of ATP.

As shown in Figure 12.7, the concentration of ATP in the RBC (~2 mmol/L) normally suppresses the activity of PFK-1. AMP, which is present at much lower concentration (~0.05 mmol/L), relieves this inhibition. Because of their relative concentrations, a small fractional conversion of ATP to AMP in the RBC yields a large relative increase in AMP concentration, which activates PFK-1. ADP also relieves the inhibition of PFK-1 by ATP, but its concentration does not change as much with energy utilization. AMP (and ADP) not only relieves the inhibition of PFK-1 by ATP but also decreases the K_m for the substrate Fru-6-P, further increasing the catalytic efficiency of the enzyme.

Through allosteric mechanisms, the activity of PFK-1 in the red cell is exquisitely sensitive to changes in the energy status of the cell, as measured by the relative concentrations of ATP, ADP and AMP. In effect, the overall activity of PFK-1, and thus the rate of glycolysis, depends on the cell's (AMP + ADP)/ATP concentration ratio. These products are interconvertible by the adenylate kinase reaction:

$$2\ \text{ADP} \rightleftarrows \text{ATP} + \text{AMP}$$

When ATP is consumed and ADP increases, AMP is formed by the adenylate kinase reaction. The increase in AMP concentrations relieves the inhibition of PFK-1 by ATP, activating glycolysis. The phosphorylation of ADP during

glycolysis, and then of AMP by the adenylate kinase reaction, gradually restores the ATP concentration or **energy charge** of the cell and, as the AMP concentration declines, the rate of glycolysis decreases to a steady-state level. Glycolysis operates at a fairly constant rate in the red cell, where ATP consumption is steady, but the activity of this pathway changes rapidly in response to ATP utilization in muscle during exercise.

Pyruvate kinase (PK)

In addition to regulation by hexokinase and PFK-1, pyruvate kinase in liver is allosterically activated by Fru-1,6-BP, the product of the PFK-1 reaction. This process, known as feed-forward regulation, may be important in the RBC to limit the accumulation of chemically reactive triose phosphate intermediates in the cytosol.

Characteristics of regulatory enzymes

Each of the three enzymes involved in regulation of glycolysis – hexokinase, PFK-1, and pyruvate kinase – has the characteristic features of a regulatory enzyme: they are dimeric or tetrameric enzymes whose structure and activity are responsive to allosteric modulators; they are present at low V_{max} in comparison with other enzymes in the pathway; and they catalyze irreversible reactions.

The regulation of glycolysis in liver, muscle, and other tissues is more complicated than in the RBC (Table 12.1) because of greater variability in the rate of fuel consumption and the interplay between carbohydrate and lipid metabolism during aerobic metabolism. In these tissues, the amount and activity of the regulatory enzymes are regulated by other allosteric effectors, by covalent modification, and by induction or repression of enzyme activity.

Table 12.1

Regulation of glycolysis in the red cell

Enzyme	Regulator
Hexokinase	Inhibited by glucose-6-P
Phosphofructokinase-1	Inhibited by ATP; activated by AMP
Pyruvate kinase	Activated by fructose-1,6-BP

Clinical box Glycolysis in tumor cells

Tumors are often said to go glycolytic, *i.e.* to increase their reliance on glycolysis as a source of energy. The increase in glycolysis might result from inhibition of mitochondrial oxidative phosphorylation as a result of hypoxia, possibly because the metabolic requirements of rapidly dividing tumor cells exceed the supply of oxygen and nutrients from blood. In these cases, the production and accumulation of lactate may become toxic to the tumor cell, contributing to necrosis and formation of a necrotic core in the tumor.

Some tumors secrete cytokines that promote angiogenesis (neovascularization), thereby increasing their fuel supply and enhancing tumor growth. Angiogenesis inhibitors, designed to

inhibit the vascularization of the tumor, are being evaluated as a non-surgical approach to tumor therapy. The ability to survive by relying on glycolysis in hypoxic environments may be an important factor in tumor survival and growth.

Synthesis of 2,3-bisphosphoglycerate (2,3-BPG)

2,3-BPG is a negative allosteric effector of the oxygen affinity of hemoglobin

2,3-Bisphosphoglycerate (Fig. 12.8) is an important by-product of glycolysis in the RBC, sometimes reaching 5 mmol/L concentration, which is comparable to the molar concentration of hemoglobin (Hb) in the RBC. 2,3-BPG is the major phosphorylated intermediate in the erythrocyte, present at even higher concentrations than ATP (1–2 mmol/L) or inorganic phosphate (1 mmol/L). 2,3-BPG is a negative allosteric effector of the O_2 affinity of Hb. It decreases the O_2 affinity of hemoglobin, promoting the release of O_2 in peripheral tissue. The presence of 2,3-BPG in the RBC explains the observation that the O_2 affinity of purified HbA is greater than that of whole RBCs. 2,3-BPG concentration increases in the RBC during adaptation to high altitude, in chronic obstructive pulmonary disease, and in anemia, promoting the release of O_2 to tissues when the O_2 tension and saturation of hemoglobin is decreased in the lung. **Fetal Hb (HbF) is less sensitive than adult Hb (HbA) to the effects of 2,3-BPG;** the higher oxygen affinity of HbF, even in the presence of 2,3-BPG, promotes efficient transfer of O_2 across the placenta from HbA to HbF (see Chapter 5).

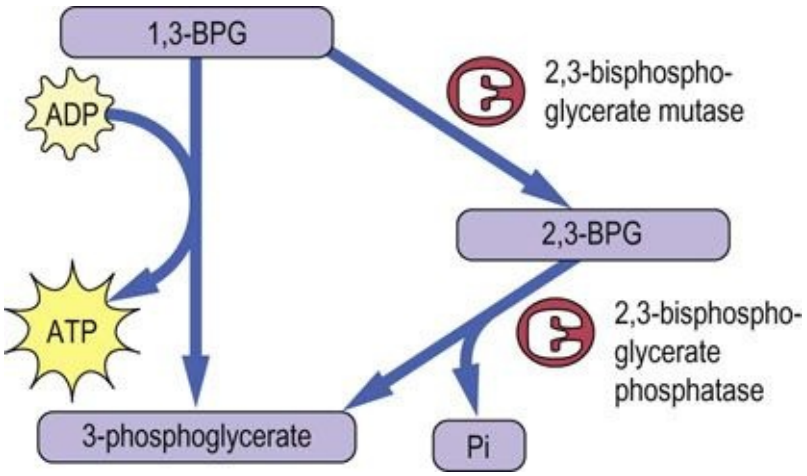


FIG. 12.8 Pathway for biosynthesis and degradation of 2,3-bisphosphoglycerate (2,3-BPG).
BPG mutase catalyzes the conversion of 1,3-BPG to 2,3-BPG. This same enzyme has bisphosphoglycerate phosphatase activity, so that it controls both the synthesis and hydrolysis of 2,3-BPG. Note that this pathway bypasses the phosphoglycerate kinase reaction, so that the overall yield of ATP per mole of glucose is decreased to zero.

The pentose phosphate pathway

Overview

The pentose phosphate pathway is the source of ribose phosphate for synthesis of RNA and DNA

The pentose phosphate pathway is a cytosolic pathway present in all cells, so named because it is the primary pathway for formation of pentose phosphates for synthesis of nucleotides for incorporation into DNA and RNA (see Chapter 30). This pathway branches from glycolysis at the level of Glc-6-P: thus, its alternative designation, the hexose monophosphate shunt. The pentose phosphate pathway is sometimes described as a shunt, rather than a pathway, because when pentoses are not needed for biosynthetic reactions, the pentose phosphate intermediates are recycled back into the mainstream of glycolysis by conversion into Fru-6-P and glyceraldehyde-3-phosphate. This rerouting is especially important in the RBC and in nondividing or quiescent cells, where there is limited need for synthesis of DNA and RNA.

NADPH is a major product of the pentose phosphate pathway in all cells

In tissues with active lipid biosynthesis, *e.g.* liver, adrenal cortex or lactating mammary glands, the NADPH is used in redox reactions required for biosynthesis of cholesterol, bile salts, steroid hormones and triglycerides. The liver also uses NADPH for hydroxylation reactions involved in the detoxification and excretion of drugs. The RBC has little biosynthetic activity, but still shunts about 10% of glucose through the pentose phosphate pathway, in this case almost exclusively for the production of NADPH. The NADPH is used primarily for the reduction of a cysteine-containing tripeptide, glutathione (GSH), an essential cofactor for antioxidant protection (Chapter 37).

The pentose phosphate pathway is divided into an irreversible redox stage, which yields both NADPH and pentose phosphates, and a reversible interconversion stage, in which excess pentose phosphates are converted into glycolytic intermediates. Both stages are important in the RBC, since it needs NADPH for reduction of glutathione, but has limited need for de novo

synthesis of pentoses.

The redox stage of the pentose phosphate pathway – synthesis of NADPH

NADPH is synthesized by two dehydrogenases, in the first and third reactions of the pentose phosphate pathway (Fig. 12.9). In the first step of the pathway, the Glc-6-P dehydrogenase (G6PDH) reaction produces NADPH by oxidation of Glc-6-P to 6-phosphogluconic acid lactone, a cyclic sugar ester. The lactone is hydrolyzed to 6-phosphogluconic acid by lactonase. Oxidative decarboxylation of 6-phosphogluconate, catalyzed by 6-phosphogluconate dehydrogenase, then yields the ketose sugar, ribulose 5-phosphate, plus 1 mole of CO_2, and the second mole of NADPH.

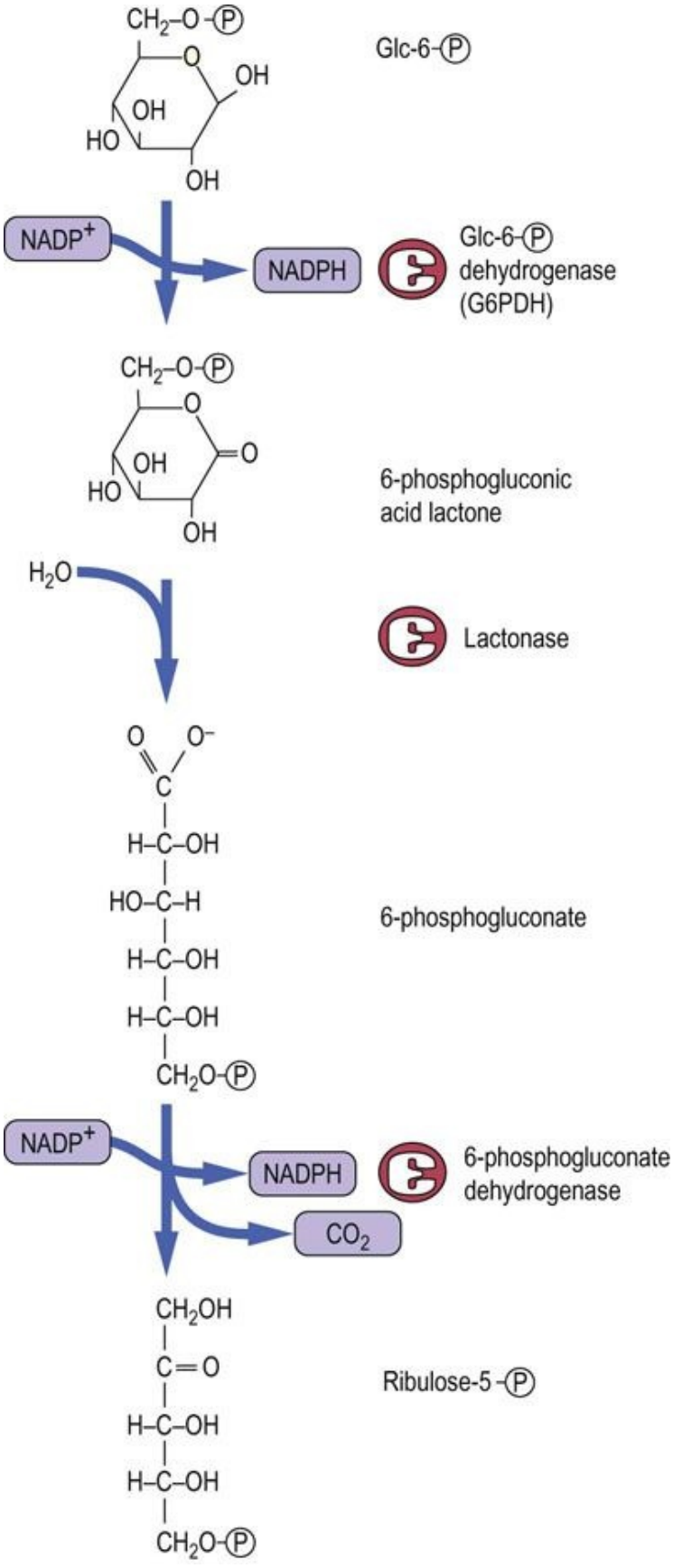

CH2–O-Ⓟ
O
OH
OH
HO
OH
Glc-6-Ⓟ
NADP+
NADPH
Glc-6-Ⓟ dehydrogenase (G6PDH)
CH2–O-Ⓟ
O
OH
HO
O
OH
6-phosphogluconic acid lactone
H2O
Lactonase
O
O–
C
H–C–OH
HO–C–H
H–C–OH
H–C–OH
CH2O-Ⓟ
6-phosphogluconate
NADP+
NADPH
CO2
6-phosphogluconate dehydrogenase
CH2OH
C=O
H–C–OH
H–C–OH
CH2O-Ⓟ
Ribulose-5-Ⓟ

FIG. 12.9 The redox stage of the pentose phosphate pathway.
A sequence of three enzymes forms 2 moles of NADPH per mole of Glc-6-P, which is converted into ribulose-5-phosphate, with evolution of CO_2.

G6PDH and 6-phosphogluconate dehydrogenase maintain a cytoplasmic ratio of $NADPH/NADP^+$ ~ 100. Interestingly, because NAD^+ is required for glycolysis, the ratio of $NADH/NAD^+$ in the cytoplasm is nearly the inverse, less than 0.01. Although the total concentrations (oxidized plus reduced forms) of NAD(H) and NADP(H) in the RBC are similar (~25 µmol/L), the cell maintains these two redox systems with similar redox potentials at such different set-points in the same cell by isolating their metabolism through the specificity of cytoplasmic dehydrogenases. **Glycolytic enzymes (GAPDH and LDH) use only NAD(H), while pentose phosphate pathway enzymes use only NADP(H).** There are no enzymes in the RBC that catalyze the reduction of NAD^+ by NADPH, so that high levels of both NAD^+ and NADPH can exist simultaneously in the same compartment.

The interconversion stage of the pentose phosphate pathway

In cells with active nucleic acid synthesis, ribulose-5-phosphate from the 6-phosphoglucose dehydrogenase reaction is isomerized to ribose-5-phosphate for synthesis of ribo-and deoxyribonucleotides for RNA and DNA (Fig. 12.10). However, in nondividing cells, the pentose phosphates are routed back to glycolysis. This is accomplished by a series of equilibrium reactions in which 3 moles of ribulose-5-phosphate are converted into 2 moles of Fru-6-P and 1 mole of glyceraldehyde-3-phosphate. Certain restrictions are imposed on the interconversion reactions – they may be carried out only by transfer of two or three carbon units between sugar phosphates. Each reaction must also involve a ketose donor and an aldose receptor. Isomerases and epimerases convert ribulose-5-phosphate to the aldose-and ketose-phosphate substrates for the interconversion stage. **Transketolase, a thiamine-dependent enzyme**, catalyzes the two-carbon transfer reactions. **Transaldolase** acts similarly to the aldolase in glycolysis, except that the three-carbon unit is transferred to another sugar, rather than released as a free triose phosphate for glycolysis.

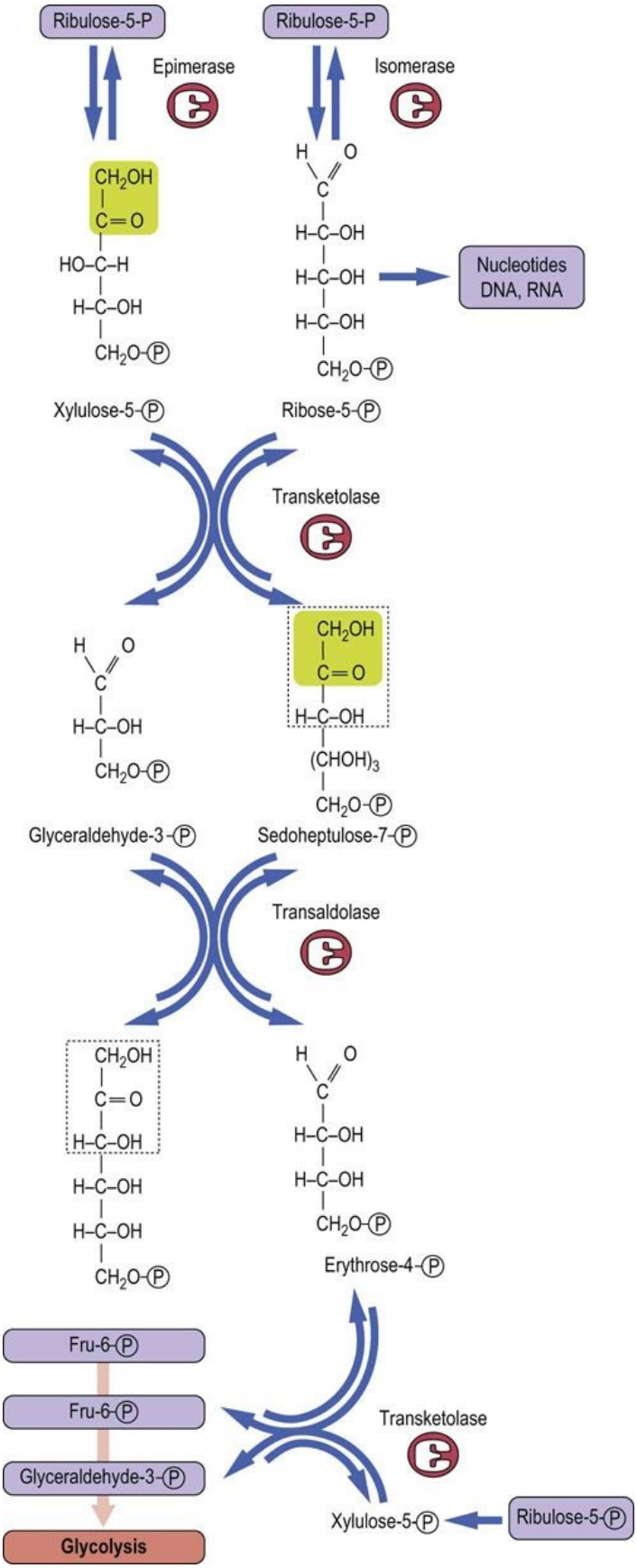

Ribulose-5-P
Epimerase
Ribulose-5-P
Isomerase
Nucleotides
DNA, RNA
Xylulose-5-Ⓟ
Ribose-5-Ⓟ
Transketolase
Glyceraldehyde-3-Ⓟ
Sedoheptulose-7-Ⓟ
Transaldolase
Erythrose-4-Ⓟ
Fru-6-Ⓟ
Fru-6-Ⓟ
Glyceraldehyde-3-Ⓟ
Glycolysis
Transketolase
Xylulose-5-Ⓟ
Ribulose-5-Ⓟ

FIG. 12.10 The interconversion stage of the pentose phosphate pathway. The carbon skeletons of three molecules of ribulose-5-phosphate are shuffled to form two molecules of Fru-6-P and one molecule of glyceraldehyde 3-phosphate, which enter into glycolysis.

As shown in Figure 12.10 and Table 12.2, two molecules of ribulose 5-phosphate, the first pentose product of the redox stage, are converted into separate products: one molecule is isomerized to the aldose sugar ribose-5-phosphate, and the other is epimerized to xylulose-5-phosphate. Transketolase then catalyzes transfer of two carbons from xylulose-5-phosphate to ribose-5-phosphate, yielding a seven-carbon ketose sugar, sedoheptulose-7-phosphate, and the three-carbon glyceraldehyde-3-phosphate. Transaldolase then catalyzes a three-carbon transfer between the two transketolase products, from sedoheptulose-7-phosphate to glyceraldehyde-3-phosphate, yielding the first glycolytic intermediate, Fru-6-P, and a residual erythrose-4-phosphate. A third molecule of xylulose-5-phosphate donates two carbons to erythrose-4-phosphate in a second transketolase reaction, yielding a second molecule of Fru-6-P and a molecule of glyceraldehyde-3-phosphate, both of which enter glycolysis.

Table 12.2

Summary of equilibrium reactions in the pentose phosphate pathway

Substrate(s)		Product(s)	Enzyme
Ribulose-5-P	⇌	Ribose-5-P	Isomerase
2 Ribulose-5-P	⇌	2 Xylulose-5-P	Epimerase
Xylulose-5-P + Ribose-5-P	⇌	Glyceraldehyde-3-P + Sedoheptulose-7-P	Transketolase
Sedoheptulose-7-P + Glyceraldehyde-3-P	⇌	Erythrose-4-P + Fructose-6-P	Transaldolase
Xylulose-5-P + Erythrose-4-P	⇌	Glyceraldehyde-3-P + Fructose-6-P	Transketolase
3 Ribulose-5-P	⇌	**Glyceraldehyde-3-P + 2 Fructose-6-P**	**SUMMARY**

Thus, three five-carbon sugar phosphates (ribulose-5-phosphate) formed in the redox stage of the pentose phosphate pathway are converted into one three-carbon (glyceraldehyde-3-phosphate) and two six-carbon (fructose-6-phosphate) intermediates for glycolysis. In the RBC, these glycolytic intermediates continue through glycolysis to lactate, illustrating that glucose is only temporarily shunted away from the mainstream of glycolysis.

Antioxidant function of the pentose phosphate pathway

The pentose phosphate pathway protects against oxidative damage in the red cell

Glutathione (GSH) is a tripeptide γ-glutamyl-cysteinyl-glycine (Fig. 12.11). It is present in cells at 2–5 mmol/L, 99% in the reduced (thiol) form, and is an essential coenzyme for protection of the cell against a range of oxidative and chemical insults (Chapter 37). Most of the NADPH formed in the red cell is used by glutathione reductase to maintain GSH in the reduced state. During its function as a coenzyme for antioxidant activities, GSH is oxidized to the disulfide form, GSSG, which is then regenerated by the action of glutathione

reductase (Fig. 12.12).

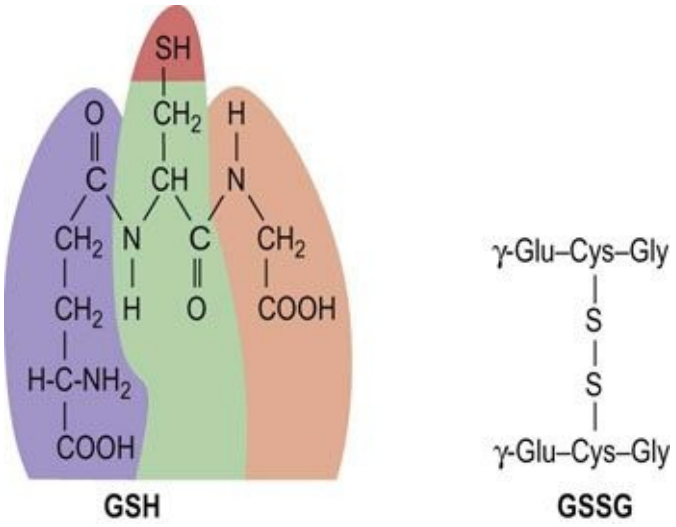


FIG. 12.11 Glutathione.
Structure of reduced glutathione (GSH) and oxidized glutathione (GSSG). Note the isopeptide bond between the γ-carboxyl, rather than the α-carboxyl, of glutamic acid and the α-amino group of cysteine.

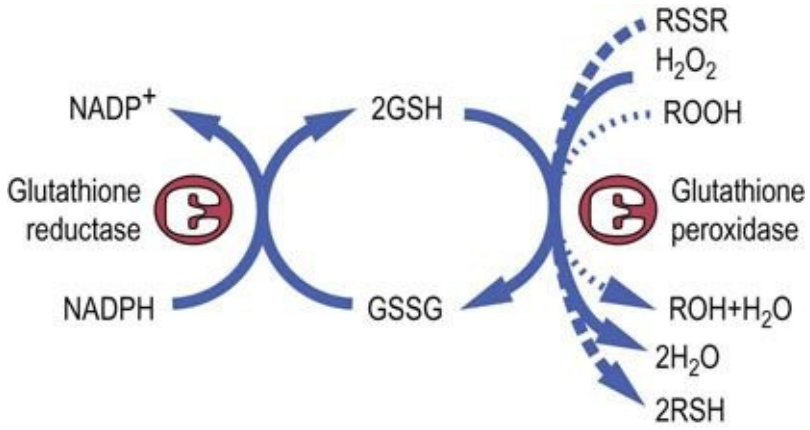


FIG. 12.12 Antioxidant activities of glutathione.
GSH is the coenzyme for glutathione peroxidases which detoxify hydrogen peroxide and organic (lipid) hydroperoxides. Hydrogen peroxide and lipid peroxides are formed spontaneously in the red cell, catalyzed by side reactions of heme iron during oxygen transport on hemoglobin (Chapter 37). GSH also reduces disulfide bonds in proteins (RSSR), formed during oxidative stress (Chapter 37), regenerating the native form of the protein (RSH).

GSH has a range of protective functions in the cell. Glutathione peroxidase (GPx) is found in all cells and uses GSH for detoxification of hydrogen peroxide and organic (lipid) peroxides in the cytosol and cell membranes (see Fig. 12.12). Because GPx contains a selenocysteine residue in its active site, selenium, which is required in trace amounts in the diet, is often described as an antioxidant nutrient (see Chapter 11).

GSH also acts as an intracellular sulfhydryl buffer, maintaining exposed –SH groups on proteins and enzymes in the reduced state. Under normal circumstances, when proteins are exposed to O_2, their free sulfhydryl groups gradually oxidize to form disulfides, either intramolecularly or by intermolecular crosslinking with other protein molecules. In the red cell, GSH maintains the –SH groups of hemoglobin in the reduced state, inhibiting disulfide crosslinking of the protein.

Clinical box Glucose-6-phosphate dehydrogenase deficiency causes hemolytic anemia

Just prior to a planned departure to the tropics, a patient visited his physician, complaining of weakness, and noting that his urine had recently become unexplainably dark. Physical examination revealed slightly jaundiced (yellow, icteric) sclera. Laboratory tests indicated a low hematocrit, a high reticulocyte count, and a significantly increased blood level of bilirubin. The patient had been quite healthy during a previous visit a month earlier when he received immunizations and prescriptions for antimalarial drugs.

Comment.

A number of drugs, particularly primaquine and related antimalarials, undergo redox reactions in the cell, producing large quantities of reactive oxygen species (ROS) (Chapter 38). The ROS cause oxidation of −SH groups in hemoglobin and peroxidation of membrane lipids. Some persons have a genetic defect in Glc-6-P dehydrogenase (G6PDH), typically yielding an unstable enzyme that has a shorter half-life in the RBC or is unusually sensitive to inhibition by NADPH. In either case,

because of the decreased activity of this enzyme and insufficient production of NADPH under stress, the cell's ability to recycle GSSG to GSH is impaired, and drug-induced oxidative stress leads to excessive damage and lysis of RBCs (hemolysis) and hemolytic anemia. Bilirubin, a brown pigment produced by heme metabolism, overloads hepatic detoxification pathways, and also accumulates in plasma and tissues, causing jaundice. If the hemolysis is severe enough, Hb spills over into the urine, resulting in hematuria and dark-colored urine. Heinz bodies, disulfide crosslinked aggregates of hemoglobin, are also apparent in blood smears. G6PDH deficiency is typically asymptomatic, except in response to an oxidative challenge, which may be induced by drugs (antimalarials, sulfa drugs), diet (fava beans) or severe infection.

There are over 200 known mutations of the *G6PDH* gene, yielding a wide variation in severity of disease. The RBC appears to be especially sensitive to oxidative stress, because, unlike other cells, it cannot synthesize and replace enzymes. Older cells, which have lower G6PDH activity, are therefore particularly affected. The activity of all enzymes in the RBC declines with the age of the cell, and cell death eventually results from inability of the cells to produce sufficient ATP for maintenance of cellular ion gradients. The gradual decline in activity of the pentose phosphate pathway in older cells is one mechanism leading to oxidative crosslinking of membrane proteins and loss of membrane elasticity, leading to entrapment and turnover of the RBC in the spleen.

Summary

This chapter describes two ancient metabolic pathways common to all cells in the body: glycolysis and the pentose phosphate pathway. The RBC, which lacks mitochondria and the capability for oxidative metabolism and obtains all of its ATP energy by glycolysis, is used as a model for introducing these pathways.

■ Anaerobic glycolysis in the RBC provides a limited amount of ATP by conversion of the six-carbon sugar glucose to two molecules of the three-carbon hydroxyacid lactate.

■ Through a series of sugar phosphate intermediates, glycolysis provides metabolites for branch points to numerous other metabolic pathways, including the pentose phosphate pathway.

■ The pentose phosphate pathway provides pentoses for synthesis of DNA and RNA in nucleated cells, and NADPH for biosynthetic reactions. NADPH is also required for maintenance of reduced glutathione, which is an essential cofactor for antioxidant defense systems that protect the cell against oxidative stress.

Active learning

1. Why was glucose selected as blood sugar during evolution, rather than other sugars, *e.g.* galactose, fructose or sucrose?
2. Describe coupled enzymatic reactions, using only red cell enzymes and a spectrometer for measuring NAD(P)(H) production or consumption, that could be used to measure blood glucose and lactate concentrations.
3. Explain the metabolic origin of acidosis in chronic obstructive pulmonary disease.

Further reading

Bar-Even, A, Flamholz, A, Noor, E, *et al.* Rethinking glycolysis: on the biochemical logic of metabolic pathways. *Nat Chem Biol*. 2012; 8:509–517.

Climent, F, Roset, F, Repiso, A, *et al.* Red cell glycolytic enzyme disorders caused by mutations: an update. *Cardiovasc Hematol Disord Drug Targets*. 2009; 9:95–106.

Granchi, C, Minutolo, F. Anticancer agents that counteract tumor glycolysis. *Chem Med Chem*. 2012; 7:1318–1350.

Katz, SE. *The art of fermentation*. White River Junction, VT: Chelsea Green Publishing; 2012.

Krol, DM, Nedley, MP. Dental caries: state of the science for the most common chronic disease of childhood. *Adv Pediatr*. 2007; 54:215–239.

Nicholson, JK, Holmes, E, Kinross, J, *et al.* Host-gut microbiota metabolic interactions. *Science*. 2012; 336:1262–1267.

Wamelink, MM, Struys, EA, Jakobs, C. The biochemistry, metabolism and inherited defects of the pentose phosphate pathway: a review. *J Inherit Metab Dis*. 2008; 31:703–717.

Weissman, JS, Coyle, W. Stool transplants: ready for prime time? *Curr Gastroenterol Rep*. 2012; 14:313–316.

Websites

American Society of Hematology, case studies on anemia. http://teachingcases.hematology.org/.

CHAPTER 13

Carbohydrate Storage and Synthesis in Liver and Muscle

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

- Describe the structure of glycogen.
- Identify the primary sites of glycogen storage in the body and the function of glycogen in these tissues.
- Outline the metabolic pathways for synthesis and degradation of glycogen.
- Describe the mechanism by which glycogen is mobilized in liver in response to glucagon, in muscle during exercise, and in both tissues in response to epinephrine.
- Explain the origin and consequences of glycogen storage diseases in liver and muscle.
- Describe the mechanism for counterregulation of glycogenolysis and glycogenesis in liver.
- Outline the pathway of gluconeogenesis, including substrates, unique enzymes and regulatory mechanisms.
- Describe the complementary roles of glycogenolysis and gluconeogenesis in maintenance of blood glucose concentration.

Introduction

The red cell and the brain have an absolute requirement for blood glucose for energy metabolism. Together, they consume about 80% of the 200 g of glucose consumed in the body per day. There are only about 10 g of glucose in the plasma and extracellular fluid volume, so that blood glucose must be replenished constantly. Otherwise, hypoglycemia develops and compromises brain function, leading to confusion and disorientation, and possibly life-threatening coma at blood glucose concentrations below 2.5 mmol/L (45 mg/dL). We absorb glucose from our intestines for only 2–3 h following a carbohydrate-containing meal, so there must be a mechanism for maintenance of blood glucose between meals.

Glycogen, a polysaccharide storage form of glucose, is our first line of defense against declining blood glucose concentration. During and immediately following a meal, glucose is converted into glycogen, a process known as **glycogenesis**, in both liver and muscle. The tissue concentration of glycogen is higher in liver than in muscle but because of the relative masses of muscle and liver, the majority of glycogen in the body is stored in muscle (Table 13.1).

Table 13.1

Tissue distribution of carbohydrate energy reserves (70-kg adult)

Tissue	Type	Amount	% of tissue mass	Calories
Liver	Glycogen	75 g	3–5%	300
Muscle	Glycogen	250 g	0.5–1.0%	1000
Blood and extracellular fluid	Glucose	10 g	–	40

Hepatic glycogenolysis and gluconeogenesis are required for maintenance of normal blood glucose concentration

Hepatic glycogen is gradually degraded between meals, by the pathway of glycogenolysis, releasing glucose to maintain blood glucose concentration. However, total hepatic glycogen stores are barely sufficient for maintenance of blood glucose concentration during a 12-h fast.

During sleep, when we are not eating, there is a gradual shift from **glycogenolysis** to de novo synthesis of glucose, also an hepatic pathway, known as **gluconeogenesis** (Fig. 13.1). Gluconeogenesis is essential for survival during fasting or starvation, when glycogen stores are depleted. The liver uses amino acids from muscle protein as the primary precursor of glucose, but also makes use of lactate from glycolysis and glycerol from fat catabolism. Fatty acids, mobilized from adipose tissue triglyceride stores, provide the energy for gluconeogenesis.

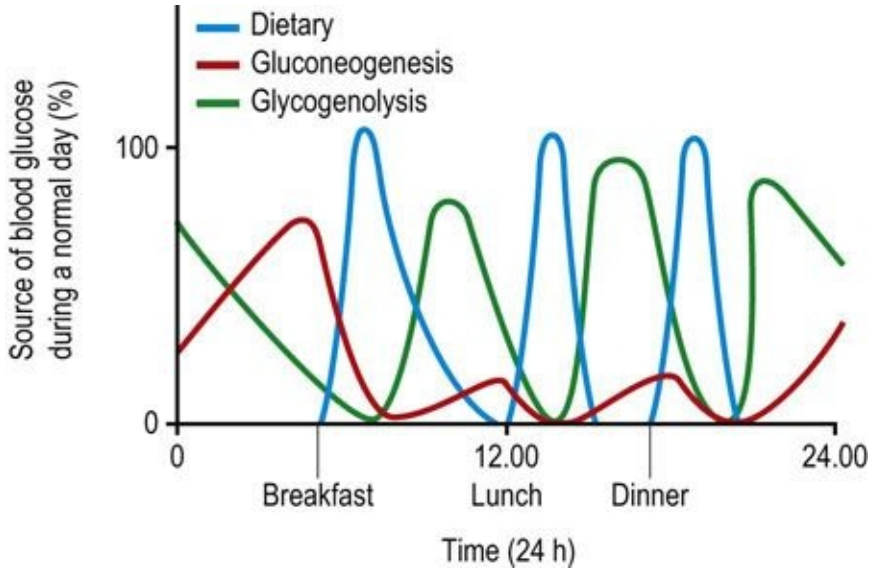


FIG. 13.1 Sources of blood glucose during a normal day.
Between meals, blood glucose is derived primarily from hepatic glycogen. Depending on the frequency of snacking, glycogenolysis and gluconeogenesis may be more or less active during the day. Late in the night or in early morning, following depletion of a major fraction of hepatic glycogen, gluconeogenesis becomes the primary source of blood glucose.

Glycogen is stored in muscle for use in energy metabolism

Muscle glycogen is not available for maintenance of blood glucose. Glucose obtained from blood and glycogen is used exclusively for energy metabolism in

muscle, especially during bursts of physical activity. Although cardiac and skeletal muscles rely on fats as their primary source of energy, some glucose metabolism is essential for efficient fat metabolism in these tissues.

This chapter describes the pathways of glycogenesis and glycogenolysis in liver and muscle, and the pathway of gluconeogenesis in liver.

Structure of glycogen

Glycogen, a highly branched glucan, is the storage form of glucose in tissues

Glycogen is a branched polysaccharide of glucose. It contains only two types of glycosidic linkages, chains of α1→4-linked glucose residues with α1→6 branches spaced about every 4–6 residues along the α1→4 chain (Fig. 13.2). Glycogen is closely related to **starch**, the storage polysaccharide of plants, but starch consists of a mixture of amylose and amylopectin. The amylose component contains only linear α1→4 chains; the amylopectin component is more glycogen-like in structure but with fewer α1→6 branches, about one per 12 α1→4-linked glucose residues. The gross structure of glycogen is dendritic in nature, expanding from a core sequence bound to a tyrosine residue in the protein **glycogenin** and developing into a final structure resembling a head of cauliflower. The enzymes of glycogen metabolism are bound to the surface of the glycogen particle; many terminal glucose molecules on the surface of the molecule provide ready access for rapid release of glucose from the glycogen polymer.

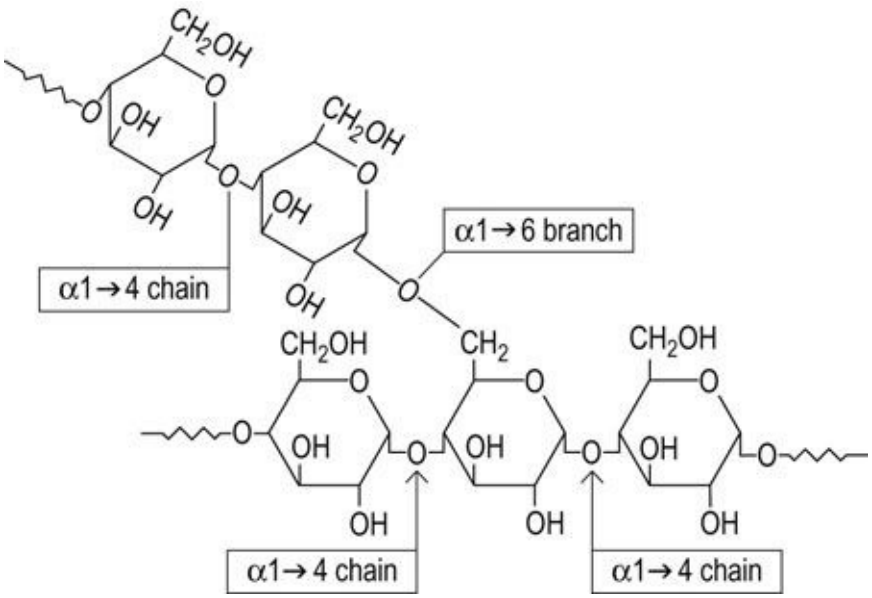


FIG. 13.2 Close-up of the structure of glycogen.

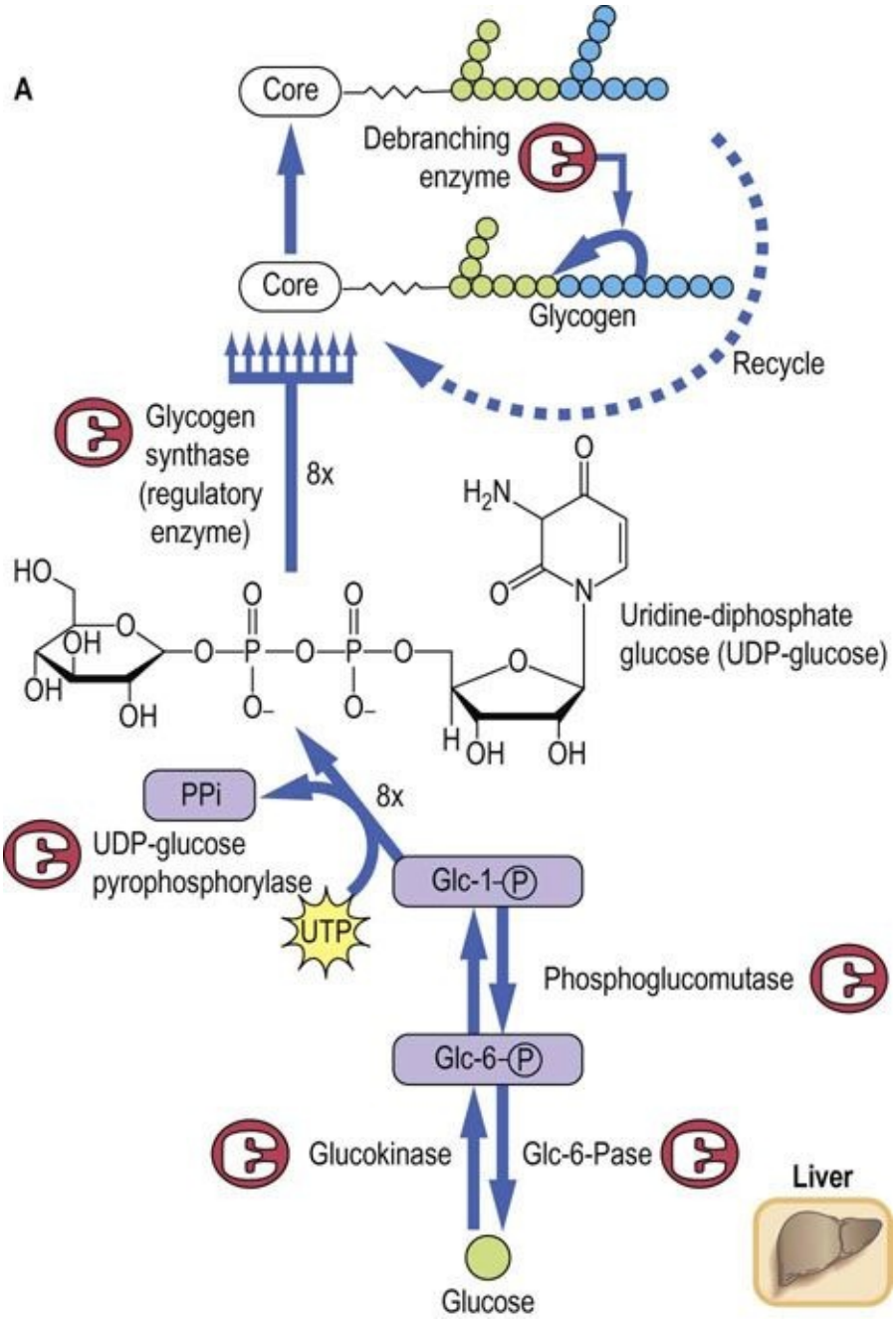
The figure shows α1→4 chains and an α1→6 branch point. Glycogen is stored as granules in liver and muscle cytoplasm.

Pathway of glycogenesis from blood glucose in liver

Glycogenesis is activated in liver and muscle following a meal

The liver is rich in the high-capacity, low-affinity (K_m >10 mmol/L) glucose transporter **GLUT-2**, making it freely permeable to glucose delivered at high concentration in portal blood during and following a meal (see Table 8.2). The liver is also rich in **glucokinase**, an enzyme that is specific for glucose and converts it into glucose 6-phosphate (Glc-6-P). Glucokinase (GK) is inducible by continued consumption of a high-carbohydrate diet. It has a high K_m, about 5–7 mmol/L, so that it is poised to increase in activity as portal glucose increases above the normal 5 mmol/L (100 mg/dL) blood glucose concentration. Unlike hexokinase, GK is not inhibited by Glc-6-P, so that the concentration of Glc-6-P increases rapidly in liver following a carbohydraterich meal, forcing glucose into all the major pathways of glucose metabolism: glycolysis, the pentose phosphate pathway, and glycogenesis (see Fig. 12.2). Glucose is channeled into glycogen, providing a carbohydrate reserve for maintenance of blood glucose during the postabsorptive state. Excess Glc-6-P in liver, beyond that needed to replenish glycogen reserves, is then funneled into glycolysis, in part for energy production but primarily for conversion into fatty acids and triglycerides, which are exported for storage in adipose tissue. Glucose that passes through the liver causes an increase in peripheral blood glucose concentration following carbohydraterich meals. This glucose is used in muscle for synthesis and storage of glycogen and in adipose tissue as a source of glycerol for triglyceride biosynthesis.

The pathway of glycogenesis from glucose (Fig. 13.3A) involves four steps:


A
Core
Debranching enzyme
Core
Glycogen
Recycle
Glycogen synthase (regulatory enzyme)
8x
Uridine-diphosphate glucose (UDP-glucose)
PPi
8x
UDP-glucose pyrophosphorylase
UTP
Glc-1-Ⓟ
Phosphoglucomutase
Glc-6-Ⓟ
Glucokinase
Glc-6-Pase
Liver
Glucose

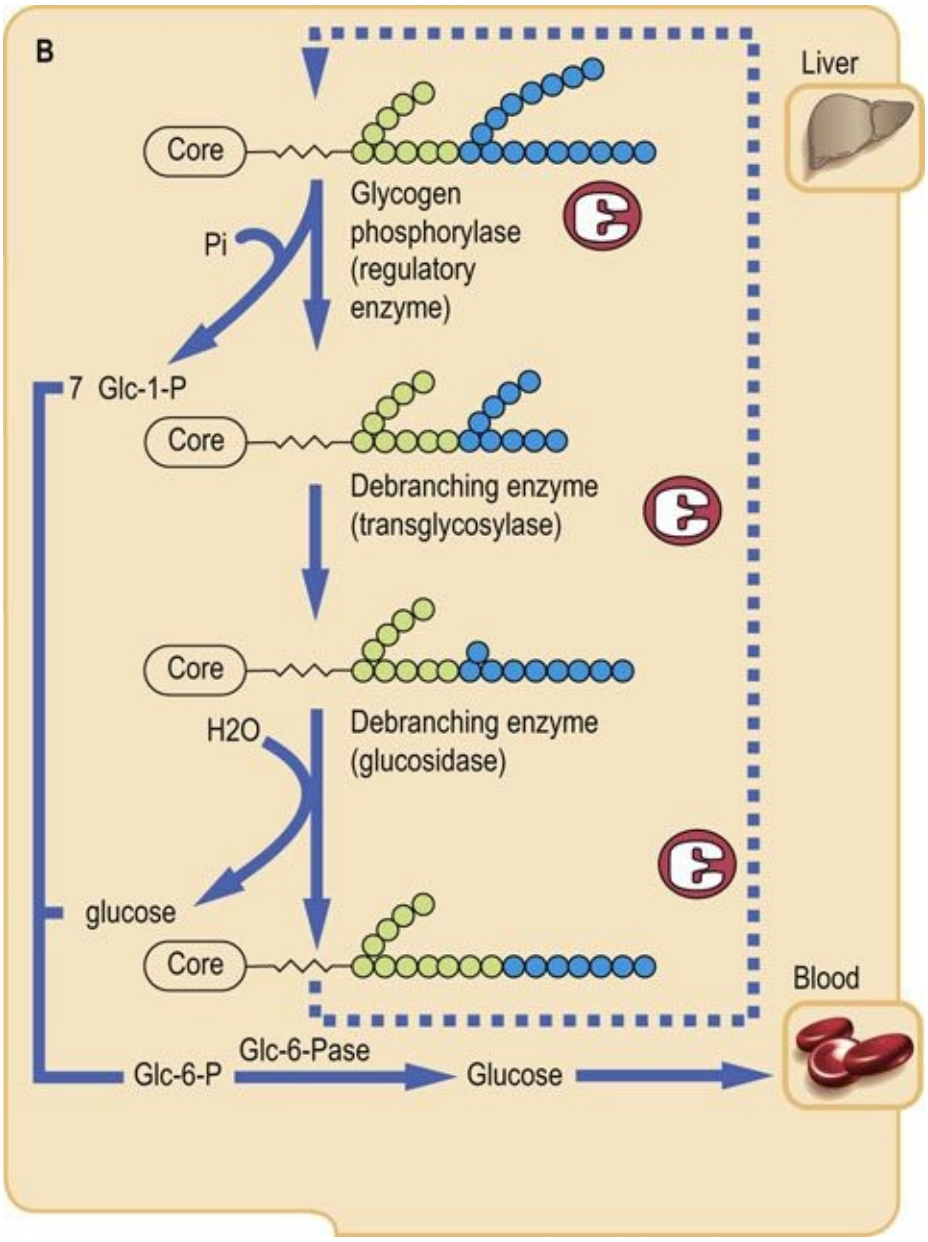


FIG. 13.3 Pathways of glycogenesis (A) and glycogenolysis (B).

■ Conversion of Glc-6-P into glucose-1-phosphate (Glc-1-P) by phosphoglucomutase.
■ Activation of Glc-1-P to the sugar nucleotide uridine diphosphate (UDP)-glucose by the enzyme UDP-glucose pyrophosphorylase.
■ Transfer of glucose from UDP-Glc to glycogen in α1→4 linkage by glycogen synthase, a member of the class of enzymes known as glycosyl transferases.
■ When the α1→4 chain exceeds eight residues in length, glycogen branching enzyme, a transglycosylase, transfers some of the α1→4-linked sugars to an

α1→6 branch, setting the stage for continued elongation of both α1→4 chains until they, in turn, become long enough for transfer by branching enzyme.

Glycogen synthase is the regulatory enzyme for glycogenesis, rather than UDP-glucose pyrophosphorylase, because UDP-glucose is also used for synthesis of other sugars, and as a glycosyl donor for synthesis of glycoproteins, glycolipids and proteoglycans (Chapters 27–29). Pyrophosphate (PPi), the other product of the pyrophosphorylase reaction, is a high energy phosphate anhydride. It is rapidly hydrolyzed to inorganic phosphate by pyrophosphatase, providing the thermodynamic driving force for biosynthesis of glycogen.

Pathway of glycogenolysis in liver

Hepatic glycogen phosphorylase provides for rapid release of glucose into blood during the postabsorptive state

As with most metabolic pathways, separate enzymes, sometimes in separate subcellular compartments, are required for the forward and reverse pathways. The pathway of glycogenolysis (Fig. 13.3B) begins with removal of the abundant, external α1→4-linked glucose residues in glycogen. This is accomplished not by a hydrolase but by **glycogen phosphorylase**, an enzyme that uses cytosolic phosphate and releases glucose from glycogen in the form of Glc-1-P. The Glc-1-P is isomerized by phosphoglucomutase to Glc-6-P, placing it at the top of the glycolytic pathway; the phosphorylase reaction, in effect, bypasses the requirement for ATP in the hexokinase or glucokinase reactions. In liver, the glucose is released from Glc-6-P by glucose-6-phosphatase (Glc-6-Pase), and the glucose exits via the GLUT-2 transporter into blood. The rate-limiting, regulatory step in glycogenolysis is catalyzed by phosphorylase, the first enzyme in the pathway.

Phosphorylase is specific for α1→4 glycosidic linkages; it cannot cleave α1→6 linkages. Further, this large enzyme cannot approach the branching glucose residues efficiently. Thus, as shown in Figure 13.3B, phosphorylase cleaves the external glucose residues until the branches are three or four residues long, then **debranching enzyme**, which has both transglycosylase and glucosidase activity, moves a short segment of glucose residues bound to the α1→6 branch to the end of an adjacent α1→4 chain, leaving a single glucose residue at the branch point. This glucose is then removed by the exo-1,6-glucosidase activity of debranching enzyme, allowing glycogen phosphorylase to proceed with degradation of the extended α1→4 chain until another branch point is approached, setting the stage for a repeat of the transglycosylase and glucosidase reactions. About 90% of the glucose is released from glycogen as Glc-1-P, and the remainder, derived from the α1→6 branching residues, as free glucose.

Clinical box Von gierke's disease: glycogen storage disease caused by

glucose-6-phosphatase deficiency

A baby girl was chronically cranky, irritable, sweaty, and lethargic, and demanded food frequently. Physical evaluation indicated an extended abdomen, resulting from an enlarged liver. Blood glucose, measured 1 h after feeding, was 3.5 mmol/L (70 mg/dL); normal value <5 mmol/L (100 mg/dL). After 4 h, when the child was exhibiting irritability and sweating, her heart rate was increased (pulse = 110), and blood glucose had declined to 2 mmol/L (40 mg/dL). These symptoms were corrected by feeding. A liver biopsy showed massive deposition of glycogen particles in the liver cytosol.

Comment.

This child cannot mobilize glycogen. Because of the severity of hypoglycemia, the most likely mutation is in hepatic Glc-6-Pase, which is required for glucose production by both glycogenolysis and gluconeogenesis. Treatment involves frequent feeding with slowly digested carbohydrate, *e.g.* uncooked starch, and nasogastric drip-feeding during the night.

Hormonal regulation of hepatic glycogenolysis

Three hormones (insulin, glucagon and cortisol) counter-regulate glycogenolysis and glycogenesis

Glycogenolysis is activated in liver in response to a demand for blood glucose, either because of its utilization during the postabsorptive state or in preparation for increased glucose utilization in response to stress. There are three major hormonal activators of glycogenolysis: glucagon, epinephrine (adrenaline), and cortisol (Table 13.2).

Table 13.2

Hormones involved in control of glycogenolysis

Hormone	Source	Initiator	Effect on glycogenolysis
Glucagon	Pancreatic α-cells	Hypoglycemia	Rapid activation
Epinephrine	Adrenal medulla	Acute stress, hypoglycemia	Rapid activation
Cortisol	Adrenal cortex	Chronic stress	Chronic activation
Insulin	Pancreatic β-cells	Hyperglycemia	Inhibition

Glucagon is a peptide hormone (3500 Da), secreted from the **α-cells** of the endocrine pancreas. Its primary function is to activate hepatic glycogenolysis for maintenance of normal blood glucose concentration (normoglycemia). Glucagon has a short half-life in plasma, about 5 min, as a result of receptor binding, renal filtration, and proteolytic inactivation in liver. Glucagon concentration in plasma therefore changes rapidly in response to the need for blood glucose. Blood glucagon increases between meals, decreases during a meal, and is chronically

increased during fasting or on a low-carbohydrate diet (Chapter 21).

Glycogenolysis is also activated in response to both acute and chronic stress. The stress may be:

- physiologic, *e.g.* in response to increased blood glucose utilization during exercise;
- pathologic, *e.g.* as a result of blood loss (shock);
- psychological, *e.g.* in response to acute or chronic threats.

Acute stress, regardless of its source, causes an activation of glycogenolysis through the action of the **catecholamine hormone epinephrine**, released from the adrenal medulla. During prolonged exercise, both glucagon and epinephrine contribute to the stimulation of glycogenolysis and maintenance of blood glucose concentration.

Increased blood concentrations of the adrenocortical steroid hormone cortisol also induce glycogenolysis. Levels of the **glucocorticoid cortisol** vary diurnally in plasma, but may be chronically elevated under continuously stressful conditions, including psychological and environmental (e.g. cold) stress.

Glucagon serves as a general model for the mechanism of action of hormones that act by way of cell surface receptors. Cortisol, which acts at the level of gene expression, will be discussed later in Chapters 35 and 39

Clinical box

McArdle's diseasea glycogen storage disease that reduces capacity for exercise

A 30-year-old man consulted his physician because of chronic arm and leg muscle pains and cramps during exercise. He indicated that he had always had some muscle weakness and, for this reason, was never active in scholastic sports, but the problem did not become severe until he recently enrolled in an exercise program to improve his health. He also noted that the pain generally disappeared after about 15–30 min, and then he could continue his exercise without discomfort. His blood glucose concentration was normal during exercise, but serum creatine kinase (MM isoform from skeletal muscle) was elevated, suggesting muscle damage. Blood glucose declined slightly during 15 min of exercise, but unexpectedly blood lactate also declined, rather than increased, even when he was experien cing muscle cramps. A biopsy indicated an unusually

high level of glycogen in muscle, suggesting a glycogen storage disease.

Comment.

This patient suffers from McArdle's disease, a rare deficiency of muscle phosphorylase activity. The actual enzyme deficiency must be confirmed by enzyme assay, since a number of other mutations could also affect muscle glycogen metabolism. During the early periods of intense exercise, the muscle obtains most of its energy by metabolism of glucose, derived from glycogen. During cramps, which normally occur during oxygen debt, most of the pyruvate produced by glycolysis is excreted into blood as lactate, leading to an increase in blood lactate concentration. In this case, however, the patient had cramps but did not excrete lactate, suggesting a failure to mobilize muscle glycogen to produce glucose. His recovery after 15–30 min results from epinephrine-mediated activation of hepatic glycogenolysis, which provides glucose to blood and relieves the deficit in muscle glycogenolysis. Treatment of McArdle's disease usually involves exercise avoidance or carbohydrate consumption prior to exercise. Otherwise, the course of the disease is uneventful.

Mechanism of action of glucagon

Glucagon activates glycogenolysis during the postabsorptive state

Glucagon binds to an hepatic plasma membrane receptor and initiates a cascade of reactions that lead to mobilization of hepatic glycogen (Fig. 13.4) during the postabsorptive state. On the inside of the plasma membrane there is a class of signal transduction proteins, known as **G-proteins**, that bind guanosine triphosphate (GTP) and guanosine diphosphate (GDP), nucleotide analogues of ATP and ADP. GDP is bound in the resting state. Binding of glucagon to the plasma membrane receptor stimulates exchange of GDP for GTP on the G-protein, and the G-protein then undergoes a conformational change that leads to dissociation of its α-subunit, which then binds to and activates the plasma membrane enzyme **adenylate cyclase**. This enzyme converts cytoplasmic ATP into **cyclic-3′,5′-AMP (cAMP)**, a soluble mediator that is described as the '**second messenger**' for action of glucagon (and other hormones). Cyclic AMP binds to the cytoplasmic enzyme **protein kinase A (PKA)**, causing dissociation of inhibitory (regulatory) subunits from the catalytic subunits of the heterodimeric enzyme, relieving inhibition of PKA (see Chapters 21 and 40), which then phosphorylates serine and threonine residues on target proteins and enzymes.

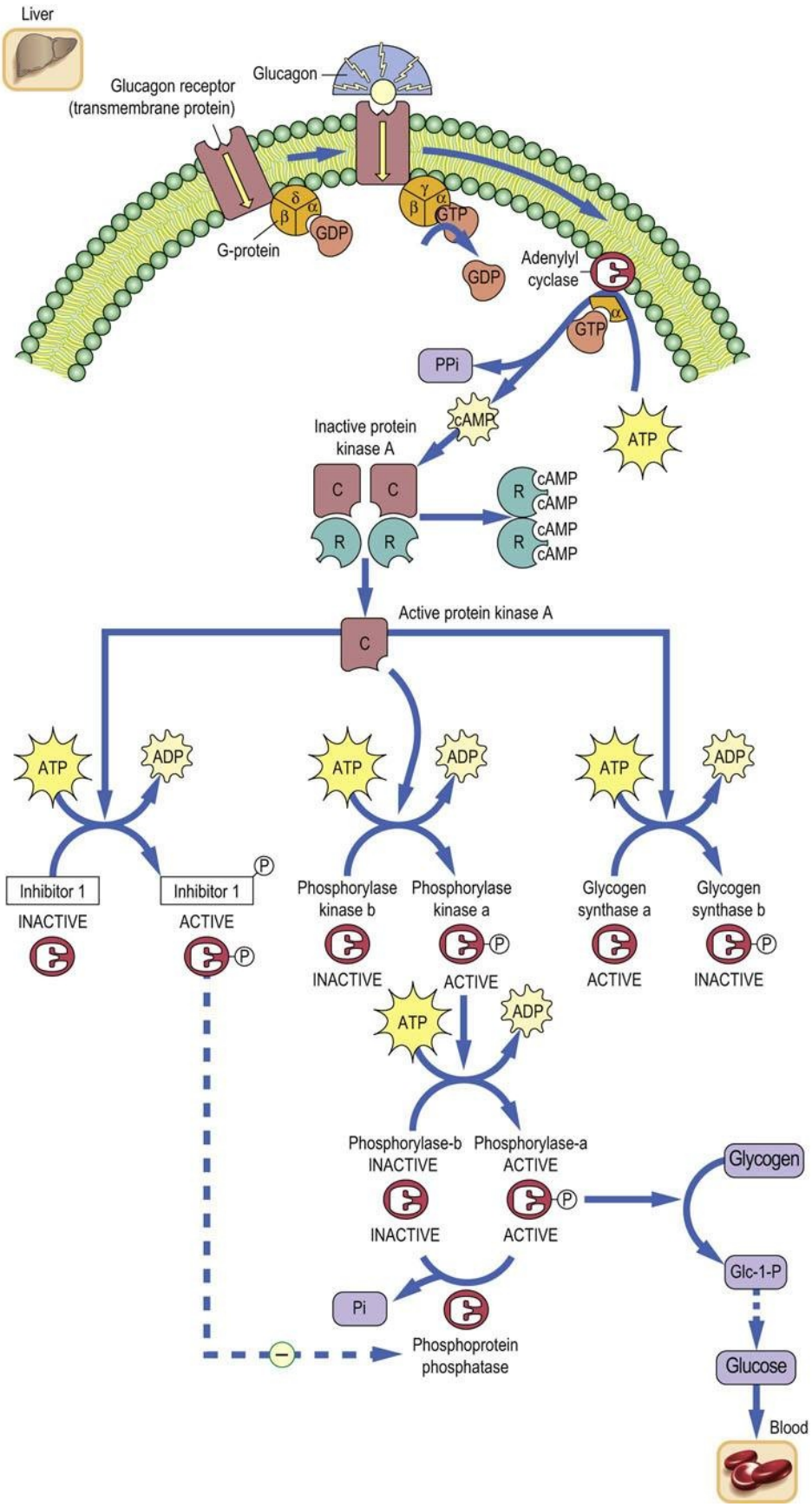

Liver
Glucagon
Glucagon receptor
(transmembrane protein)
G-protein
GDP
GTP
GDP
Adenylyl
cyclase
GTP
PPi
cAMP
ATP
Inactive protein
kinase A
C
C
R
R
cAMP
cAMP
cAMP
cAMP
Active protein kinase A
C
ATP
ADP
ATP
ADP
ATP
ADP
Inhibitor 1
INACTIVE
Inhibitor 1
ACTIVE
P
Phosphorylase
kinase b
Phosphorylase
kinase a
INACTIVE
ACTIVE
Glycogen
synthase a
Glycogen
synthase b
ACTIVE
INACTIVE
ATP
ADP
Phosphorylase-b
INACTIVE
Phosphorylase-a
ACTIVE
INACTIVE
ACTIVE
Glycogen
Glc-1-P
Glucose
Blood
Pi
Phosphoprotein
phosphatase

FIG. 13.4 Cascade amplification system.
Mobilization of hepatic glycogen by glucagon. A cascade of reactions amplifies the hepatic response to glucagon binding to its plasma membrane receptor. cAMP is known as the second messenger of glucagon action. PKA indirectly activates phosphorylase via phosphorylase kinase and directly inactivates glycogen synthase. C, catalytic subunits; R, regulatory (inhibitory) subunits; PKA, protein kinase A.

Advanced concepts box G-proteins

G-proteins are plasma-membrane, guanosine-nucleotide-binding proteins that are involved in signal transduction for a wide variety of hormones (Fig. 13.4; see also Chapter 39). In some cases they stimulate (Gs) and in other cases they inhibit (Gi) protein kinases and protein phosphorylation. G-proteins are closely associated with hormone receptors in plasma membranes and consist of α, β, and γ subunits. The G_{α}-subunit binds GDP in the resting state. Following hormone binding (ligation), the receptor recruits G-proteins, stimulating exchange of GDP for GTP on the G_{α}-subunit. GTP binding leads to release of the β-and γ-subunits, and the α-subunit is then free to bind to and activate adenylate cyclase. The hormonal response is amplified following receptor binding, because a single receptor can activate many α-subunits. Hormonal responses are also turned off at the level of receptors and G-proteins by two mechanisms:

- The G_{α}-subunit has a sluggish guanosine triphosphate phosphatase (GTPase) activity that hydrolyzes GTP, with a half-time measured in minutes, so that it gradually dissociates from, and thereby ceases to activate, adenylate cyclase.
- Phosphorylation of the hormone receptor by protein kinase A decreases its affinity for the hormone, a process described as desensitization or hormone resistance.

Advanced concepts box Protein kinase a is very sensitive to small changes in camp concentration

As illustrated in Figure 13.4, cAMP-dependent PKA is a tetrameric enzyme with two different types of subunits (R_2C_2); the catalytic C-subunit has protein kinase activity, and the regulatory R-subunit inhibits the protein kinase activity. The R-subunit has a sequence of amino acids that would normally be recognized and phosphorylated by the C-subunit, except that this sequence in R contains an alanine, rather than a serine or threonine, residue. Binding of two molecules of cAMP to each R-subunit results in conformational changes that lead to dissociation of a ($cAMP_2$-R_2) dimer from the C-subunits. The monomeric, active C-subunits then proceed to phosphorylate serine and threonine residues in target enzymes. PKA is not a typical allosteric enzyme, in that the binding of the allosteric effector (cAMP) causes subunit dissociation; however, the complete activation of PKA involves cooperative binding of four molecules of cAMP to two R-subunits. PKA is fully activated at submicromolar concentrations of cAMP, so that it is exquisitely sensitive to small changes in adenylate cyclase activity in response to glucagon.

The pathway for activation of glycogen phosphorylase (see Fig. 13.4) involves phosphorylation of many molecules of phosphorylase kinase by PKA, which then phosphorylates and activates many molecules of glycogen phosphorylase. The net effect of these sequential steps, beginning with activation of many molecules of adenylate cyclase by G-proteins, is a '**cascade amplification**' system, not unlike that of a series of amplifiers in a radio or stereo set, resulting in a massive increase in signal strength within seconds after glucagon binding to the hepatocyte plasma membrane. Phosphorylation of phosphorylase activates glycogenolysis, leading to production of Glc-6-P in liver, which is then hydrolyzed to glucose and exported into blood. Another target of PKA is

inhibitor-1, a phosphoprotein phosphatase inhibitor protein, which is activated by phosphorylation. Phosphorylated inhibitor-1 inhibits cytoplasmic phosphoprotein phosphatases, which would otherwise reverse the phosphorylation of enzymes and quench the response to glucagon (see Fig. 13.4).

Glycogenolysis and glycogenesis are counterregulated by protein kinase A, which activates phosphorylase and inhibits glycogen synthase

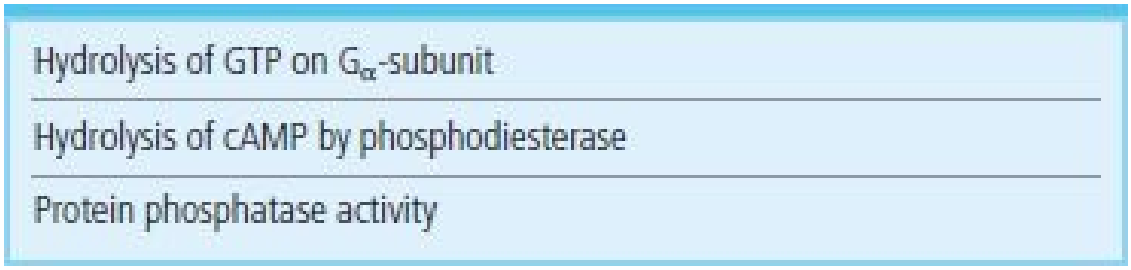
Glycogenolysis and glycogenesis are opposing pathways. Theoretically, Glc-1-P produced by phosphorylase could be rapidly activated to UDP-glucose and reincorporated into glycogen. To prevent this wasteful or **futile cycle**, PKA also acts directly on glycogen synthase, in this case inactivating the enzyme. Thus, the activation of phosphorylase (glycogenolysis) is coordinated with inactivation of glycogen synthase (glycogenesis). Other hepatic biosynthetic pathways, including protein, cholesterol, fatty acid, and triglyceride synthesis, as well as glycolysis, are also regulated by phosphorylation of key regulatory enzymes, generally limiting biosynthetic reactions and focusing liver metabolism in response to glucagon on the provision of glucose to blood for maintenance of vital body functions (see Chapter 21).

Perhaps in order to balance the cascade of events amplifying the response to glucagon, there are multiple, redundant mechanisms to insure rapid termination of the hormonal response (Table 13.3). In addition to the slow **GTPase** activity of the G_{α}-subunit, there is also a **phosphodiesterase** in the cell that hydrolyzes cAMP to AMP, permitting reassociation of the inhibitory and catalytic subunits of PKA, decreasing its protein kinase activity. There are also **phosphoprotein phosphatases** that remove the phosphate groups from the active, phosphorylated forms of phosphorylase kinase and phosphorylase. The decrease in cAMP concentration and PKA activity also leads to decreased phosphorylation of inhibitor-1, permitting increased activity of phosphoprotein phosphatases. Thus, an array of mechanisms act in concert to insure that hepatic glycogenolysis declines rapidly in response to increasing blood glucose and decreasing blood glucagon concentrations following a meal.

Table 13.3

Several mechanisms are involved in terminating the hormonal

response to glucagon



There are a number of autosomal recessive genetic diseases affecting glycogen metabolism (Table 13.4). These diseases, known as **glycogen storage diseases**, are characterized by accumulation of glycogen granules in tissues, which eventually compromises tissue function. Predictably, glycogen storage diseases affecting hepatic glycogen metabolism are characterized by fasting hypoglycemia and may be life-threatening, while defects in muscle glycogen metabolism are characterized by rapid muscle fatigue during exercise.

Table 13.4

Major classes of glycogen storage diseases

Type	Name	Enzyme deficiency	Structural or clinical consequences
I	Von Gierke's	Glc-6-Pase	Severe postabsorptive hypoglycemia, lactic acidemia, hyperlipidemia
II	Pompe's	Lysosomal α-glucosidase	Glycogen granules in lysosomes
III	Cori's	Debranching enzyme	Altered glycogen structure, hypoglycemia
IV	Andersen's	Branching enzyme	Altered glycogen structure
V	McArdle's	Muscle phosphorylase	Excess muscle glycogen deposition, exercise-induced cramps and fatigue
VI	Hers'	Liver phosphorylase	Hypoglycemia, not as severe as type I

Mobilization of hepatic glycogen by epinephrine

Epinephrine activates glycogenolysis during stress, increasing blood glucose concentration

Epinephrine works through several distinct receptors on different cells. The best studied of these receptors are the α-and β-adrenergic receptors; they recognize different features of the epinephrine molecule, bind epinephrine with different affinities, work by different mechanisms, and are inhibited by different classes of drugs. During severe hypoglycemia, glucagon and epinephrine work together to magnify the glycogenolytic response in liver. However, even when blood glucose is normal, epinephrine is released in response to real or perceived threats, causing an increase in blood glucose to support a 'fight or flight' response. **Caffeine** in coffee and **theophylline** in tea are inhibitors of phosphodiesterase and also cause an increase in hepatic cAMP and blood glucose. Like epinephrine, caffeine, administered in the form of a few strong cups of coffee, can also make us alert and responsive – and aggressive.

Epinephrine action on hepatic glycogenolysis proceeds by two pathways. One of these, through the epinephrine **β-adrenergic receptor**, is similar to that for glucagon, involving a plasma membrane epinephrine-specific receptor, G-proteins, and cAMP. The epinephrine response augments the effects of glucagon during severe hypoglycemia (metabolic stress) and also explains, in part, the rapid heartbeat, sweating, tremors, and anxiety associated with hypoglycemia. Epinephrine also works simultaneously through an α-adrenergic receptor, but by a different mechanism. Binding to α-receptors also involves G-proteins, common elements in hormone signal transduction, but in this case, the G-protein is specific for activation of a membrane isozyme of **phospholipase C (PLC)**, which is specific for cleavage of a membrane phospholipid, **phosphatidylinositol bisphosphate (PIP_2)** (Fig. 13.5). Both products of PLC action, **diacylglycerol (DAG)** and **inositol trisphosphate (IP_3)**, act as second messengers of epinephrine action. DAG activates **protein kinase C (PKC)** which, like PKA, initiates phosphorylation of serine and threonine residues on target proteins. IP_3 promotes the transport of Ca^{2+} into the cytosol. Ca^{2+} then binds to the cytoplasmic protein calmodulin, which binds to and activates

phosphorylase kinase, leading to cAMP-independent phosphorylation and activation of phosphorylase. A Ca^{2+}–calmodulin-dependent protein kinase and other enzymes are also activated, either by phosphorylation or by association with the **Ca^{2+}–calmodulin complex** (Fig. 13.5). Thus, a range of metabolic pathways is activated in response to stress, especially those involved in the mobilization of energy reserves.

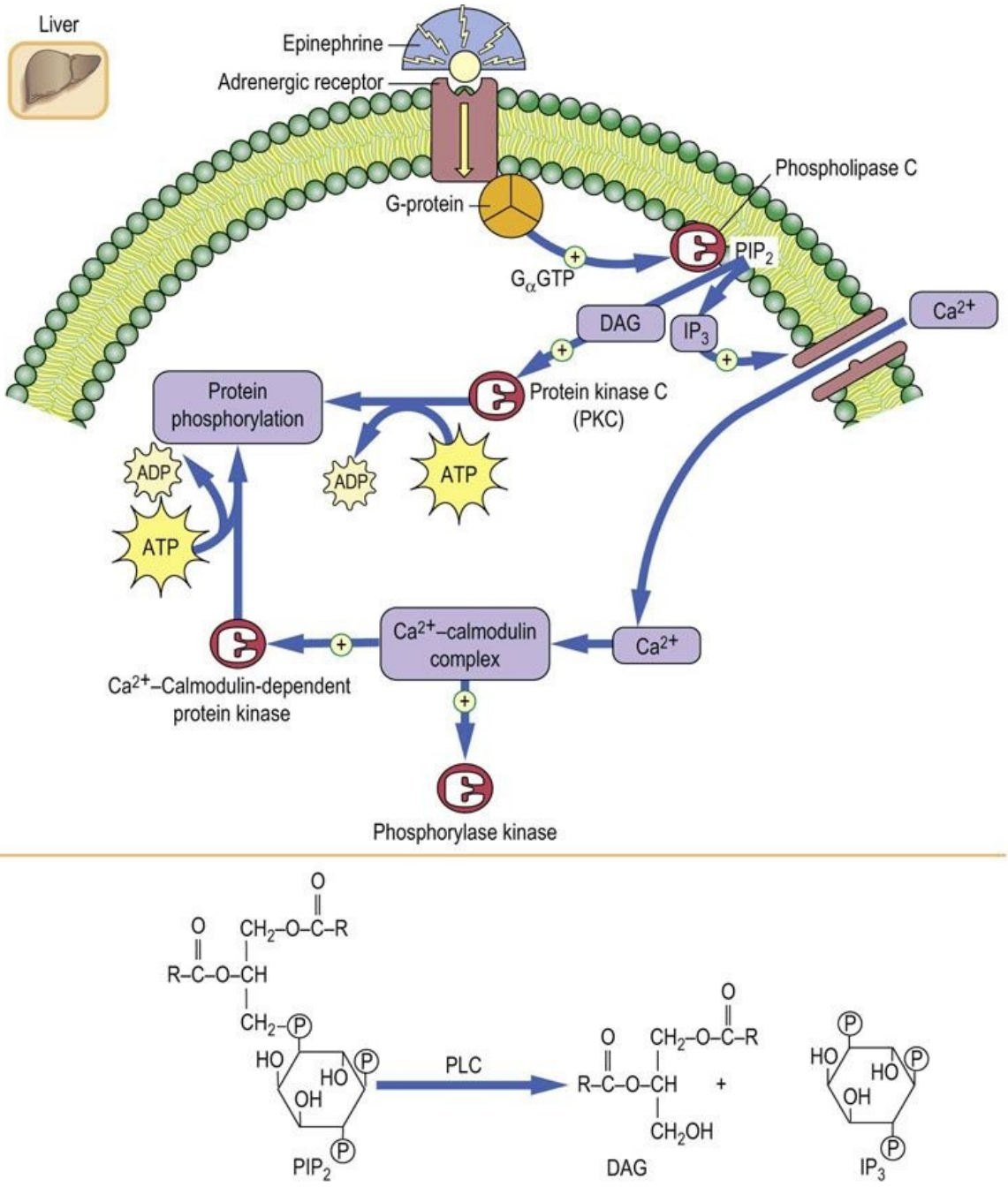


FIG. 13.5 Mechanism of activation of glycogenolysis in liver via the α-adrenergic receptor.
Diacylglycerol (DAG) and inositol trisphosphate (IP_3) are second messengers mediating the α-adrenergic response. PIP_2, phosphatidylinositol bisphosphate; PKC, protein kinase C. See also Chapter 40.

Glycogenolysis in muscle

Muscle lacks a glucagon receptor and glucose-6-phosphatase, and retains glucose for energy metabolism, even during hypoglycemia

The tissue localization of hormone receptors provides tissue specificity to hormone action. Thus, only those tissues with glucagon receptors respond to glucagon. Muscle may be rich in glycogen, even during hypoglycemia, but it lacks both the glucagon receptor and Glc-6-Pase. Therefore muscle glycogen cannot be mobilized to replenish blood glucose. Muscle glycogenolysis is activated in response to epinephrine through the cAMP-dependent **β-adrenergic receptor**, but the glucose is metabolized through glycolysis for energy production. This occurs not only during 'fight or flight' situations but also in response to metabolic demands during prolonged exercise. There are also two important hormone-independent mechanisms for activation of glycogenolysis in muscle (Fig. 13.6). First, the influx of Ca^{2+} into the muscle cytoplasm in response to nerve stimulation activates the basal, unphosphorylated form of phosphorylase kinase by action of the Ca^{2+}–calmodulin complex. This hormone-independent activation of phosphorylase provides for rapid activation of glycogenolysis during short bursts of exercise, even in the absence of epinephrine action. A second mechanism for activation of muscle glycogenolysis involves direct allosteric activation of phosphorylase by AMP. Increased usage of ATP during a rapid burst of muscle activity leads to rapid accumulation of ADP, which is converted in part into AMP by action of the enzyme **myokinase (adenylate kinase)**, which catalyzes the reaction:

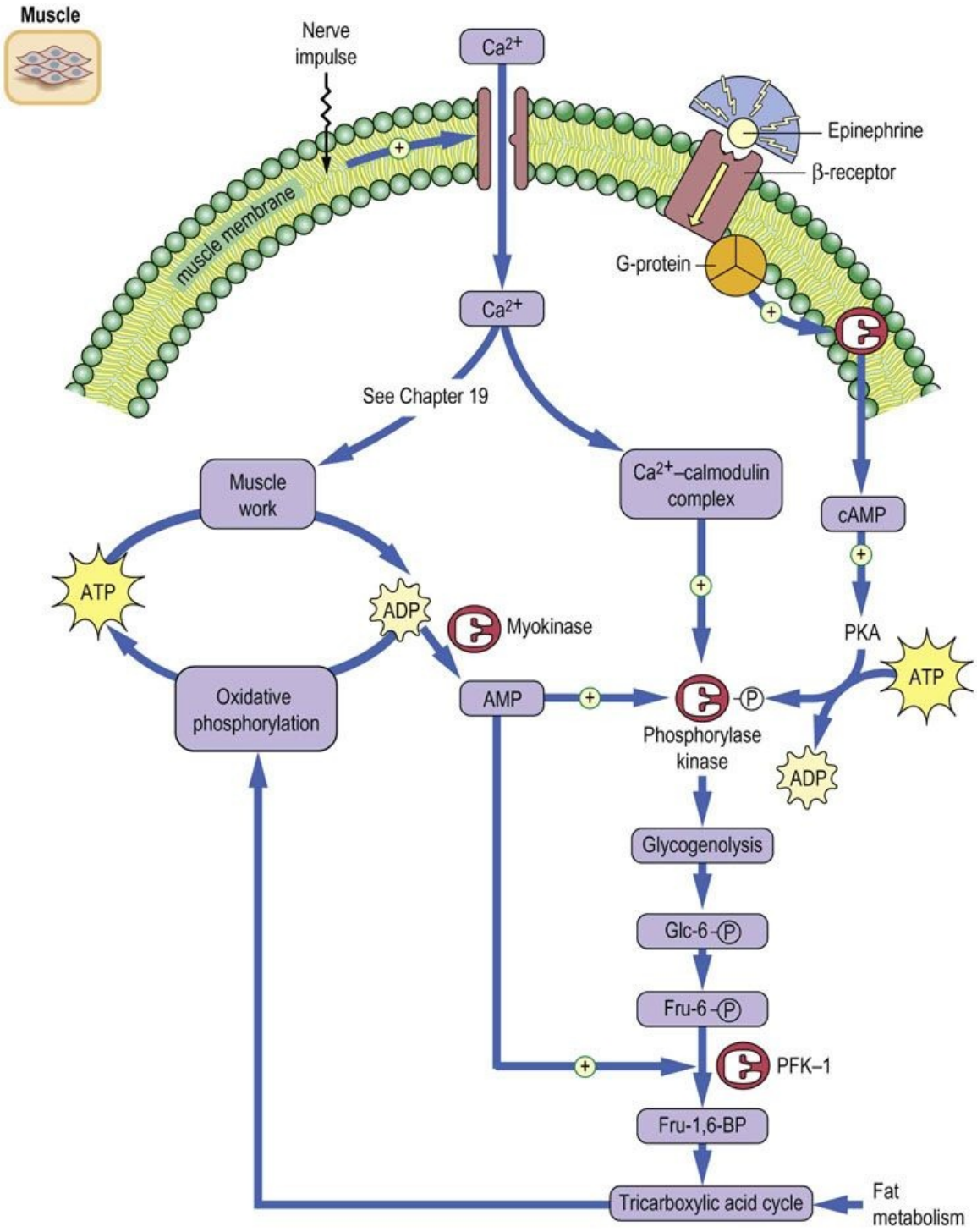


FIG. 13.6 Regulation of protein kinase A (PKA) in muscle.
Activation of glycogenolysis and glycolysis in muscle during exercise. PFK-1, phosphofructokinase-1. Compare Figure 8.4.

$$2\ ADP \rightleftarrows ATP + AMP$$

Advanced concepts box Maximal inhibition of glycogen synthase is achieved only through sequential action of several kinases

When both glucagon and epinephrine are acting on liver, the activation of glycogenolysis and inhibition of glycogenesis is mediated by at least three kinases: protein kinase A (PKA), protein kinase C (PKC), and Ca^{2+}–calmodulin activated protein kinase. All three of these protein kinases phosphorylate key serine and threonine residues in regulatory enzymes. These and other protein kinases work in concert with one another in a process known as sequential or hierarchical phosphorylation, leading to phosphorylation of up to nine amino acid residues on glycogen synthase. Maximal inhibition of glycogen synthase is achieved only through the sequential activity of several kinases. In some cases, certain serine or threonine residues must be phosphorylated in a specific sequence by cooperative action of different kinases, *i.e.* phosphorylation of one site by one enzyme requires prior phosphorylation of another site by a separate enzyme.

AMP activates both the basal and phosphorylated forms of phosphorylase, enhancing glycogenolysis in either the absence or presence of hormonal stimulation. AMP also relieves inhibition of phosphofructokinase-1 (PFK-1) by ATP (see Chapter 12), stimulating the utilization of glucose through glycolysis for energy production. The stimulatory effects of Ca^{2+} and AMP insure that the muscle can respond to its energy needs, even in the absence of hormonal input.

Regulation of glycogenesis

Insulin opposes the action of glucagon and stimulates gluconeogenesis

Glycogenesis, and energy storage in general, occurs during and immediately following meals. Glucose and other carbohydrates, rushing into the liver from the intestines via the portal circulation, are efficiently trapped to make glycogen. Excess glucose proceeds to the peripheral circulation, where it is taken up into muscle and adipose tissue for energy reserves or storage. We normally eat sitting down, rather than during exercise, so that the opposing pathways of uptake and storage versus mobilization and utilization of energy supplies are temporally compartmentalized functions in our lives.

Energy storage is under the control of the **polypeptide hormone insulin**, which is stored in **β-cells** in the pancreatic islets of Langerhans (Chapter 21). Insulin is secreted into blood following a meal, tracking blood glucose concentration. It has two primary functions in carbohydrate metabolism: first, insulin reverses the actions of glucagon in phosphorylation of proteins, turning off glycogen phosphorylase and activating glycogen synthase, promoting glucose storage; second, it stimulates the uptake of glucose into peripheral tissues (muscle and adipose tissue), facilitating synthesis and storage of glycogen and triglycerides. Insulin also acts at the level of gene expression, stimulating the synthesis of enzymes involved in carbohydrate metabolism and storage and conversion of glucose into triglycerides. It also acts as a growth hormone, stimulating protein turnover, both synthesis and degradation.

Protein tyrosine phosphorylation, rather than serine and threonine phosphorylation, is a characteristic feature of insulin and growth factor activity. Insulin binding to its transmembrane receptor (Fig. 13.7) stimulates aggregation of receptors and promotes **tyrosine kinase** activity in the intracellular domain of the receptor. The insulin receptor **autophosphorylates** its tyrosine residues, enhancing its protein tyrosine kinase activity, and phosphorylates tyrosine residues in other intracellular effector proteins, which then activate secondary pathways. Among these are kinases that phosphorylate serine and threonine residues on proteins, but at sites and on proteins distinct from those phosphorylated by PKA and PKC. Insulin-dependent activation of GTPase, phosphodiesterase and phosphoprotein phosphatases also checks the action of

glucagon, which is typically present at high concentration in the blood at mealtimes, *i.e.* several hours since the last meal.

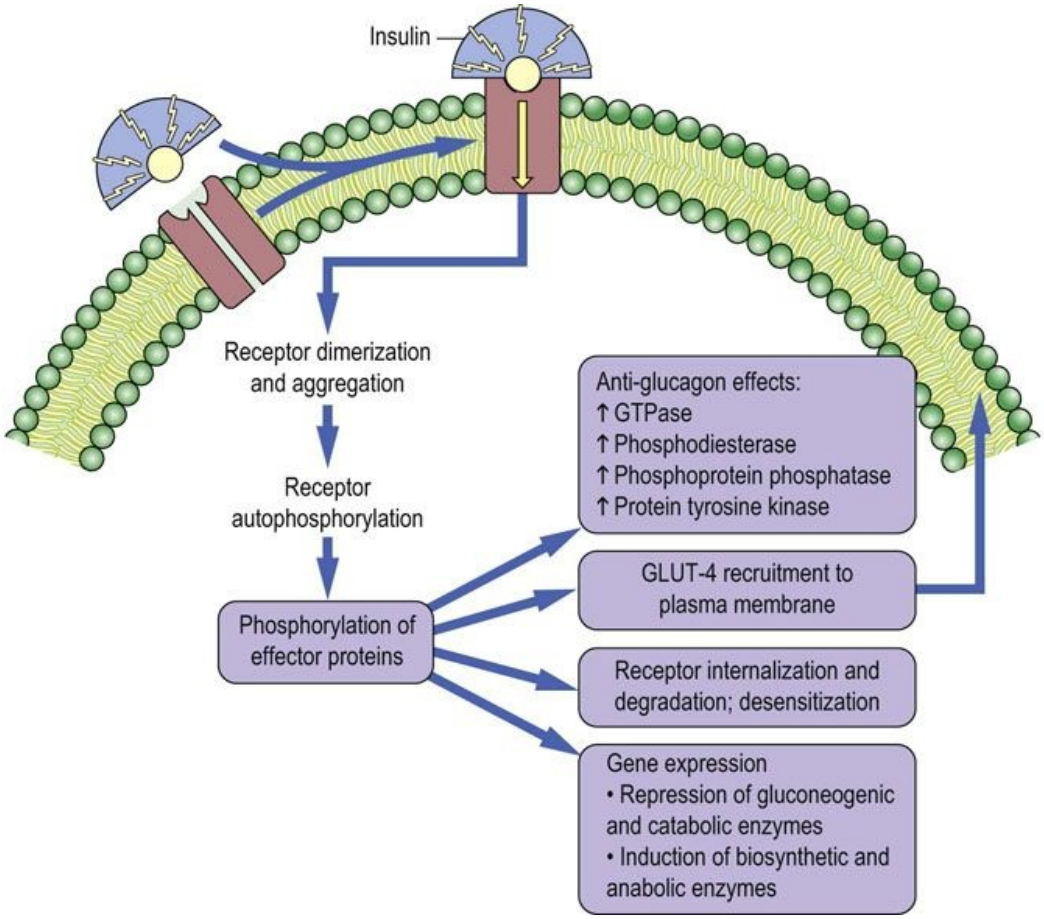


FIG. 13.7 Mechanisms of insulin action.
Regulatory effects of insulin on hepatic and muscle carbohydrate metabolism. (See also Chapter 21.)

The liver also appears to be directly responsive to ambient blood glucose concentration, since hepatic glycogen synthesis increases following a meal, even in the absence of hormonal input. Thus, the increase in hepatic glycogenesis begins more rapidly than the increase in insulin concentration in blood. Perfusion of liver with glucose solutions in vitro, in the absence of insulin, also leads to inhibition of glycogenolysis and activation of glycogenesis. This appears to occur by direct allosteric inhibition of phosphorylase by glucose and secondary stimulation of protein phosphatase activity.

Most, if not all, cells in the body are responsive to insulin in some way, but

the major sites of insulin action, on a mass basis, are muscle and adipose tissue. These tissues normally have low levels of cell surface glucose transporters, restricting the entry of glucose – they rely mostly on lipids for energy metabolism. In muscle and adipose tissue, insulin receptor tyrosine kinase activity induces movement of glucose transporter-4 (**GLUT-4**; see Table 8.2) from intracellular vacuoles to the cell surface, increasing glucose transport into the cell. The glucose is then used in muscle for synthesis of glycogen, and in adipose tissue to produce glyceraldehyde-3-phosphate which is converted to glycerol-3-phosphate for synthesis of triglycerides (Chapter 16). The insulin-stimulated, GLUT-4-mediated uptake of glucose into muscle and adipose tissue is the primary mechanism limiting the increase in blood glucose following a meal.

Clinical box large child born of a diabetic mother

A baby boy, born of a poorly controlled, chronically hyperglycemic, diabetic mother, was large and chubby (macrosomic) at birth (5 kg) but appeared otherwise normal. He declined rapidly, however, and within 1 h showed all the symptoms of hypoglycemia, similar to the case of the baby girl born of a malnourished mother. The difference, in this case, was that the boy was obviously on the heavy side, rather than thin and malnourished.

Comment.

This child has experienced a chronically hyperglycemic environment during uterine development. He adapted by increasing endogenous insulin production, which has a growth hormone-like activity, resulting in macrosomia. At birth, when placental delivery of glucose ceases, he has a normal blood glucose concentration and a substantial supply of hepatic glycogen. However, chronic hyperinsulinemia prior to birth probably represses gluconeogenic enzymes, and his high blood insulin concentration at birth promotes glucose uptake into muscle and adipose tissue. In the absence of a maternal source of glucose, insulin-induced hypoglycemia leads to a stress response, which was corrected by glucose infusion. After 1–2 days, his ample body mass will provide

a good reservoir for synthesis of blood glucose from muscle protein.

Gluconeogenesis

Gluconeogenesis is required to maintain blood glucose during fasting and starvation

Unlike glycogenolysis, which can be turned on rapidly in response to hormonal stimulation, gluconeogenesis increases more slowly, depending on changes in gene expression, and reaches maximal activity over a period of hours (see Fig. 13.1); it becomes the primary source of our blood glucose concentration about 8 hours into the postabsorptive state (Chapter 21). Gluconeogenesis requires both a source of energy for biosynthesis and a source of carbons for formation of the backbone of the glucose molecule. The energy is provided by metabolism of fatty acids released from adipose tissue. The carbon skeletons are provided from three primary sources:

- Lactate produced in tissues such as the red cell and muscle.
- Amino acids derived from muscle protein.
- Glycerol released from triglycerides during lipolysis in adipose tissue.

Among these, **muscle protein is the major precursor of blood glucose during fasting and starvation** – the rate of gluconeogenesis is often limited by the availability of substrate, including the rate of proteolysis in muscle or, in some cases, muscle mass. During prolonged fasting, malnutrition or starvation, we lose both adipose and muscle mass. The fat is used both for the general energy needs of the body and to support gluconeogenesis, while most of the amino acids in protein are converted into glucose; urinary nitrogen (urea) excretion is also increased.

Gluconeogenesis from lactate

Gluconeogenesis uses lactate, amino acids and glycerol as substrates for synthesis of glucose; fatty acids provide the energy

Gluconeogenesis from lactate is conceptually the opposite of anaerobic glycolysis but proceeds by a slightly different pathway, involving both mitochondrial and cytosolic enzymes (Fig. 13.8). During hepatic gluconeogenesis lactate is converted back into glucose, using, in part, the same glycolytic enzymes involved in conversion of glucose into lactate. The lactate

cycle involving the liver, red cells, and muscle, known as the **Cori cycle**, is discussed in detail in Chapter 21. At this point, we focus on the metabolic pathway for conversion of lactate to glucose.

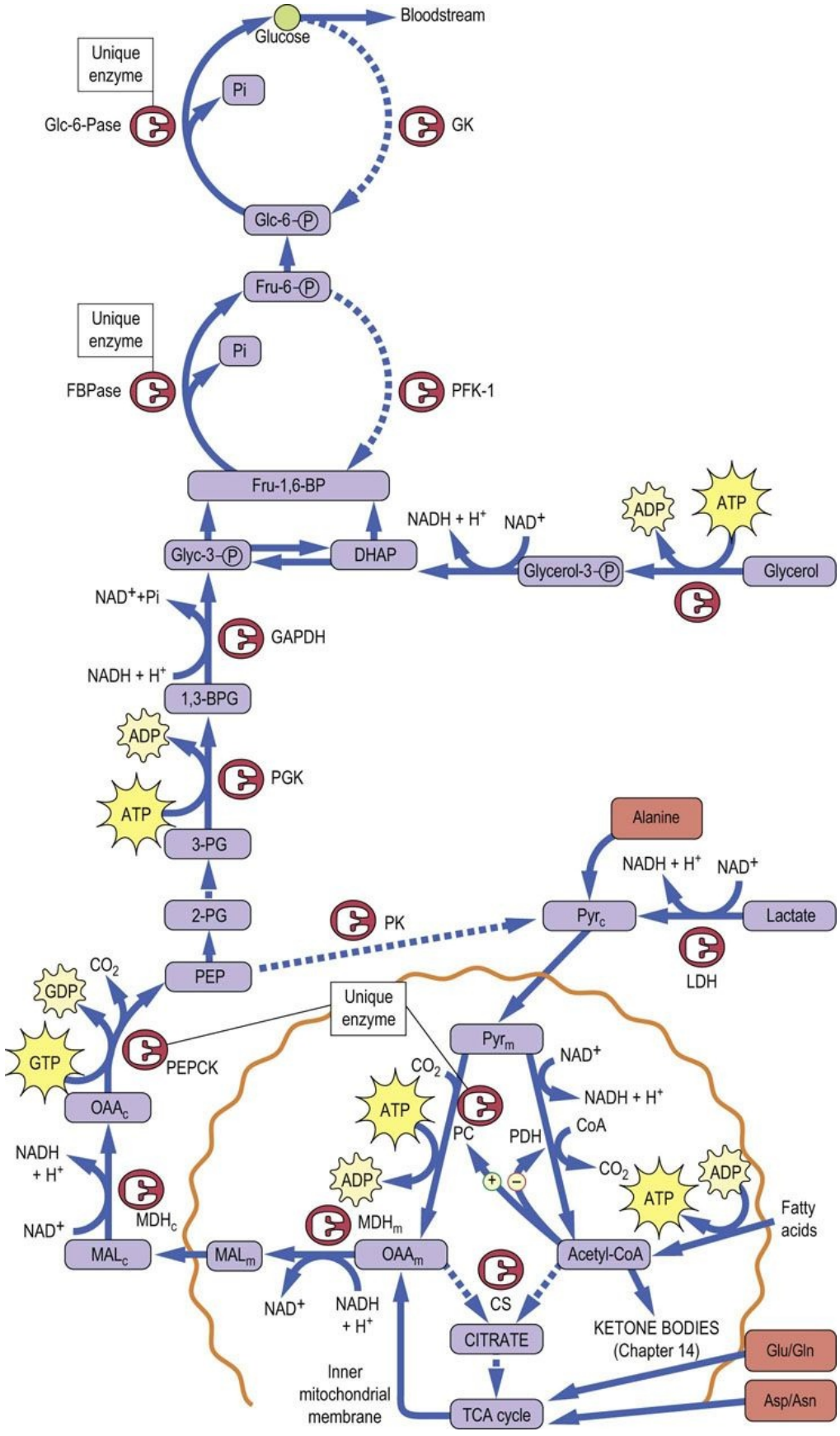

Bloodstream
Glucose
Unique enzyme
Pi
Glc-6-Pase
GK
Glc-6-Ⓟ
Fru-6-Ⓟ
Unique enzyme
Pi
FBPase
PFK-1
Fru-1,6-BP
NADH + H+
NAD+
ADP
ATP
Glyc-3-Ⓟ
DHAP
Glycerol-3-Ⓟ
Glycerol
NAD+ +Pi
GAPDH
NADH + H+
1,3-BPG
ADP
PGK
ATP
3-PG
2-PG
PK
Alanine
NADH + H+
NAD+
Pyrc
Lactate
LDH
CO2
PEP
GDP
GTP
PEPCK
Unique enzyme
Pyrm
NAD+
CO2
NADH + H+
ATP
PC
PDH
CoA
CO2
ADP
ATP
ADP
Fatty acids
NADH + H+
NAD+
MDHc
MDHm
OAAc
MALc
MALm
OAAm
Acetyl-CoA
NAD+
NADH + H+
CS
CITRATE
KETONE BODIES (Chapter 14)
Glu/Gln
Asp/Asn
Inner mitochondrial membrane
TCA cycle

FIG. 13.8 Pathway of gluconeogenesis.
Gluconeogenesis is the reverse of glycolysis. Unique enzymes overcome the irreversible kinase reactions of glycolysis. **Compartments:** c, cytoplasmic; imm, inner mitochondrial membrane; m, mitochondrial. **Enzymes:** CS, citrate synthase; Fru-1,6-BPase, fructose-1,6-bisphosphatase; GAPDH, glyceraldehyde-3-P dehydrogenase; Glc-6-Pase, glucose-6-phosphatase; GK, glucokinase; MDH, malate dehydrogenase; PC, pyruvate carboxylase; PDH, pyruvate dehydrogenase; PGK, phosphoglycerate kinase. **Substrates:** 2,3-BPG, bisphosphoglycerate; DHAP, dihydroxyacetone phosphate; Fru-1,6-BP, fructose-1,6-bisphosphate; Glyc-3-P, glyceraldehyde 3-phosphate; MAL, malate; OAA, oxaloacetate; PEP, phosphoenolpyruvate; PEPCK, PEP carboxykinase; Pyr, pyruvate; 3-PG, 3-phosphoglycerate. *Solid lines*: active during gluconeogenesis. *Dotted lines*: inactive during gluconeogenesis.

A critical problem in the reversal of glycolysis is overcoming the irreversibility of three kinase reactions: **glucokinase (GK), phosphofructokinase-1 (PFK-1), and pyruvate kinase (PK)**. The fourth kinase in glycolysis, phosphoglycerate kinase (PGK), catalyzes a freely reversible, equilibrium reaction: a substrate-level phosphorylation reaction, transferring a high-energy acyl phosphate in 1,3-bisphosphoglycerate to an energetically similar pyrophosphate bond in ATP. **To circumvent the three irreversible reactions in glycolysis, the liver uses four unique enzymes: pyruvate carboxylase (PC) in the mitochondrion and phosphoenolpyruvate carboxykinase (PEPCK) in the cytoplasm to bypass PK, fructose-1,6-bisphosphatase (Fru-1,6-BPase) to bypass PFK-1, and Glc-6-Pase to bypass GK** (see Fig. 13.8). Gluconeogenesis from lactate involves, first, its conversion into phosphoenolpyruvate (PEP), a process requiring investment of two ATP equivalents because of the high energy of the enol–phosphate bond in PEP. Lactate is first converted into pyruvate by lactate dehydrogenase (LDH), and then enters the mitochondrion, where it is converted to oxaloacetate by PC, using **biotin** and ATP. Oxaloacetate is reduced to malate by the TCA cycle enzyme, malate dehydrogenase, exits the mitochondrion, and is then reoxidized to oxaloacetate by cytosolic malate dehydrogenase. The cytosolic oxaloacetate is then decarboxylated by PEPCK, using GTP as a co-substrate, yielding PEP. The energy for synthesis of PEP from oxaloacetate is derived from both the hydrolysis of GTP and the decarboxylation of oxaloacetate.

Glycolysis may now proceed backwards from PEP until it reaches the next irreversible reaction, PFK-1. This enzyme is bypassed by a simple hydrolysis reaction, catalyzed by Fru-1,6-BPase without production of ATP, reversing the PFK-1 reaction and producing Fru-6-P. Similarly, the bypass of GK is

accomplished by hydrolysis of Glc-6-P by Glc-6-Pase, without production of ATP. The free glucose is then released into blood.

Gluconeogenesis is fairly efficient – the liver can make a kilogram of glucose per day by gluconeogenesis, and actually does so in poorly controlled, hyperglycemic diabetic patients. Normal glucose production, in the absence of dietary carbohydrate, is ~200 g/day, almost a half-pound of glucose. Gluconeogenesis from pyruvate is moderately expensive, requiring a net expenditure of the equivalent of 4 moles of ATP per mole of pyruvate converted into glucose, *i.e.* 2 mole ATP at the PC reaction and 2 mole of GTP at the PEPCK reaction. The ATP and GTP are provided by oxidation of fatty acids (Chapter 15).

Gluconeogenesis from amino acids and glycerol

Most amino acids are glucogenic (Chapter 19): *i.e.* following deamination, their carbon skeletons can be converted into glucose. **Alanine and glutamine are the major amino acids exported from muscle for gluconeogenesis.** Their relative concentrations in venous blood from muscle exceed their relative concentration in muscle protein, indicating considerable reshuffling of muscle amino acids to provide gluconeogenic substrates. As discussed in more detail in Chapter 19, alanine is converted directly into pyruvate by the enzyme alanine aminotransferase (**alanine transaminase, ALT**), and then gluconeogenesis proceeds as described for lactate. Other amino acids are converted into tricarboxylic acid cycle (TCA cycle) intermediates, then to malate for gluconeogenesis. Aspartate, for example, is converted into oxaloacetate by aspartate aminotransferase (**aspartate transaminase**, AST), and glutamate into α-ketoglutarate by glutamate dehydrogenase. Some glucogenic amino acids are converted by less direct routes into alanine or intermediates in the tricarboxylic acid cycle for gluconeogenesis. The amino groups of these amino acids are converted into urea, via the **urea cycle** in hepatocytes, and the urea is excreted in urine (Chapter 19).

Glycerol enters gluconeogenesis at the level of triose phosphates (see Fig. 13.8). Following release of glycerol and fatty acids from adipose tissue into plasma, the glycerol is taken up into liver and phosphorylated by **glycerol kinase**. Following action of glycerol-3-phosphate dehydrogenase (see Fig. 9.7), glycerol enters the gluconeogenic pathway as dihydroxyacetone phosphate. Only the glycerol component of fats can be converted into glucose. The incorporation

of glycerol into glucose requires only 2 mole of ATP per mole of glucose produced.

Clinical box child born of malnourished mother may have hypoglycemia

A baby girl was born at 39 weeks of gestation to a young, malnourished mother. The child was also thin and weak at birth and within 1 h after birth was showing signs of distress, including rapid heartbeat and respiration. Her blood glucose was 3.5 mmol/L (63 mg/dL) at birth, and declined rapidly to 1.5 mmol/L (27 mg/dL) by 1 h, when she was becoming unresponsive and comatose. Her condition was markedly improved by infusion of a glucose solution, followed by a carbohydraterich diet. She improved gradually over the next 2 weeks before discharge from the hospital.

Comment.

During development in utero, the fetus obtains glucose exogenously, from the placental circulation. However, following birth, the child relies at first on mobilization of hepatic glycogen and then on gluconeogenesis for maintenance of blood glucose. Because of the malnourished state of the mother, this child was born with negligible hepatic glycogen reserves. Thus, she was unable to maintain blood glucose homeostasis postpartum and rapidly declined into hypoglycemia, initiating a stress response. After surviving the transient hypoglycemia, she probably still lacked adequate muscle mass to provide a sufficient supply of amino acids for gluconeogenesis. Infusion of glucose, followed by a carbohydraterich diet, would address these deficits, but may not correct more serious damage from prolonged malnutrition during fetal development.

Glucose cannot be synthesized from fatty acids!

As discussed in Chapter 15, metabolism of fatty acids involves their conversion in two carbon oxidation steps to form acetyl-CoA, which is then metabolized in the tricarboxylic acid cycle following condensation with oxaloacetate to form citrate. While the carbons of acetate are theoretically available for gluconeogenesis by conversion to malate, during the pathway from citrate to malate, two molecules of CO_2 are eliminated, at the isocitrate and α-ketoglutarate dehydrogenase reactions. Thus, although energy is produced in the tricarboxylic acid cycle, the two carbons invested for gluconeogenesis from acetyl-CoA are lost as CO_2. For this reason, acetyl-CoA, and therefore even-chain fatty acids, cannot serve as substrates for *net* gluconeogenesis. However, odd-chain and branched-chain fatty acids, which form propionyl-CoA, can serve as minor precursors for gluconeogenesis. Propionyl-CoA is first carboxylated to methylmalonyl-CoA, which undergoes racemase and mutase reactions to form succinyl-CoA, a tricarboxylic acid cycle intermediate (see Chapter 15). Succinyl-CoA is converted into malate, which exits the mitochondrion and is oxidized to oxaloacetate. Following decarboxylation by PEPCK, the three carbons of propionate are conserved in PEP and glucose.

Regulation of gluconeogenesis

Fructose-2,6-bisphosphate allosterically counterregulates glycolysis and gluconeogenesis

Like glycogen metabolism in liver, gluconeogenesis is regulated primarily by hormonal mechanisms. In this case, the regulatory process involves counterregulation of glycolysis and gluconeogenesis, largely by phosphorylation/dephosphorylation of enzymes, under control of glucagon and insulin. The primary control point is at the regulatory enzymes PFK-1 and Fru-1,6-BPase which, in liver, are exquisitely sensitive to the allosteric effector **fructose 2,6-bisphosphate (Fru-2,6-BP)**. Fru-2,6-BP is an activator of PFK-1 and an inhibitor of Fru-1,6-BPase, counterregulating the two opposing pathways. As shown in Figure 13.9, Fru-2,6-BP is synthesized by an unusual, **bifunctional enzyme, phosphofructokinase-2/fructose-2,6-bisphosphatase (PFK-2/Fru-2,6-BPase)** which has both kinase and phosphatase activities. In the phosphorylated state, effected by glucagon through protein kinase A, this

enzyme displays Fru-2,6-BPase activity, which reduces the level of Fru-2,6-BP. The decrease in Fru-2,6-BP simultaneously decreases the stimulation of glycolysis at PFK-1 and relieves inhibition of gluconeogenesis at Fru-1,6-BPase. In this way, glucagon-mediated phosphorylation of PFK-2/Fru-2,6-BP places the liver cell in a gluconeogenic mode. The coordinate, allosterically mediated increase in Fru-1,6-BPase and decrease in PFK-1 activities ensure that glucose made by gluconeogenesis is not consumed by glycolysis in a futile cycle, but released into blood by Glc-6-Pase. Similarly, any flux of glucose from glycogen through glycogenolysis, also induced by glucagon, is diverted to blood, rather than to glycolysis, by inhibition of PFK-1. PK is also inhibited by phosphorylation by protein kinase A (PKA), providing an additional site for inhibition of glycolysis (Fig 13.9).

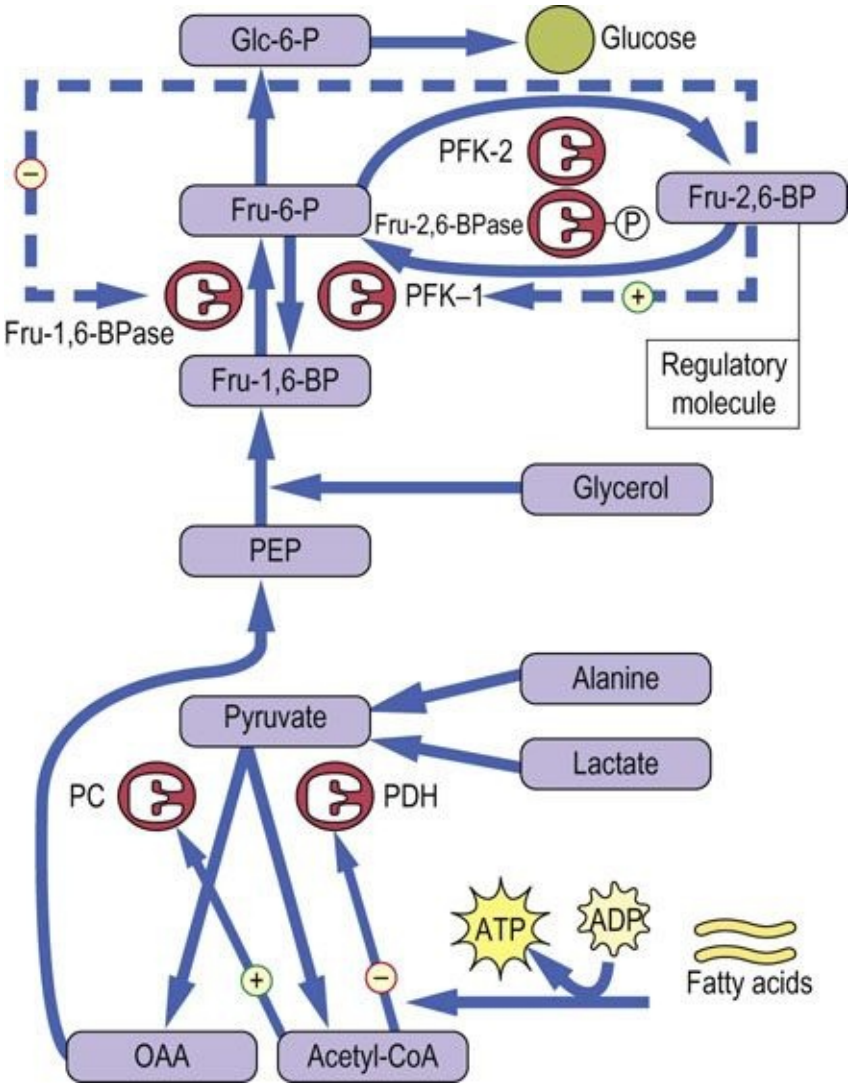


FIG. 13.9 Regulation of gluconeogenesis.
Gluconeogenesis is regulated by hepatic levels of Fru-2,6-BP and acetyl-CoA. The upper part of the diagram focuses on the reciprocal regulation of Fru-1,6-BPase and PFK-1 by Fru-2,6-BP and the lower part on the reciprocal regulation of pyruvate dehydrogenase (PDH) and pyruvate carboxylase (PC) by acetyl-CoA.

When glucose enters the liver following a meal, **insulin** mediates the dephosphorylation of PFK-2/Fru-2,6-BPase, turning on its PFK-2 activity. The resultant increase in Fru-2,6-BP activates PFK-1 and inhibits Fru-1,6-BPase activity. Gluconeogenesis is inhibited and glucose entering the liver is then incorporated into glycogen or routed into glycolysis for lipogenesis. Thus, liver metabolism following a meal is focused on synthesis and storage of both carbohydrate and lipid energy reserves, which are used later, in the postabsorptive state, for maintenance of blood glucose and fatty acid

homeostasis.

Gluconeogenesis is also regulated in the mitochondrion by acetyl-CoA. The influx of fatty acids from adipose tissue, stimulated by glucagon to support gluconeogenesis (see Chapter 15), leads to an increase in hepatic acetyl-CoA, which is both an inhibitor of pyruvate dehydrogenase (PDH) and an essential allosteric activator of pyruvate carboxylase (PC) (see Fig. 13.8). In this way, fat metabolism inhibits the oxidation of pyruvate and favors its use for gluconeogenesis in liver. In muscle during the fasting state, glucose utilization for energy metabolism is limited both by the low level of GLUT-4 in the plasma membranes (because of the low plasma insulin concentration) and by inhibition of PDH by acetyl-CoA. Active fat metabolism and high levels of acetyl-CoA in muscle promote the excretion of a significant fraction of pyruvate as lactate, even in the resting state. The carbon skeleton of glucose is returned to the liver via the Cori cycle (Chapter 21), and recycling of pyruvate into glucose, in effect, conserves muscle protein.

Conversion of fructose and galactose to glucose

As discussed in detail in Chapter 27, fructose is metabolized almost exclusively in the liver. It enters glycolysis at the level of triose phosphates, bypassing the regulatory enzyme, PFK-1. Following consumption of fruit juices, Gatorade® or foods containing high fructose corn syrup, large amounts of pyruvate may be forced on the mitochondrion for use in energy metabolism or fat biosynthesis. During a gluconeogenic state, this fructose may also proceed toward Glc-6-P, providing a convenient source of blood glucose. Gluconeogenesis from galactose is equally efficient, since Glc-1-P, derived from galactose 1-phosphate (Chapter 27), is readily isomerized to Glc-6-P by phosphoglucomutase. Fructose and galactose are good sources of glucose, independent of glycogenolysis and gluconeogenesis.

Summary

■ Glycogen is stored in two tissues in the body for different reasons: in liver for short-term maintenance of blood glucose homeostasis, and in muscle as a source of energy. Glycogen metabolism in these tissues responds rapidly to both allosteric and hormonal control.

■ In liver, the balance between glycogenolysis and glycogenesis is regulated by the balance between concentrations of glucagon and insulin in the circulation, which controls the state of phosphorylation of enzymes. Phosphorylation of enzymes under the influence of glucagon directs glycogen mobilization and is the most common condition in the liver, *e.g.* during sleep and between meals.

■ Increases in blood insulin during and after meals promote dephosphorylation of the same enzymes, leading to glycogenesis. Insulin also promotes glucose uptake into muscle and adipose tissue for glycogen and triglyceride synthesis following a meal.

■ Epinephrine increases phosphorylation of liver enzymes, enabling a burst in hepatic glycogenolysis and an increase in blood glucose for stress responses.

■ Muscle is responsive to epinephrine, but not to glucagon; in this case the glucose produced by glycogenolysis is used for muscle energy metabolism – fight or flight. In addition, muscle glycogenolysis is responsive to intracellular Ca^{2+} and AMP concentrations, providing a mechanism for coupling glycogenolysis to normal energy consumption during exercise.

■ Gluconeogenesis takes place primarily in liver, and is designed for maintenance of blood glucose during the fasting state. It is essential after 12 h of fasting, when the majority of hepatic glycogen has been consumed.

■ The major substrates for gluconeogenesis are lactate, amino acids, and glycerol; fatty acid metabolism provides the energy. The major control point is at the level of phosphofructokinase-1 (PFK-1), which is activated by the allosteric effector Fru-2,6-BP.

■ The synthesis of Fru-2,6-BP is under control of the bifunctional enzyme, PFK-2/Fru-2,6-BPase, whose kinase and phosphatase activities are regulated by phosphorylation/dephosphorylation, under hormonal control by insulin and glucagon.

■ During fasting and active gluconeogenesis, glucagon mediates phosphorylation and activation of the phosphatase activity of this enzyme, leading to a decrease in the level of Fru-2,6-BP and a corresponding decrease in glycolysis;

carbohydrate degradation is inhibited and fats become the primary energy source during fasting and starvation. Oxidation of pyruvate is also inhibited in the mitochondrion by inhibition of PDH by acetyl-CoA, derived from fat metabolism.

■ Following a meal, the decrease in phosphorylation of enzymes enhances PFK-2 activity; the increase in Fru-2,6-BP concentration activates PFK-1 and promotes glycolysis, providing pyruvate, which is converted to acetyl-CoA for lipogenesis. The actions of insulin, glucagon, and epinephrine illustrate many of the fundamental principles of hormone action (Table 13.5).

Table 13.5

General features of hormone action

Tissue specificity, determined by receptor distribution
Multistep, cascade amplification
Intracellular second messengers
Coordinate counterregulation of opposing pathways
Augmentation and/or opposition by other hormones
Multiple mechanisms of termination of response

Hormonal regulation of glucose metabolism illustrates fundamental principles of hormone action (see Chapter 39).

Hormonal regulation of glucose metabolism illustrates fundamental principles of hormone action (see Chapter 39).

Active learning

1. The inactivation of glycogenesis in response to epinephrine occurs in a single step by the action of PKA on glycogen synthase, while the activation of glycogenolysis involves an intermediate enzyme, phosphorylate kinase, which phosphorylates phosphorylase. Discuss the metabolic advantages of the two-step activation of glycogenolysis.

2. Investigate the use of inhibitors of gluconeogenesis for treatment of type 2 diabetes.
3. Glucose-6-phosphatase is essential for production of glucose in liver, but is not a cytosolic enzyme. Describe the activity and subcellular localization of this enzyme and the final stages of the pathway for production of glucose in the liver.

Further reading

Angelini, C, Semplicini, C. Enzyme replacement therapy for Pompe disease. *Curr Neurol Neurosci Rep*. 2012; 12:70–75.

Hicks, J, Wartchow, E, Mierau, G. Glycogen storage diseases: a brief review and update on clinical features, genetic abnormalities, pathologic features, and treatment. *Ultrastruct Pathol*. 2011; 35:183–196.

Johnson, RJ, Murray, R. Fructose, exercise, and health. *Curr Sports Med Rep*. 2010; 9:253–258.

Obel, LF, Müller, MS, Walls, AB, *et al.* Brain glycogen – new perspectives on its metabolic function and regulation at the subcellular level. *Front Neuroenergetics*. 2012; 4:3.

Roach, PJ, Depaoli-Roach, AA, Hurley, TD, *et al.* Glycogen and its metabolism: some new developments and old themes. *Biochem J*. 2012; 441:763–787.

Schutz, Y. Protein turnover, ureagenesis and gluconeogenesis. *Int J Vitam Nutr Res*. 2011; 81:101–107.

Stanhope, KL. Role of fructose-containing sugars in the epidemics of obesity and metabolic syndrome. *Annu Rev Med*. 2012; 63:329–343.

Stanley, CA. Hypoglycemia in the neonate. *Pediatr Endocrinol Rev*. 2006; 4(Suppl 1):76–81.

Websites

Gluconeogenesi. www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb1/part2/9-glycogen.ppt

Glycoge. www.rpi.edu/dept/bcbp/molbiochem/MBWeb/mb1/part2/10-gluconeo.ppt

CHAPTER 14

The Tricarboxylic Acid Cycle

L. William Stillway

Learning objectives

After reading this chapter you should be able to:

- Outline the sequence of reactions in the tricarboxylic acid (TCA) cycle and explain the purpose of the cycle.
- Identify the four oxidative enzymes in the TCA cycle and their products.
- Identify the two intermediates required in the first step of the TCA cycle and their metabolic sources.
- Identify four important metabolic intermediates synthesized from TCA cycle intermediates.
- Describe how the TCA cycle is regulated by substrate supply, allosteric effectors, covalent modification, and protein synthesis.
- Explain why there is no net synthesis of glucose from acetyl-CoA.
- Explain the concept of ‘suicide substrate’ as applied to the TCA cycle.

Introduction

Located in the mitochondrion, the tricarboxylic acid (TCA) cycle, also known as the Krebs or citric acid cycle, is a shared pathway for metabolism of all fuels. It oxidatively strips electrons from acetyl-CoA, which is the common product of catabolism of fat, carbohydrate and proteins, producing the majority of the reduced coenzymes that are used for the generation of adenosine triphosphate (ATP) in the electron transport chain. Although the TCA cycle does not use oxygen in any of its reactions, it requires oxidative metabolism in the mitochondrion for reoxidation of reduced coenzymes. The TCA cycle has two major functions: energy production and biosynthesis (Fig. 14.1).

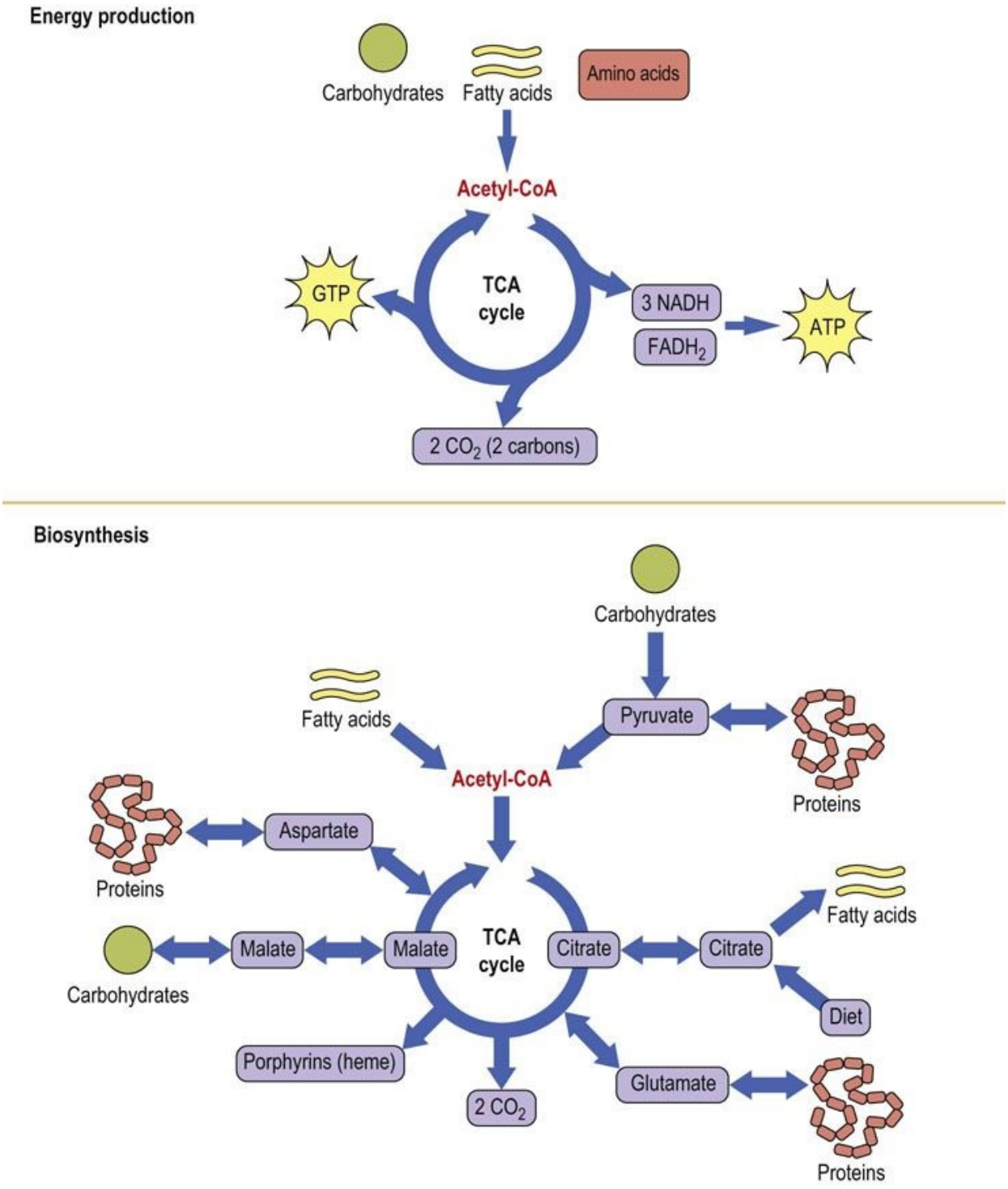


FIG. 14.1 Amphibolic nature of the TCA cycle.
The TCA cycle provides energy and metabolites for cellular metabolism. Because of the catabolic (top) and anabolic (bottom) nature of the TCA cycle, it is described as amphibolic. Acetyl-CoA is the common intermediate between metabolic fuels and the TCA cycle. $FADH_2$, reduced flavin adenine dinucleotide; GDP, guanosine diphosphate; NADH, reduced nicotinamide adenine dinucleotide.

Functions of the tricarboxylic acid cycle

Four oxidative steps provide free energy for ATP synthesis

A common end product of carbohydrate, fatty acid and amino acid metabolism, acetyl-CoA (Fig. 14.2) is oxidized in the TCA cycle to produce reduced coenzymes by four redox reactions per turn of the cycle. Three produce reduced nicotinamide adenine dinucleotide (NADH) and another produces reduced flavin adenine dinucleotide ($FADH_2$) (Fig. 9.4). These reduced nucleotides provide energy for ATP synthesis by the electron transport system (see Chapter 9). One high-energy phosphate, guanosine triphosphate (GTP), is also produced in the cycle by substrate-level phosphorylation. Nearly all metabolic carbon dioxide is produced by decarboxylation reactions catalyzed by pyruvate dehydrogenase and TCA cycle enzymes in the mitochondrion.

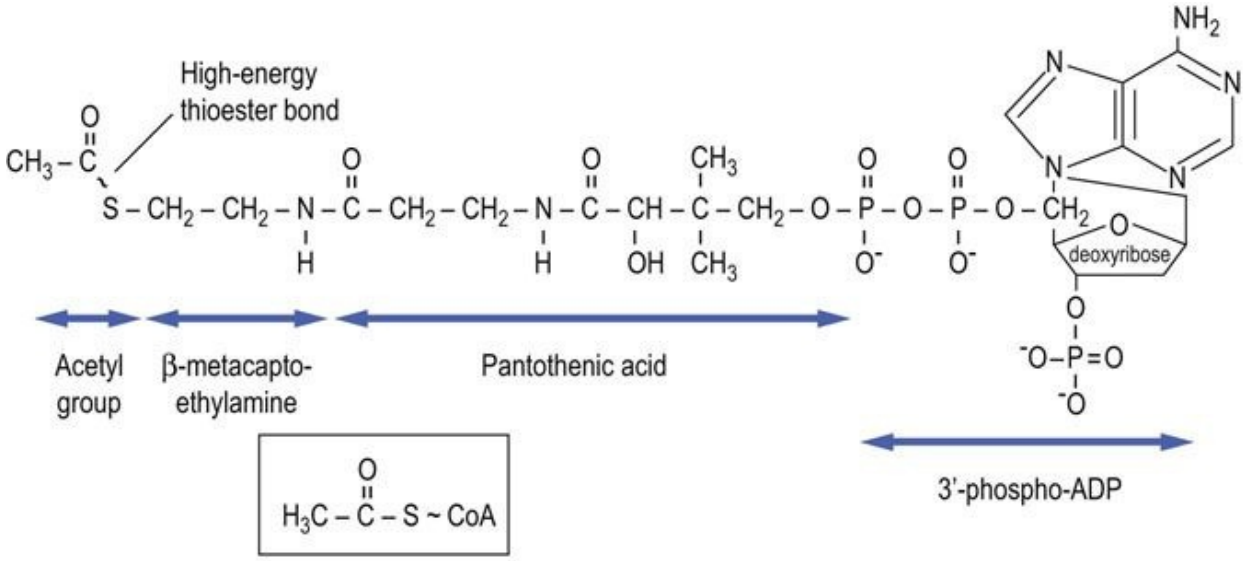


FIG. 14.2 Structure of acetyl-CoA.
Coenzyme A is an adenine nucleotide, contains a pantothenic acid moiety, and terminates in a thiol group. The acetyl group is bound to the thiol group in a high-energy thioester linkage.

The TCA cycle provides a common ground for interconversion of fuels and metabolites

In addition to its role in catabolism, the TCA cycle (see Fig. 14.1) participates in the synthesis of glucose from amino acids and lactate during starvation and fasting (gluconeogenesis; Chapter 13). It is also involved in the conversion of carbohydrates to fat following a carbohydrate-rich meal (Chapter 16). It is a source of nonessential amino acids, such as aspartate and glutamate, which are synthesized directly from TCA cycle intermediates. One TCA cycle intermediate, succinyl coenzyme A (succinyl-CoA), serves as a precursor to porphyrins (heme) in all cells, but especially in bone marrow and liver (Chapter 30). Biosynthetic reactions proceeding from the TCA cycle require the input of carbons from intermediates other than acetyl-CoA. Such reactions are known as anaplerotic (building up) reactions.

Acetyl-CoA is a common product of many catabolic pathways

The TCA cycle begins with acetyl-CoA (Fig. 14.2), which has three major metabolic precursors. Carbohydrates undergo glycolysis to yield pyruvate (Chapter 12), which can be taken up by mitochondria and oxidatively decarboxylated to acetyl-CoA by the pyruvate dehydrogenase complex. During lipolysis, triacylglycerols are converted to glycerol and free fatty acids, which are taken up by cells and transported into mitochondria, where they undergo oxidation to acetyl-CoA (Chapter 15). Lastly, proteolysis of tissue proteins releases constituent amino acids, many of which are metabolized to acetyl-CoA and TCA cycle intermediates (Chapter 19).

The first version of the TCA cycle, proposed by Krebs in 1937, began with pyruvic acid, not acetyl-CoA. Pyruvic acid was decarboxylated and condensed with oxaloacetic acid through an unknown mechanism to form citric acid. The key intermediate, acetyl-CoA, was not identified until years later. It is tempting to begin the TCA cycle with pyruvic acid, unless it is recognized that fatty acids and many amino acids form acetyl-CoA by pathways that bypass pyruvate. It is for this reason that the TCA cycle is said to begin with acetyl-CoA, not pyruvic acid.

The TCA cycle is located in the mitochondria l matrix

Localization of the TCA cycle in the mitochondrial matrix is important

metabolically; this allows identical intermediates to be used for different purposes inside and outside mitochondria. Acetyl-CoA, for example, cannot cross the inner mitochondrial membrane. The main fate of mitochondrial acetyl-CoA is oxidation in the TCA cycle but, in the cytoplasm, it is used for biosynthesis of fatty acids and cholesterol.

Metabolic defects in the TCA cycle are rare

Metabolic defects involving enzymes of the TCA cycle are rare, because functioning of the cycle is absolutely essential to sustain life. Products of energy-producing pathways must be metabolized in the TCA cycle for efficient production of ATP. Any defect in the TCA cycle will limit ATP production, and cells deprived of ATP either die rapidly or are severely impaired functionally. Tissues that use oxygen at rapid rates, such as the central nervous system and muscle, are most susceptible to such defects.

Pyruvate carboxylase

Pyruvate may be directly converted to four different metabolites

Pyruvate is at a crossroads in metabolism. It may be converted in one step to lactate (lactate dehydrogenase), to alanine (alanine aminotransferase, ALT), to oxaloacetate (pyruvate carboxylase), and to acetyl-CoA (pyruvate dehydrogenase complex) (Fig. 14.3). Depending on metabolic circumstances, pyruvate may be routed toward gluconeogenesis (Chapter 13), fatty acid biosynthesis (Chapter 16) or the TCA cycle itself. Pyruvate carboxylase, like most other carboxylases, uses CO_2 and the coenzyme biotin (Fig. 14.4), a water-soluble vitamin, and ATP to drive the carboxylation reaction. The enzyme is a tetramer of identical subunits, each of which contains an allosteric site that binds acetyl-CoA, a positive heterotropic modifier. In fact, pyruvate carboxylase has an absolute requirement for acetyl-CoA; the enzyme does not work in its absence. An abundance of mitochondrial acetyl-CoA acts as a signal for the generation of additional oxaloacetate. For example, when lipolysis is stimulated, intramitochondrial acetyl-CoA levels rise, allosterically activating pyruvate carboxylase to produce additional oxaloacetate for gluconeogenesis (Chapter 13).

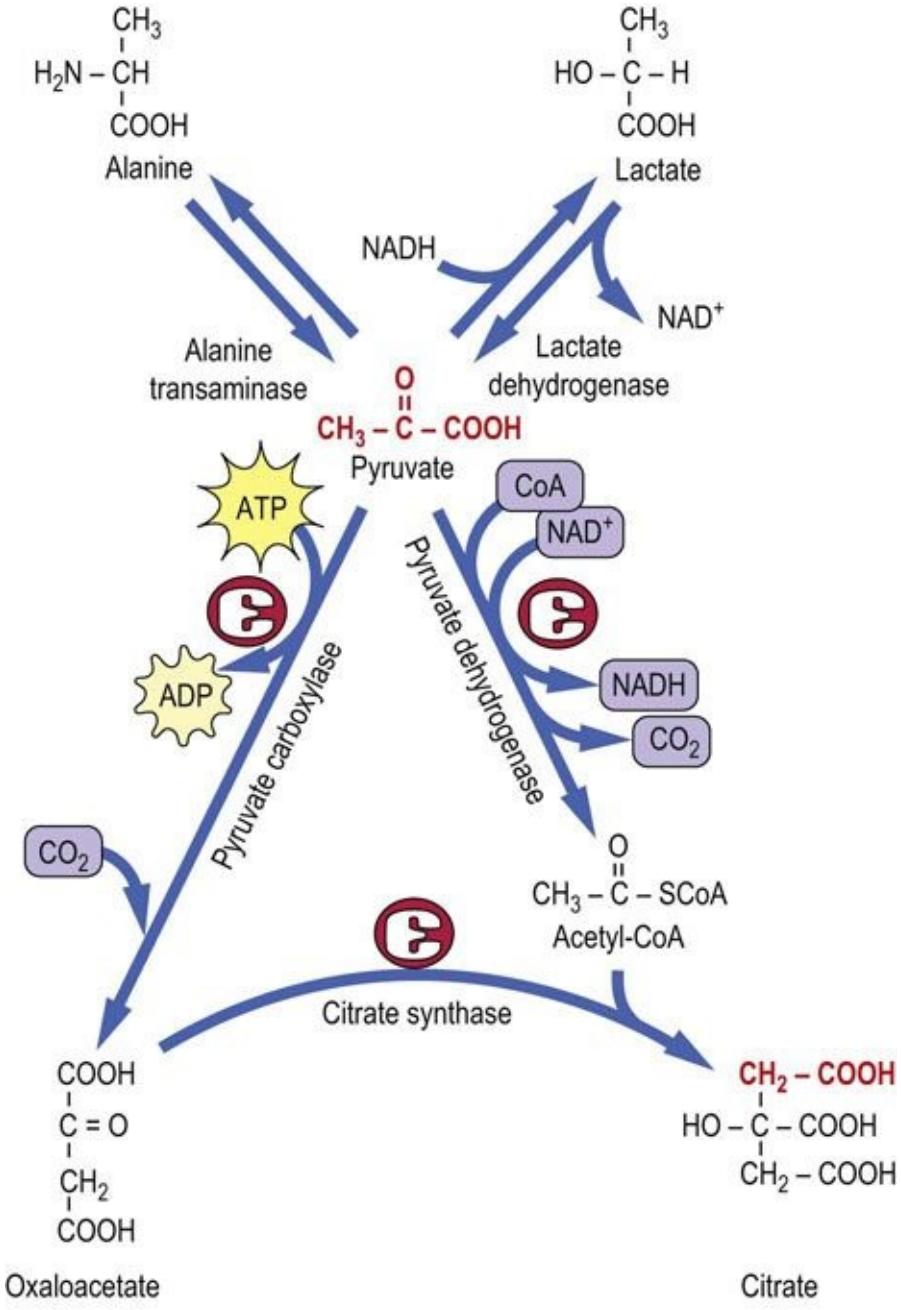


FIG. 14.3 Pyruvate is at the crossroads of metabolism.
Pyruvate is readily formed from lactate or alanine. Acetyl-CoA and oxaloacetate are derived from pyruvate through the catalytic action of pyruvate dehydrogenase and pyruvate carboxylase, respectively. ADP, adenosine diphosphate.

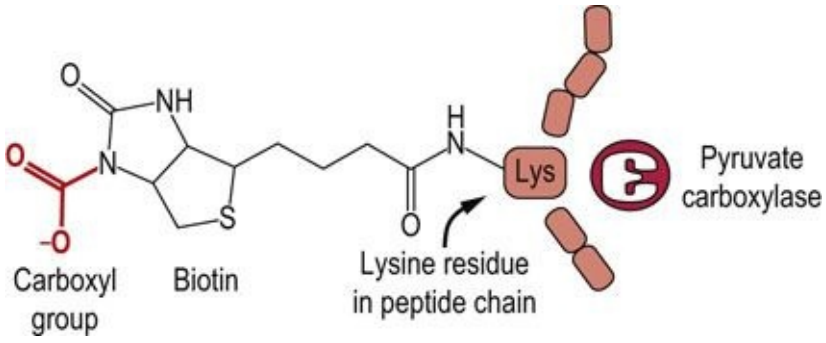


FIG. 14.4 The carboxy-biotin intermediate.
Pyruvate carboxylase catalyzes carboxylation of pyruvate to oxaloacetate. The coenzyme, biotin, is covalently bound to pyruvate carboxylase, and transfers the carbon originating from CO_2 to pyruvate (see Chapter 11).

Clinical test box Measuring lactate

Lactic acid is measured in a clinical setting, because its accumulation can result in rapid death. Lactic acid is produced metabolically by the reversible reduction of pyruvate with NADH by the enzyme lactate dehydrogenase (LDH). Both lactate and pyruvate coexist in metabolic systems, and the ratio of pyruvate : lactate is roughly proportional to the cytosolic ratio of NAD^+/NADH. Both lactate and pyruvate contribute to the acidity of biological fluid; however, lactate is usually present at higher concentrations and is more easily measured. Blood lactate may increase in chronic obstructive lung disease and during intense exercise. Its measurement is usually indicated when there is metabolic acidosis, characterized by an elevated anion gap, $[Na^+] - ([Cl^-] + [HCO_3^-])$, indicating the presence of an unknown anion(s) in plasma. Although rare, lactic acidosis can be caused by metabolic defects in energy-producing pathways, such as some of the glycogen storage diseases or in any enzyme in the pathways from pyruvate to the generation of ATP, including the pyruvate dehydrogenase complex, TCA cycle, electron transport system or ATP synthase.

The pyruvate dehydrogenase complex

The pyruvate dehydrogenase complex (PDC) serves as a bridge between carbohydrates and the TCA cycle (Fig. 14.5). PDC is one of several α-ketoacid dehydrogenases having analogous reaction mechanisms, including α-ketoglutarate dehydrogenase in the TCA cycle and α-ketoacid dehydrogenases associated with the catabolism of leucine, isoleucine and valine. Its irreversibility explains in part why acetyl-CoA cannot yield a net synthesis of glucose (below). The complex functions as a unit consisting of three principal enzymes:

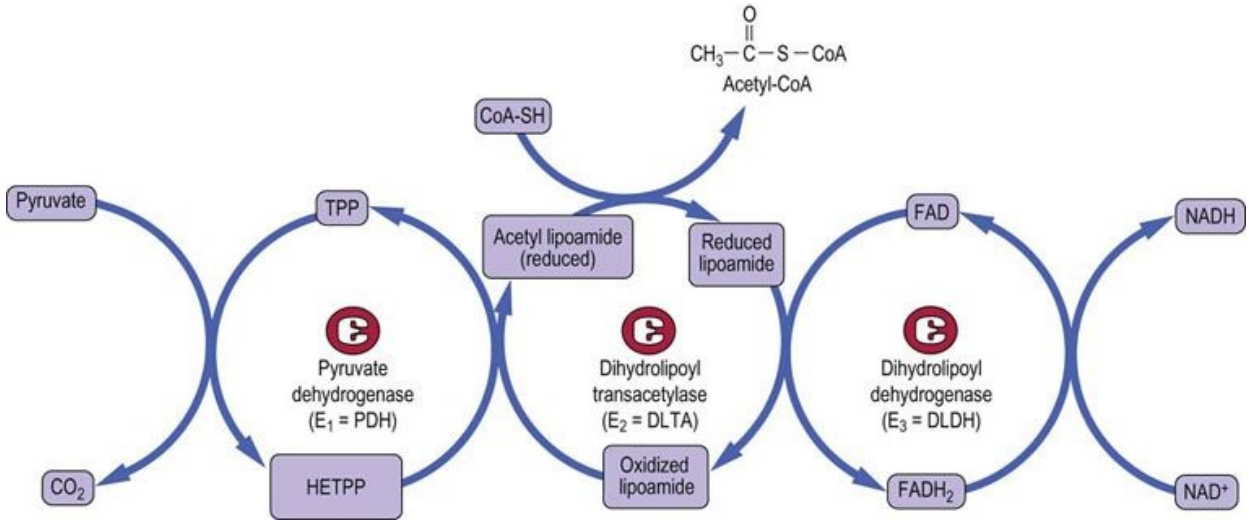


FIG. 14.5 Mechanism of action of the pyruvate dehydrogenase complex.
The three enzyme components of the pyruvate dehydrogenase complex are pyruvate dehydrogenase (E_1 = PDH), dihydrolipoyl transacetylase (E_2 = DLTA) and dihydrolipoyl dehydrogenase (E_3 = DLDH). Pyruvate is first decarboxylated by the thiamine pyrophosphate-containing enzyme (E_1), forming CO_2 and hydroxyethyl-thiamine pyrophosphate (HETPP). Lipoamide, the prosthetic group on E_2, serves as a carrier in the transfer of the 2-carbon unit from HETPP to coenzyme A (CoA). The oxidized, cyclic disulfide form of lipoamide accepts the hydroxyethyl group from HETPP. The lipoamide is reduced and the hydroxyethyl group converted to an acetyl group during this transfer reaction, forming acetyldihydrolipoamide. Following transfer of the acetyl group to CoA, E_3 reoxidizes the lipoamide, using FAD, and the $FADH_2$ is in turn oxidized by NAD^+, yielding NADH. The net reaction is: Pyr + NAD^+ + CoA-SH → acetyl-CoA + NADH + H^+ + CO_2

- pyruvate dehydrogenase
- dihydrolipoyl transacetylase
- dihydrolipoyl dehydrogenase.

Intermediates are tethered to the transacetylase component of the complex during the reaction sequence (Figs 14.5 and 14.6). This optimizes the catalytic efficiency of the enzyme since substrate does not equilibrate into solution.

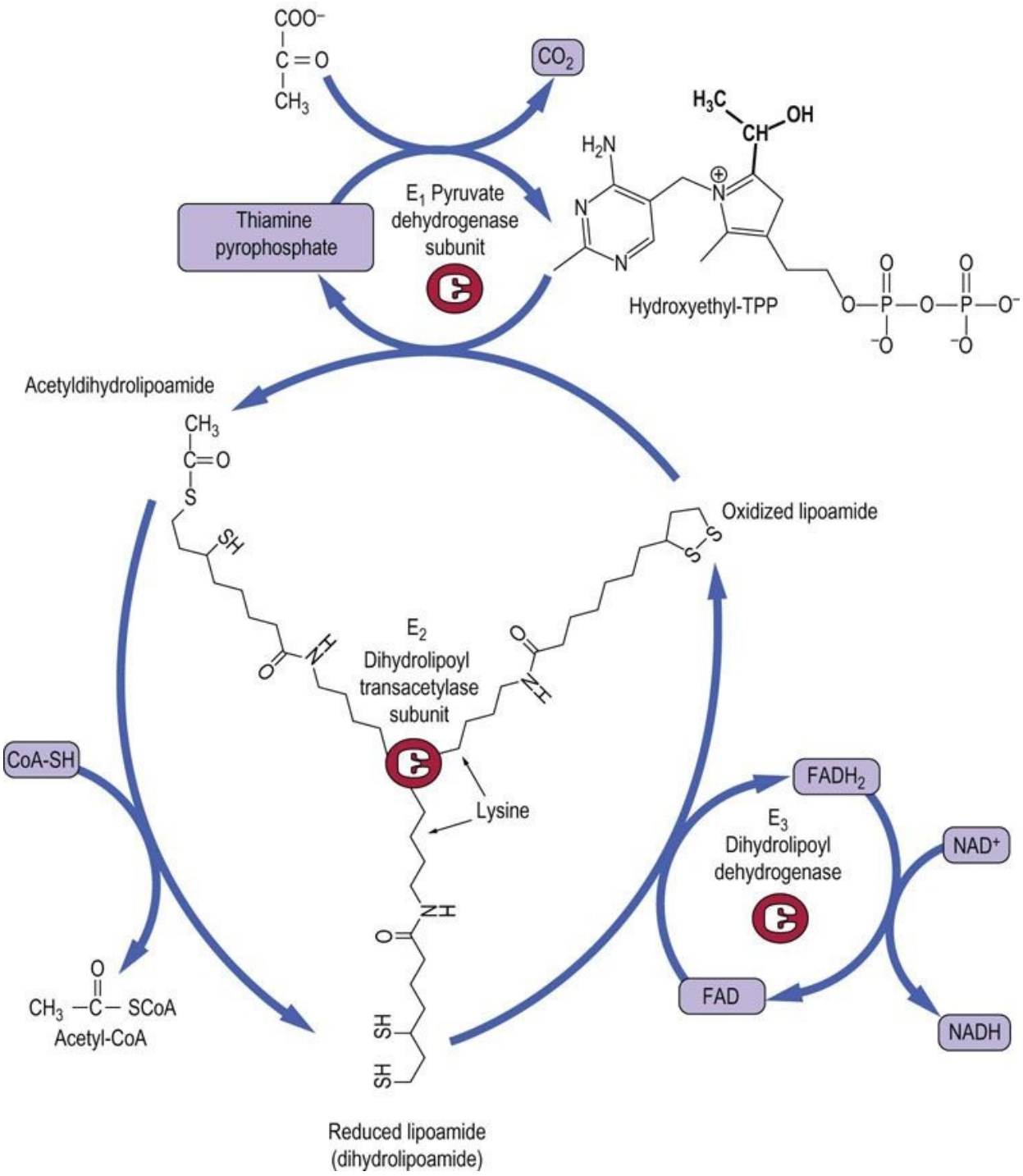


FIG. 14.6 Lipoic acid in the pyruvate dehydrogenase complex.
The coenzyme lipoamide is attached to a lysine residue in the transacetylase subunit of pyruvate dehydrogenase. Lipoamide moves from one active site to another on the transacetylase subunit in a ‘swinging arm’ mechanism. The structures of thiamine pyrophosphate (TPP) and lipoamide are shown.

Two additional enzymes of the complex, pyruvate dehydrogenase kinase and pyruvate dehydrogenase phosphatase, regulate its activity by covalent modification via reversible phosphorylation/dephosphorylation. There are four known isoforms of the kinase, and two of the phosphatase; the relative amounts of each are cell specific.

Five coenzymes are required for PDC activity: thiamine pyrophosphate, lipoamide (lipoic acid bound in amide linkage to protein), CoA, FAD, and NAD^+. Four vitamins are required for their synthesis: thiamine, pantothenic acid, riboflavin and nicotinamide. Deficiencies in any of these vitamins have obvious effects on energy metabolism. For example, increases in cellular concentrations of pyruvate and α-ketoglutarate are found in **beriberi** because of thiamine deficiency (Chapter 11). In this case, all the proteins are available but the relevant coenzyme is not, and the conversions of pyruvate to acetyl-CoA and α-ketoglutarate to succinyl-CoA are significantly decreased. Symptoms include cardiac and skeletal muscle weakness and neurologic disease. Thiamine deficiency is common in **alcoholism**, because distilled spirits are devoid of vitamins, and symptoms of beriberi are often observed.

Clinical box Pyruvate dehydrogenase complex deficiency

Most children with this enzyme deficiency present in infancy with delayed development and reduced muscle tone often associated with ataxia and seizures. Some infants have congenital malformations of the brain.

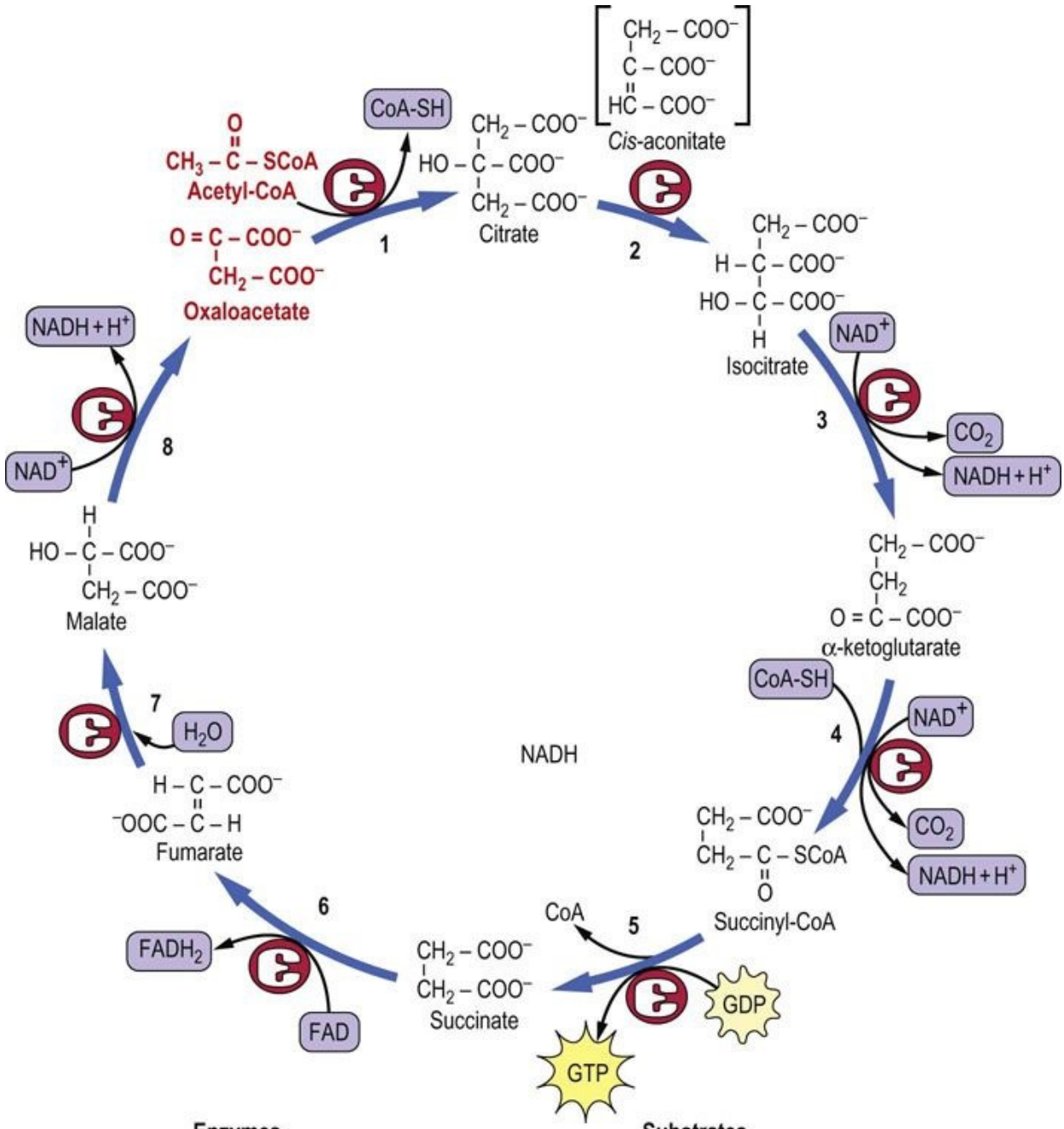
Comment.

Without mitochondrial oxidation, pyruvate is reduced to lactate. The ATP yield from anaerobic glycolysis is less than a tenth of that produced from complete oxidation of glucose via the tricarboxylic acid cycle. The diagnosis is suggested by elevated lactate, but with a normal lactate/pyruvate ratio, *i.e.* no evidence of hypoxia. A ketogenic diet and severe restriction of protein (<15%) and carbohydrate (<5%) improve mental development. Such treatment ensures that the cells use acetyl-CoA from fat metabolism. A few children show a reduction in plasma lactate on treatment with large doses of thiamine, but the outlook is generally poor.

Enzymes and reactions of the tricarboxylic acid cycle

The TCA cycle is a sequence of reactions for oxidation of acetyl-CoA to CO_2 and reduced nucleotides

The TCA cycle is a sequence of eight enzymatic reactions (Fig. 14.7), beginning with condensation of acetyl-CoA with oxaloacetate (OAA) to form citrate. The oxaloacetate is regenerated on completion of the cycle. Of the four oxidations in the cycle, two involve decarboxylations. Three dehydrogenases produce NADH and one produces $FADH_2$. GTP, a high-energy phosphate, is produced at one step by substrate-level phosphorylation.



Enzymes

(1) Citrate synthase
(2) Aconitase
(3) Isocitrate dehydrogenase
(4) α-ketoglutarate dehydrogenase
(5) Succinyl-CoA synthase (succinate thiokinase)
(6) Succinate dehydrogenase
(7) Fumarase
(8) Malate dehydrogenase

Substrates

1→2 Citrate
2→3 Isocitrate
3→4 α-ketoglutarate
4→5 Succinyl-CoA
5→6 Succinate
6→7 Fumarate
7→8 Malate
8→1 Oxaloacetate

FIG. 14.7 Enzymes and intermediates of the TCA cycle.

Citrate synthase

Citrate synthase begins the TCA cycle by catalyzing the condensation of acetyl-CoA and oxaloacetate to form citric acid. The reaction is driven by cleavage of the high-energy thioester bond of citroyl-CoA, an intermediate in the reaction. A later TCA cycle enzyme, succinyl-CoA synthetase, utilizes the high-energy thioester bond in succinyl-CoA to produce GTP, a high-energy phosphate (Fig. 14.7).

Advanced concept box Toxicity of fluoroacetate – a suicide substrate

Fluoroacetate, originally isolated from plants, is a potent toxin. It is activated as fluoroacetyl-CoA and then condenses with oxaloacetate to form fluorocitrate (Fig. 14.8). Death results from inhibition of the TCA cycle by 2-fluorocitrate, a strong inhibitor of aconitase. Fluoroacetate is an example of a 'suicide substrate', a compound that is not toxic per se but is metabolically activated to a toxic product. Thus, the cell is said to commit suicide by converting an apparently harmless substrate to a lethal toxin. Similar processes are involved in the activation of many environmental procarcinogens to carcinogens that induce mutations in DNA.

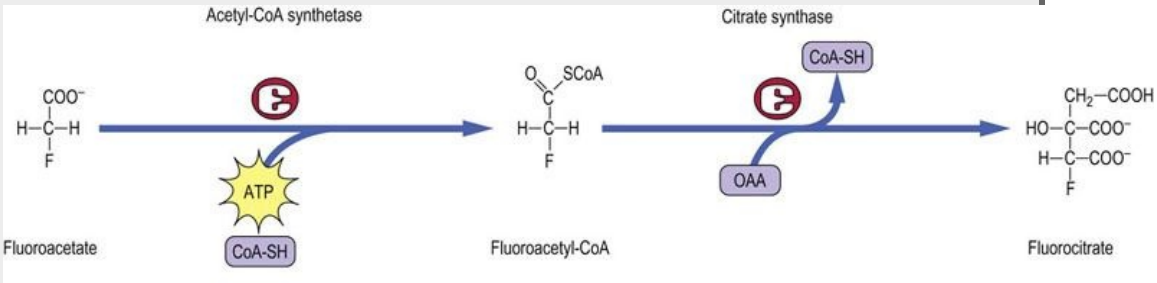


FIG. 14.8 Toxicity of fluoroacetate – a suicide substrate.
Fluoroacetate is a competitive inhibitor of aconitase. OAA, oxaloacetate.

Aconitase

Aconitase is an iron-sulfur protein (Chapter 9) that isomerizes citrate to isocitrate through the enzyme-bound intermediate *cis*-aconitate. The two-step reaction is reversible and involves dehydration followed by hydration. Although citrate is a symmetric molecule, aconitase works specifically on the oxaloacetate end of citrate, not the end derived from acetyl-CoA (Fig. 14.9). Such stereochemical specificity occurs because of the geometry of the active site of aconitase (Figs 14.10 and 14.11). A cytosolic protein with aconitase activity, known as IRE-BP (iron-response element binding protein), functions in the regulation of iron storage.

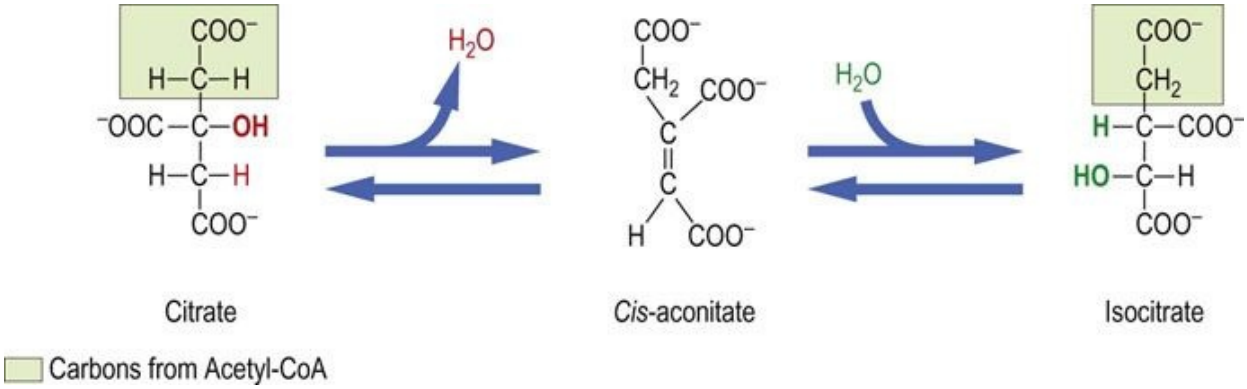


FIG. 14.9 Specificity of isomerization during the aconitase reaction.

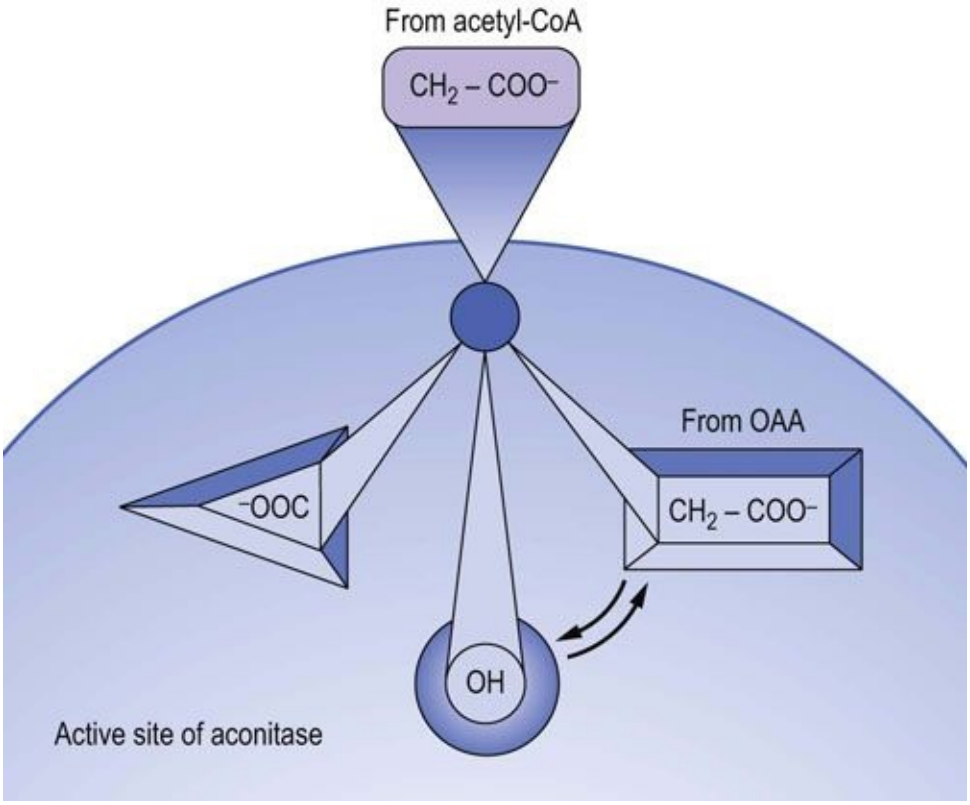


FIG. 14.10 Stereochemistry of the aconitase reaction.
Aconitase converts achiral citrate to a specific chiral form of isocitrate. Binding of the adjacent C-3 hydroxyl (OH) and carboxylate (COO^-) groups of citrate on the enzyme surface places the carboxymethyl (—CH_2—COO^-) group, derived from the oxaloacetate end of the molecule, in touch with the third binding locus in the active site of aconitase. This assures the transfer of the OH group to the CH_2 group derived from oxaloacetate, indicated by arrows, rather than that derived from the acetyl group. OAA, oxaloacetate.

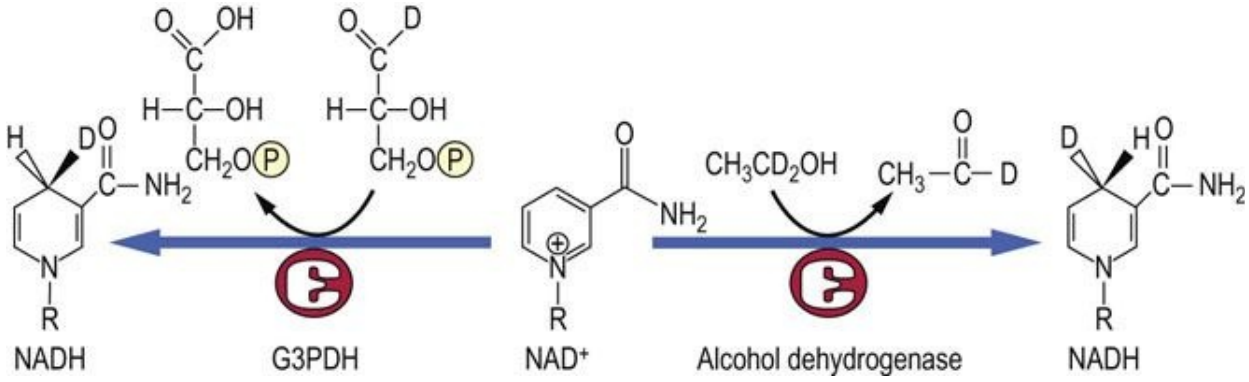


FIG. 14.11 Stereochemistry of the reduction of NAD^+ by dehydrogenases. Alcohol dehydrogenase places the hydrogen ion on the front face of the nicotinamide ring, while glyceraldehyde-3-phosphate dehydrogenase (G3PDH) places the hydrogen on the back face of the ring. The two positions can be discriminated using deuterated (D) substrates.

Advanced concept box
Stereospecificity of enzymes

Aconitase catalyzes isomerization at the oxaloacetate end of the citrate molecule. However, citrate has no asymmetric centers; it is achiral. How does aconitase know 'which end is up'? The answer lies in the nature of citrate binding to the active site of aconitase, a process known as three-point attachment. As shown in Figure 14.10, because of the geometry of the active site of aconitase, there is only one way for citrate to bind. This 'three-point binding' places the oxaloacetate carbons in the proper orientation for the isomerization reaction, while the carbons derived from acetyl-CoA are excluded from the active site.

Although citrate is a symmetric or achiral molecule, it is termed 'prochiral' because it is converted to a chiral molecule, isocitrate. Similar types of three-point binding processes are involved in transaminase reactions that produce exclusively L-amino acids from ketoacids. The reduction of the nicotinamide ring by NAD(H)-dependent dehydrogenases is also stereospecific. Some dehydrogenases place the added hydrogen exclusively on the front face of the nicotinamide ring (viewed with the amide group to the right), while others add hydrogen only to the back face (see Fig. 14.11 and Chapter 6).

Isocitrate dehydrogenase and α-ketoglutarate dehydrogenase

Isocitrate dehydrogenase and the α-ketoglutarate dehydrogenase complex catalyze two sequential oxidative decarboxylation reactions in which NAD^+ is

reduced to NADH, and CO_2 is released. The first of these enzymes, isocitrate dehydrogenase, catalyzes the conversion of isocitrate to α-ketoglutarate. It is an important regulatory enzyme that is inhibited under energy-rich conditions by high levels of NADH and ATP, and is activated when NAD^+ and ADP are produced by metabolism. Inhibition of this enzyme following a carbohydrate meal causes intramitochondrial accumulation of citrate, which is then exported to the cytosol for lipogenesis (Chapter 16). Citrate is also an allosteric effector, inhibiting phosphofructokinase-1 (Chapter 13) and activating acetyl-CoA carboxylase (Chapter 16).

The second dehydrogenase, the α-ketoglutarate dehydrogenase complex, catalyzes the oxidative decarboxylation of α-ketoglutarate to NADH, CO_2 and succinyl-CoA, a high-energy thioester compound. Like the pyruvate dehydrogenase complex, this enzyme complex contains three subunits having the same designations as pyruvate dehydrogenase (E_1, E_2 and E_3). E_3 is identical in the two complexes and is encoded by the same gene. The reaction mechanisms and the cofactors thiamine pyrophosphate, lipoate, CoA, FAD and NAD^+ are the same. Both enzymes begin with an α-keto acid, pyruvate or α-ketoglutarate, and both form the CoA esters, acetyl-CoA or succinyl-CoA, respectively.

At this point, the net carbon yield of the TCA cycle is zero, *i.e.* two carbons were introduced as acetyl-CoA and two carbons were liberated as CO_2. Note, however, that because of the asymmetry of the aconitase reaction, neither of the CO_2 molecules produced in this first round trip through the TCA cycle originates from the carbons of the acetyl-CoA, because they are derived from the oxaloacetate end of the citrate molecule. Both of the carbons that originated from acetyl-CoA remain in TCA cycle intermediates, and may appear in compounds produced in biosynthetic reactions branching from the TCA cycle, including glucose, aspartic acid and heme. However, because of the loss of two CO_2 molecules at this point, there is no net synthesis of these metabolites from acetyl-CoA.

Animals cannot perform net synthesis of glucose from acetyl-CoA. This is an especially important concept in the understanding of starvation, diabetes and ketogenesis, because large amounts of acetyl-CoA are generated from fatty acids, but it does not yield a net synthesis of glucose. ‘Net’ synthesis is invoked, because labeled carbons from acetyl-CoA eventually appear in glucose, making it appear that glucose is synthesized from acetyl-CoA. However, the investment of the two carbons of acetyl-CoA is dissipated by the two decarboxylation

reactions in the TCA cycle. Net synthesis also means that in order for continued operation of the TCA cycle for gluconeogenesis or biosynthesis of metabolites, anaplerotic reactions must supply carbons to the cycle in forms other than acetyl-CoA (see below).

Clinical box Deficiencies in pyruvate metabolism in the TCA cycle

A 7-month-old-child showed progressive neurologic deterioration characterized by loss of coordination and muscle tone. He was unable to keep his head upright and had great difficulty moving his limbs, which were limp. He also suffered from unrelenting acidosis. Administration of thiamine had no effect. Measurements showed that he had elevated blood levels of lactate, α-ketoglutarate and branched-chain amino acids. The child died a week later. Liver, brain, kidney, skeletal muscle, and heart were examined postmortem, and all gluconeogenic enzymes were shown to have normal activities, but both pyruvate dehydrogenase and α-ketoglutarate dehydrogenase were deficient. The defective component was shown to be dihydrolipoyl dehydrogenase (E_3), which is a single gene component required by all of the α-ketoacid dehydrogenases.

Comment.

This is an example of one of the many variants of **Leigh's disease**, which is a group of disorders that are all characterized by lactic acidosis. Lactic acid accumulates under anaerobic conditions or because of any enzyme defect in the pathway from pyruvate to the synthesis of ATP. In this case, there are defects in both the pyruvate dehydrogenase and α-ketoglutarate complexes, as well as other α-keto acid dehydrogenase complexes required for the catabolism of branched-chain amino acids. The failure of aerobic metabolism leads to increases in blood levels of lactate, α-ketoglutarate and branched-chain amino acids. Tissues dependent on aerobic metabolism, such as brain and muscle, are most severely affected, so that the clinical picture includes impaired motor function, neurologic disorders and mental retardation. These

diseases are rare, but deficiencies in pyruvate carboxylase and all the components of the pyruvate dehydrogenase complex (PDH) have been described, including the associated kinase and phosphatase enzymes (Fig. 14.12).

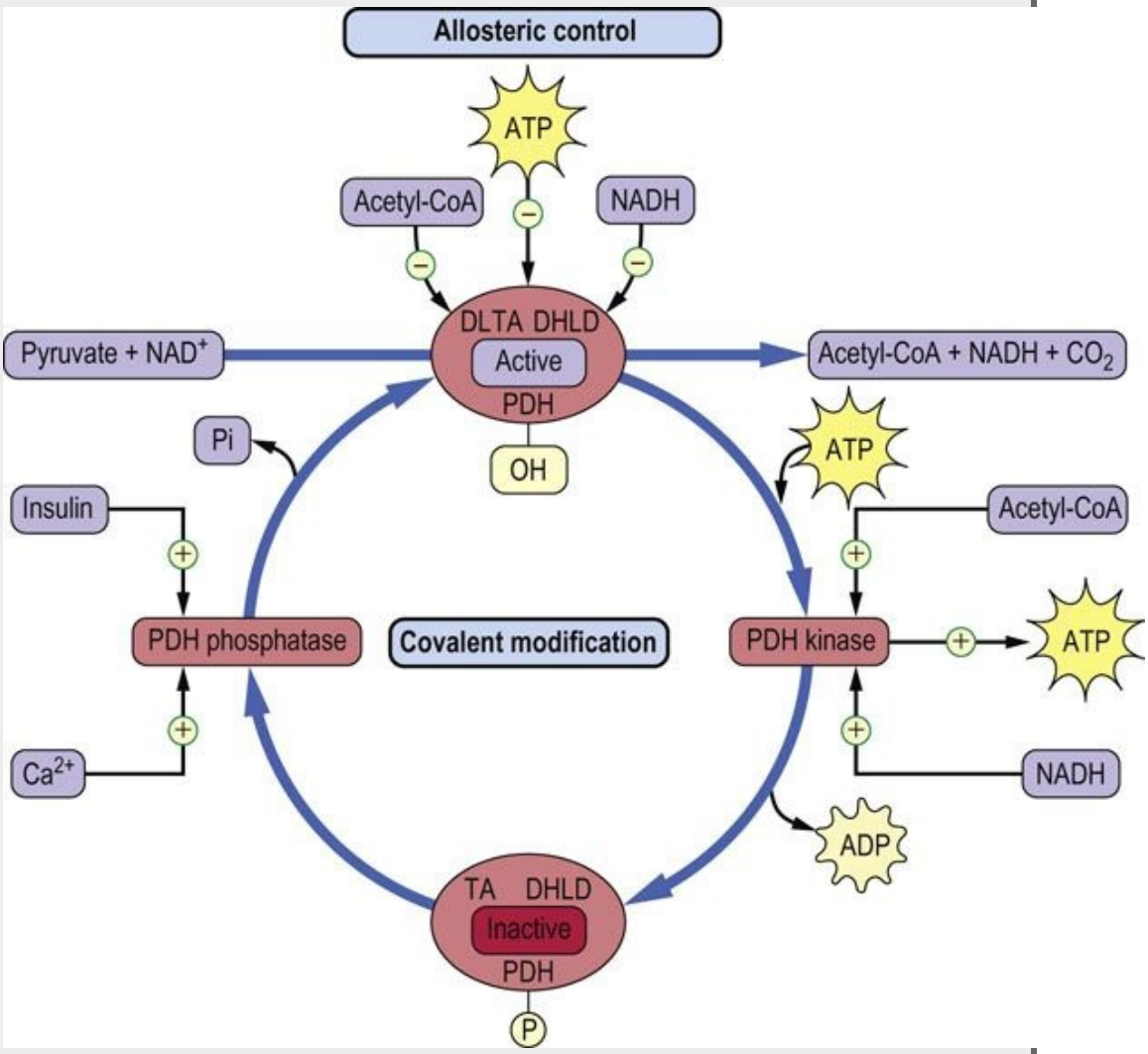


FIG. 14.12 Regulation of the pyruvate dehydrogenase complex. The pyruvate dehydrogenase complex regulates the flux of pyruvate into the TCA cycle. NAD(H), ATP and acetyl-CoA exert both allosteric and covalent control of enzyme activity. PDH, pyruvate dehydrogenase; TA, dihydrolipoyl transacetylase; DHLD, dihydrolipoamide dehydrogenase subunit.

Succinyl-CoA synthetase

Succinyl-CoA synthetase (succinate thiokinase) catalyzes the conversion of energy-rich succinyl-CoA to succinate and free CoA. The free energy of the thioester bond in succinyl-CoA is conserved by formation of GTP from GDP and inorganic phosphate (Pi). Because a high-energy thioester serves as the driving force for the synthesis of GTP, this is a substrate-level phosphorylation reaction, like the reactions catalyzed by phosphoglycerate kinase and pyruvate kinase in glycolysis (Chapter 12). GTP is used by enzymes such as phosphoenol-pyruvate carboxykinase (PEPCK) in gluconeogenesis (Chapter 13), and in several steps in protein synthesis (Chapter 34) and cell signaling (Chapter 13) but is also readily equilibrated with ATP by the enzyme nucleoside diphosphate kinase:

$$\text{GTP} + \text{ADP} \leftrightharpoons \text{GDP} + \text{ATP}$$

The next three reactions in the TCA cycle illustrate a common theme in metabolism for introducing a carbonyl group into a molecule:

- Introduction of a double bond.
- Addition of water across the double bond to form an alcohol.
- Oxidation of the alcohol to a ketone.

This same sequence occurs in the form of enzyme-bound intermediates during the oxidation of fatty acids (Chapter 15).

Succinate dehydrogenase

Succinate dehydrogenase is a flavoprotein containing the prosthetic group FAD. As described in Chapter 9, this enzyme is embedded in the inner mitochondrial membrane where it is a part of complex II (succinate-Q reductase). The reaction involves oxidation of succinate to the *trans*-dicarboxylic acid fumarate, with reduction of FAD to $FADH_2$.

Fumarase

Fumarase stereospecifically adds water across the *trans* double bond of fumarate to form the α-hydroxy acid, L-malate.

Malate dehydrogenase

Malate dehydrogenase catalyzes the oxidation of L-malate to oxaloacetate, producing NADH, completing one round trip through the TCA cycle. The oxaloacetate may then react with acetyl-CoA, continuing the cycle of reactions.

Advanced concept box The malonate block

The malate dehydrogenase reaction played an important role in the elucidation of the cyclic nature of the TCA cycle. Addition of tricarboxylic acids (citrate, aconitate) and α-ketoglutarate was known to catalyze pyruvate metabolism – we now know that this is the result of formation of catalytic amounts of oxaloacetate from these intermediates. In 1937, Krebs found that malonate, the 3-carbon dicarboxylic acid homologue of succinate and competitive inhibitor of succinate dehydrogenase, blocked metabolism of pyruvate by minced muscle preparations. He also showed that malonate inhibition of pyruvate metabolism led to accumulation not only of succinate but also of citrate and α-ketoglutarate, suggesting that succinate was a product of pyruvate metabolism and that the tricarboxylic acids might be intermediates in this process. Interestingly, fumarate and oxaloacetate also stimulated pyruvate oxidation and led to accumulation of citrate and succinate during malonate block, suggesting that the 3-and 4-carbon acids might combine to form the tricarboxylic acids. The experiments with fumarate indicated that there were two paths between fumarate and succinate, one involving reversal of the succinate dehydrogenase reaction, which was inhibited during malonate block, and the other involving conversion of fumarate to succinate through a series of organic acids. These observations, combined with Krebs' experience a few years earlier in characterization of the urea cycle (Chapter 19), led to his description of the TCA cycle.

Energy yield from the tricarboxylic acid cycle

During the course of the TCA cycle, each mole of acetyl-CoA generates sufficient reduced nucleotide coenzymes for synthesis of ~9 moles ATP by oxidative phosphorylation.

$$3\ NADH \rightarrow 7.5\ ATP$$

$$1\ FADH_2 \rightarrow 1.5\ ATP$$

Together with the GTP synthesized by substrate-level phosphorylation in the succinyl-CoA synthetase (succinate thiokinase) reaction, a total of ~10 ATP equivalents is available per mole of acetyl-CoA. Thus, complete metabolism of a mole of glucose through glycolysis, the pyruvate dehydrogenase complex and the TCA cycle yields ≈30–32 moles ATP (Table 14.1). (The actual ATP yield depends on the route of transport of redox equivalents to the mitochondrion, *i.e.* about 5 moles of ATP by the malate aspartate shuttle and about 3 moles of ATP by the glycerol phosphate shuttle (Chapter 9).) In contrast, only 2 moles of ATP (net) are recovered by anaerobic glycolysis in which glucose is converted to lactate (Chapter 12).

Table 14.1

ATP yield from glucose during oxidative metabolism

Reaction	Mechanism	Moles ATP/mole Glc
Hexokinase	Phosphorylation	–1
Phosphofructokinase	Phosphorylation	–1
G3PDH	NADH, oxidative phosphorylation	+5 (+3)*
Phosphoglycerate kinase	Substrate-level phosphorylation	+2
Pyruvate kinase	Substrate-level phosphorylation	+2
Pyruvate dehydrogenase	NADH, oxidative phosphorylation	+5
Isocitrate dehydrogenase	NADH, oxidative phosphorylation	+5
α-ketoglutarate dehydrogenase	NADH, oxidative phosphorylation	+5
Succinyl-CoA synthetase	Substrate-level phosphorylation (GTP)	+2
Succinate dehydrogenase	$FADH_2$, oxidative phosphorylation	+3
Malate dehydrogenase	NADH, oxidative phosphorylation	+5
TOTAL		**32 (30)***

The yields of ATP shown are approximate, because they are measured experimentally with live, isolated mitochondria and there is some variability. Recent work suggests that the actual yields of ATP from NADH and $FADH_2$ are about 2.5 and 1.5, respectively, yielding approximately 30–32 moles of ATP per mole of glucose. The oxidation of glucose in a bomb calorimeter yields 2870 kJ/mol (686 cal/mol), while the synthesis of ATP requires 31 kJ/mol (7.3 kcal/mol). Aerobic metabolism of glucose is therefore about 40% efficient (2870 kJ/mol glucose/31 kJ/mole ATP = 93 theoretical moles of ATP/mol glucose; 36/93 = 39%).

*Electrons from cytosolic NADH can result in the synthesis of about 5 moles of ATP per mole of glucose via the malate–aspartate shuttle, but only about 3 via the glycerol-3-phosphate shuttle per mole of glucose (Chapter 9).

Anaplerotic (‘building up’) reactions

As shown in Figure 14.1, many TCA cycle intermediates participate in biosynthetic processes, which deplete TCA cycle intermediates. For example, the synthesis of 1 mole of heme requires 8 moles of succinyl-CoA. The TCA cycle would cease to function if the intermediates were not replenished, because acetyl-CoA cannot yield a net synthesis of oxaloacetate. Anaplerotic (building up) reactions provide the TCA cycle with intermediates other than acetyl-CoA to maintain activity of the cycle. Pyruvate carboxylase is a prime example of an enzyme that catalyzes an anaplerotic reaction. It converts pyruvate to oxaloacetate, which is required for initiation of the cycle. Malic enzyme in the cytoplasm also converts pyruvate to malate, which can enter the mitochondrion as a substrate for the TCA cycle. Aspartate is also a precursor of oxaloacetate by a transamination reaction, and α-ketoglutarate can be produced through an aminotransferase reaction from glutamate, as well as by the glutamate dehydrogenase reaction. Several other ‘glucogenic’ amino acids (Chapter 19) may also serve as sources of pyruvate or TCA cycle intermediates, guaranteeing that the cycle never stalls because of a lack of intermediates.

Regulation of the tricarboxylic acid cycle

Pyruvate dehydrogenase and isocitrate dehydrogenase regulate TCA cycle activity

There are several levels of control of the TCA cycle. In general, the overall activity of the cycle depends on the availability of NAD^+ for the dehydrogenase reactions. This, in turn, is linked to the rate of NADH consumption by the electron transport system, which ultimately depends on the rate of ATP utilization and production of ADP by metabolism (see Table 14.1). Thus, as ATP is used for metabolic work, ADP is produced, then NADH is consumed by the electron transport system for ATP production, and NAD^+ is produced. The TCA cycle is activated, fuels are consumed, and more NADH is produced so that more ATP may be made. The mitochondrial level of NAD^+ provides a link between work (ATP utilization) and fuel consumption (Chapter 9).

There are several regulatory enzymes that affect the activity of the TCA cycle. The activity of the pyruvate dehydrogenase complex, and therefore the supply of acetyl-CoA from glucose, lactate and alanine, is regulated by allosteric and covalent modifications (see Fig. 14.12). The products of the pyruvate dehydrogenase reaction, NADH and acetyl-CoA, as well as ATP, act as negative allosteric effectors of the enzyme complex. In addition, the pyruvate dehydrogenase complex has associated kinase and phosphatase enzymes that modulate the degree of phosphorylation of regulatory serine residues in the complex. NADH, acetyl-CoA and ATP activate the kinase, which phosphorylates and inactivates the enzyme complex. In contrast, when these three compounds are low in concentration, the enzyme complex is activated allosterically and by dephosphorylation by the phosphatase. This is an important regulatory process during fasting and starvation, when gluconeogenesis is essential to maintain blood glucose concentration. Active fat metabolism during fasting leads to increased NADH and acetyl-CoA in the mitochondrion, which leads to inhibition of pyruvate dehydrogenase and blocks the utilization of carbohydrate for energy metabolism in the liver. Under this condition, pyruvate, from such intermediates as lactate and alanine, is directed toward gluconeogenesis. Conversely, insulin stimulates pyruvate dehydrogenase by activating the phosphatase in response to dietary carbohydrates. This directs carbohydrate-derived carbons into fatty acids (lipogenesis) via citrate synthase

(Chapter 16). Ca^{2+} also affects PDC phosphatase activity, in response to the increase in intracellular Ca^{2+} during muscle contraction (Chapter 20).

Oxaloacetate is required for entry of acetyl-CoA into the TCA cycle but, at times, the availability of oxaloacetate appears to regulate the activity of the cycle. This occurs especially during fasting when levels of ATP and NADH, derived from fat metabolism, are increased in the mitochondrion. The increase in NADH shifts the malate : oxaloacetate equilibrium toward malate, directing TCA cycle intermediates toward malate, which is exported to the cytosol for gluconeogenesis (Chapter 13). Meanwhile, acetyl-CoA derived from fat metabolism is directed toward synthesis of ketone bodies because of the lack of oxaloacetate, regenerating CoA-SH and leading to the increase in ketone bodies in plasma during fasting (Chapter 15).

Isocitrate dehydrogenase is a major regulatory enzyme within the TCA cycle. It is subject to allosteric inhibition by ATP and NADH and stimulation by ADP and NAD^+. During consumption of a high carbohydrate diet under resting conditions, the demand for ATP is diminished and the level of carbohydrate-derived intermediates increases. Under these circumstances, increased insulin levels stimulate the pyruvate dehydrogenase complex, and the accumulation of ATP and NADH inhibits isocitrate dehydrogenase, causing a mitochondrial accumulation of citrate. The citrate is then exported to the cytosol for synthesis of fatty acids, which are exported from the liver for storage in adipose tissue as triglycerides. With an increase in energy demand, *e.g.* during muscle contraction, NAD^+ and ADP accumulate, and they stimulate isocitrate dehydrogenase.

Induction and repression, as well as proteolysis of enzyme proteins, such as pyruvate carboxylase and those in the pyruvate dehydrogenase complex and the TCA cycle, also play an important regulatory role. In fact, all of the TCA cycle and associated enzymes are synthesized in the cytoplasm and transported through a complex series of steps into the mitochondrion. Regulation can occur at the level of translation, transcription and intracellular transport. Diet, for example, is known to control expression of four pyruvate dehydrogenase kinases; one of them is induced in response to a high-fat diet and is repressed in response to a high-carbohydrate diet. Unfortunately, the regulation of the TCA cycle at genetic and transport levels is not as well understood, although it is clearly important for understanding the pathogenesis of a wide range of contemporary health problems, such as diabetes and obesity. Three TCA cycle enzymes, succinate dehydrogenase, fumarate hydratase and isocitrate

dehydrogenase, are described as tumor suppressors because genetic defects in these enzymes are associated with human cancers (see Cardaci and Ciriolo and Yang *et al.* in Further Reading).

Summary

■ The TCA cycle is the central, common pathway by which fuels are oxidized, and it also participates in major biosynthetic pathways.
■ In its oxidative role, major products of the TCA cycle are GTP and the reduced coenzymes NADH and $FADH_2$, which furnish large amounts of free energy for the synthesis of ATP by oxidative phosphorylation.
■ In its biosynthetic role, the TCA cycle provides essential intermediates for the synthesis of glucose, fatty acids, amino acids and heme, as well as the ATP required for their biosynthesis.
■ The activity of the TCA cycle is tightly regulated by substrate supply, by allosteric effectors and control of gene expression so that fuel consumption is coordinated with energy production.

Active learning

1. In beriberi, the vitamin thiamine is deficient. Which intermediates would accumulate, and why?
2. Based on rates of oxygen consumption, which tissues would be the most critically impaired because of genetically defective enzymes of the TCA cycle?
3. Compare the regulation of the pyruvate dehydrogenase complex to the regulation of cytosolic enzymes by phosphorylation/dephosphorylation reactions.
4. Predict the consequences of deficiencies in TCA cycle enzymes such as succinate dehydrogenase, fumarase or malate dehydrogenase.

Further reading

Cardaci, S, Ciriolo, MR. TCA cycle defects and cancer: When metabolism tunes redox state. *Int J Cell Biol*. 2012; 2012:161837.

Marin-Valencia, I, Roe, CR, Pascual, JM. Pyruvate carboxylase deficiency: mechanisms, mimics and anaplerosis. *Mol Genet Metab*. 2010; 101:9–17.

Patel, KP, O'Brien, TW, Subramony, SH, *et al*. The spectrum of pyruvate dehydrogenase complex deficiency: clinical, biochemical and genetic features in 371 patients. *Mol Genet Metab*. 2012; 106:385–

394.

Qi, F, Pradhan, RK, Dash, RK, *et al.* Detailed kinetics and regulation of mammalian 2-oxoglutarate dehydrogenase. *BMC Biochem*. 2011; 12:53.

Sudheesh, NP, Ajith, TA, Janardhanan, KK, *et al.* Palladium alpha-lipoic acid complex formulation enhances activities of Krebs cycle dehydrogenases and respiratory complexes I–IV in the heart of aged rats. *Food Chem Toxicol*. 2009; 47:2124–2128.

Yang, M, Soga, T, Pollard, PJ, *et al.* The emerging role of fumarate as an oncometabolite. *Front Oncol*. 2012; 2:85.

Websites

www.johnkyrk.com/krebs.html

www.wiley.com/legacy/college/boyer/0470003790/animations/tca/tca.htm

www.npr.org/blogs/krulwich/2011/09/14/140428189/lord-save-me-from-the-krebs-cycle

CHAPTER 15

Oxidative Metabolism of Lipids in Liver and Muscle

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

■ Describe the pathway for activation and transport of fatty acids to the mitochondrion for catabolism.

■ Outline the sequence of reactions involved in oxidation of fatty acids in the mitochondrion.

■ Describe the general features of pathways for oxidation of unsaturated, odd-chain and branched-chain fatty acids.

■ Explain the rationale for the pathway of ketogenesis and identify the major intermediates and products of this pathway.

■ Describe the mechanism by which hormonal activation of lipolysis in adipose tissue is coordinated with activation of gluconeogenesis in liver during fasting.

Introduction

Fats are normally the major source of energy in liver and muscle, and in other tissues, with two exceptions: brain and red cells

Triglycerides are the storage and transport form of fats; fatty acids are the immediate source of energy. Fatty acids are released from triglyceride stores in adipose tissue, transported in plasma in association with albumin, and delivered to cells for metabolism. The catabolism of fatty acids is entirely oxidative; after they have been transported through the cytoplasm, their oxidation proceeds in both the peroxisome and the mitochondrion, primarily by a cycle of reactions known as **β-oxidation**. Carbons are released, two at a time, from the carboxyl end of the fatty acid; the major end products are acetyl coenzyme A (acetyl-CoA) and the reduced forms of the nucleotides $FADH_2$ and NADH. In muscle, the acetyl-CoA is metabolized via the tricarboxylic acid cycle and oxidative phosphorylation to produce ATP. In liver, acetyl-CoA is converted largely to ketone bodies (**ketogenesis**), which are water-soluble lipid derivatives that, like glucose, are exported for use in other tissues. Fat metabolism is controlled primarily by the rate of triglyceride hydrolysis (lipolysis) in adipose tissue, which is regulated by hormonal mechanisms involving insulin and **glucagon, epinephrine, and cortisol.** These hormones coordinate the metabolism of carbohydrate, lipid and protein throughout the body (see Chapter 21).

Activation of fatty acids for transport into the mitochondrion

Fatty acids are activated by formation of a high energy thioester bond with Coenzyme A

Fatty acids do not exist to a significant extent in free form in the body – salts of fatty acids are soaps; they would dissolve cell membranes. In blood, fatty acids are bound to albumin, which is present at ~0.5 mmol/L concentration (35 mg/mL) in plasma. Each molecule of albumin can bind 6–8 fatty acid molecules. In the cytosol, fatty acids are bound to a series of fatty acid-binding proteins and enzymes. As the priming step for their catabolism, the fatty acids are activated to their CoA derivative, using ATP as the energy source (Fig. 15.1). The carboxyl group is first activated to an enzyme-bound, high-energy acyl-adenylate intermediate, formed by reaction of the carboxyl group of the fatty acid with ATP. The acyl group is then transferred to CoA by the same enzyme, **fatty acyl-CoA synthetase**. This enzyme is commonly known as fatty acid **thiokinase**, because ATP is consumed in the formation of the thioester bond in acyl-CoA.

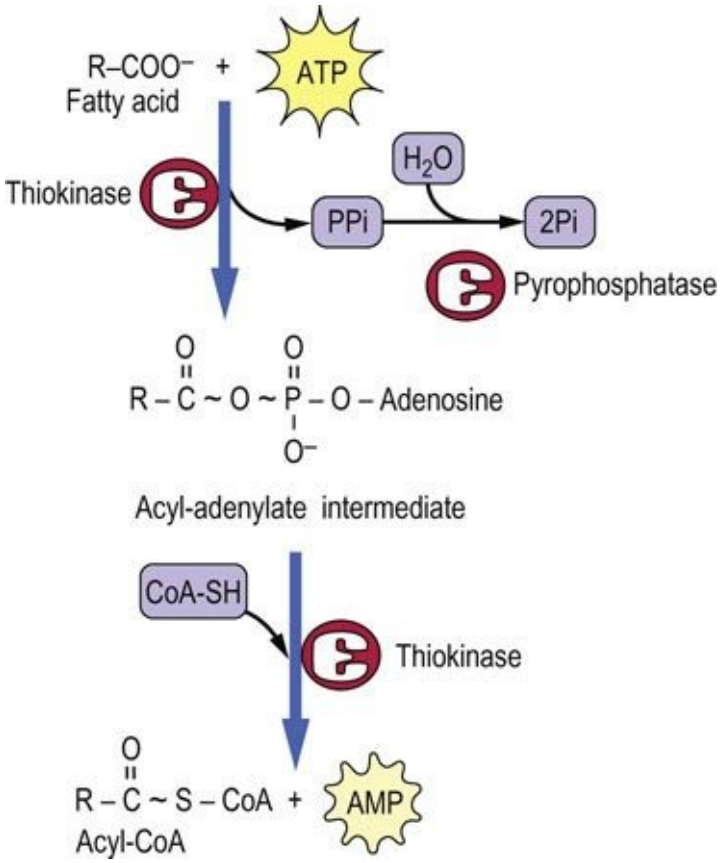


FIG. 15.1 Activation of fatty acids by fatty acyl-CoA synthetase (thiokinase). ATP forms an enzyme-bound acyl-adenylate intermediate, which is discharged by CoA-SH to form acyl-CoA. AMP, adenosine monophosphate; CoA-SH, coenzyme A; PPi, inorganic pyrophosphate.

The length of the fatty acid dictates where it is activated to CoA

Short-and medium-chain fatty acids (Table 15.1) can cross the mitochondrial membrane by passive diffusion, and are activated to their CoA derivative within the mitochondrion. Very long-chain fatty acids from the diet are shortened to long-chain fatty acids in peroxisomes. Long-chain fatty acids are the major components of storage triglycerides and dietary fats. They are activated to their CoA derivatives in the cytoplasm and are transported into the mitochondrion via the carnitine shuttle.

Table 15.1

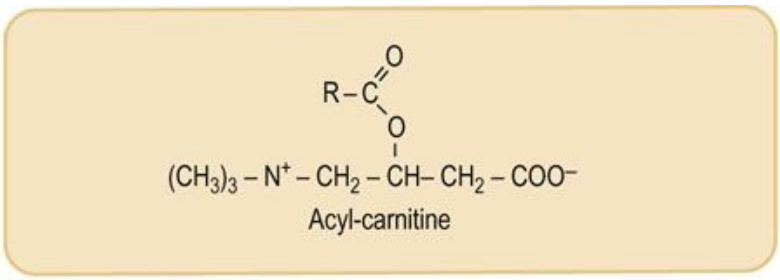
Metabolism of the four classes of fatty acids.

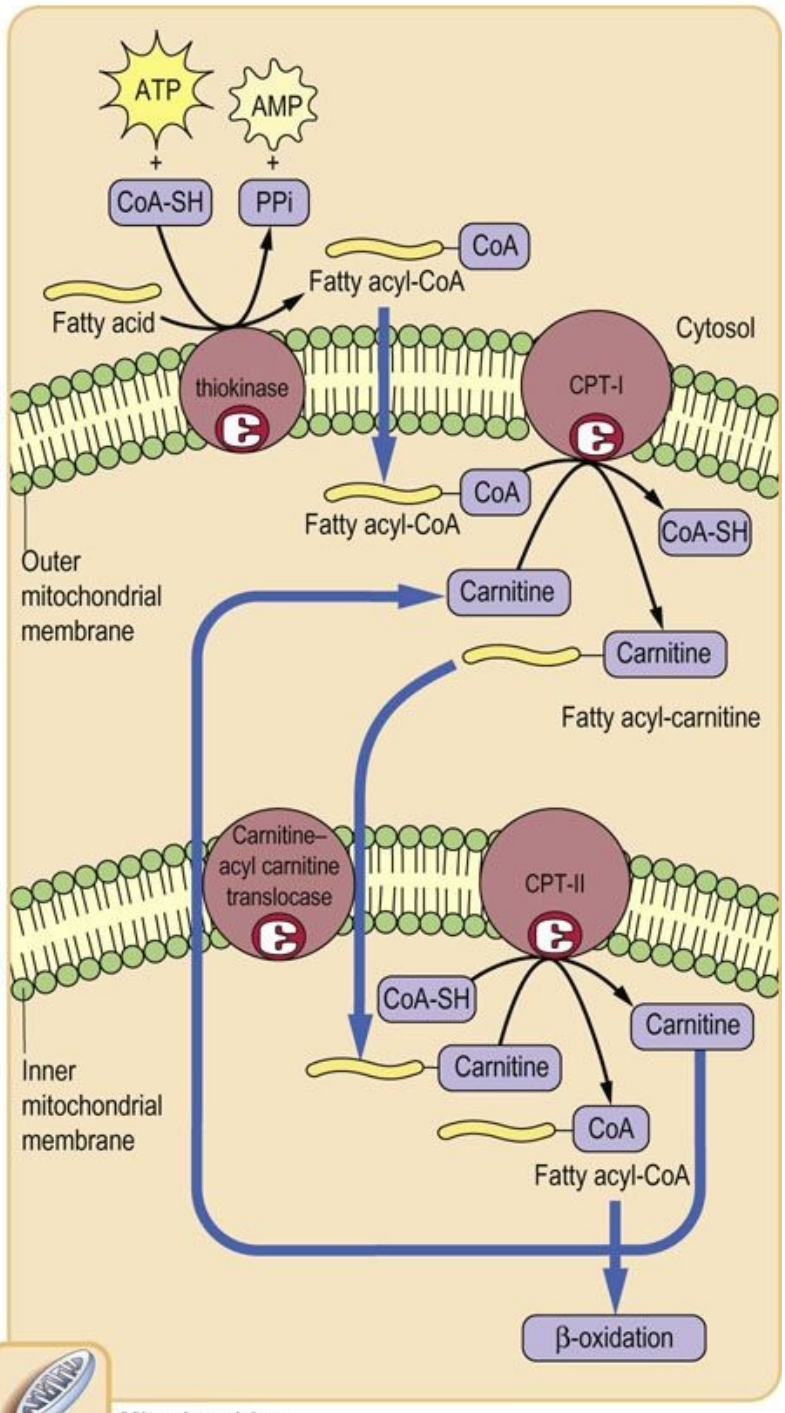
Size class	Number of carbons	Site of catabolism	Membrane transport
Short-chain	2–4	Mitochondrion	Diffusion
Medium-chain	4–12	Mitochondrion	Diffusion
Long-chain	12–20	Mitochondrion	Carnitine cycle
Very long-chain	>20	Peroxisome	Unknown

The carnitine shuttle

The carnitine shuttle bypasses the impermeability of the mitochondrial membrane to Coenzyme A

CoA is a large, polar, nucleotide derivative (Fig. 14.2), and cannot penetrate the mitochondrial inner membrane. Thus, for the transport of long-chain fatty acids, the fatty acid is first transferred to the small molecule, carnitine, by carnitine palmitoyl transferase-I (**CPT-I**), located in the outer mitochondrial membrane. An **acyl-carnitine transporter** or translocase in the inner mitochondrial membrane mediates transfer of the acyl-carnitine into the mitochondrion, where **CPT-II** regenerates the acyl-CoA, releasing free carnitine. The carnitine shuttle (Fig. 15.2) operates by an antiport mechanism in which free carnitine and the acyl-carnitine derivative move in opposite directions across the inner mitochondrial membrane. The shuttle is an important site in the regulation of fatty acid oxidation. As discussed in the next chapter, the carnitine shuttle is inhibited by **malonyl-CoA** after the ingestion of carbohydrate-rich meals. Malonyl-CoA prevents the futile cycle in which newly synthesized fatty acids would be oxidized in the mitochondrion.





Mitochondrion

FIG. 15.2 Transport of long-chain fatty acids into the mitochondrion.
The three components of the carnitine pathway include carnitine palmitoyl transferases (CPTs) in the outer and inner mitochondrial membranes and the carnitine-acyl carnitine translocase.

Oxidation of fatty acids

Mitochondrial β-oxidation

Oxidation of the β-carbon (C-3) facilitates sequential cleavage of acetyl units from the carboxyl end of fatty acids

Fatty acyl-CoAs are oxidized in a cycle of reactions involving oxidation of the β-carbon to a ketone: hence the term β-oxidation (Figs 15.3 and 15.4). The oxidation is followed by cleavage between the α-and β-carbons by a thiolase, rather than hydrolase, reaction; in this way the high energy of the thioester bond is preserved to provide the thermodynamic driving force for subsequent reactions. One mole each of acetyl-CoA, $FADH_2$ and NADH is formed during each cycle, along with a fatty acyl-CoA with two fewer carbon atoms. For a 16-carbon fatty acid, such as palmitate, the cycle is repeated seven times, yielding 8 moles of acetyl-CoA (see Fig. 15.3), plus 7 moles of $FADH_2$ and 7 moles of NADH + H^+. This process occurs in the mitochondrion, and the reduced nucleotides are used directly for synthesis of ATP by oxidative phosphorylation (Table 15.2).

Table 15.2

Comparative energy yield from glucose and palmitate.

Substrate	Molecular weight	Net ATP yield (mol/mol)	ATP (mol/g)	Caloric value Cal/g (kJ)
Glucose	180	36–38	0.2	4 (17)
Palmitate	256	129	0.5	9 (37)

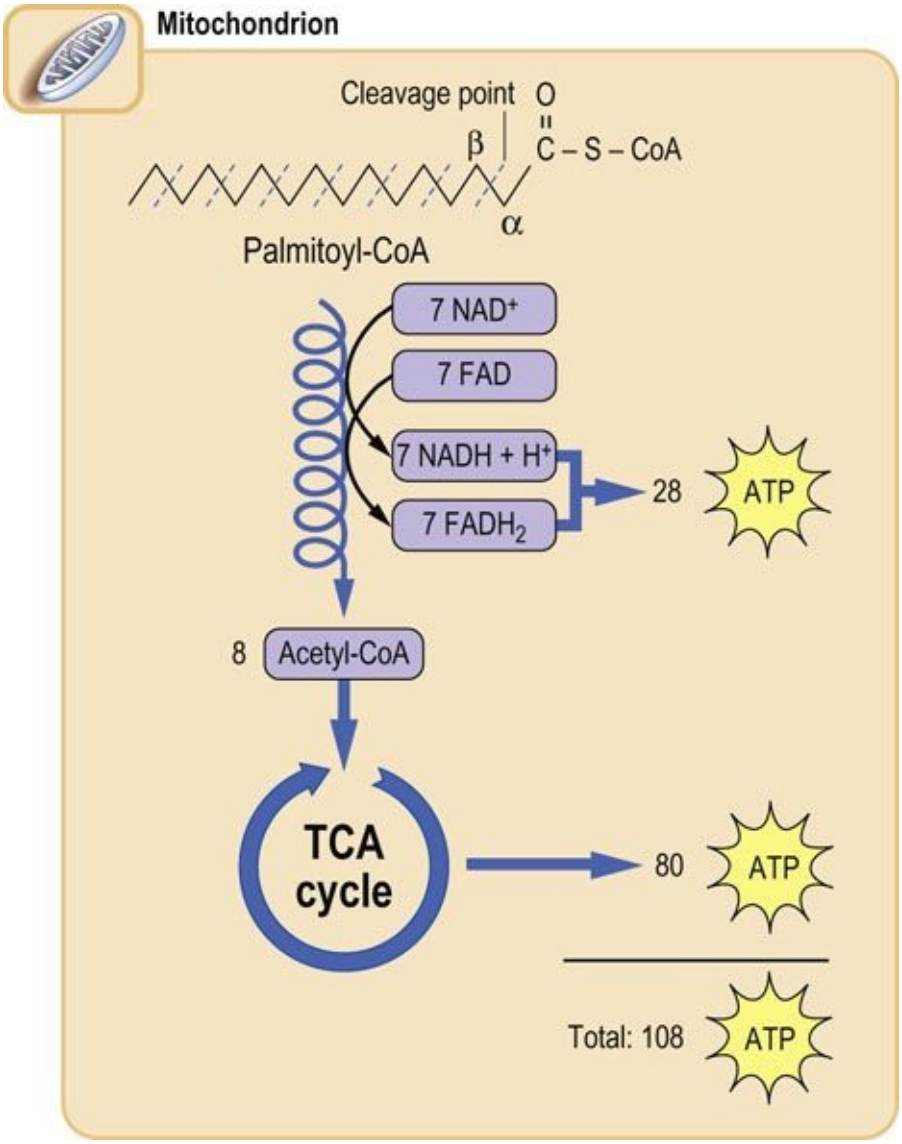


FIG. 15.3 Overview of β-oxidation of palmitate.
In a cycle of reactions, the carbons of the fatty acyl-CoA are released in two-carbon acetyl-CoA units; the yield of 28 ATP from this β-oxidation is nearly equivalent to that from complete oxidation of glucose. In liver, the acetyl-CoA units are then used for synthesis of ketone bodies, and in other tissues they are metabolized in the TCA cycle to form ATP. The complete oxidation of palmitate yields a net 106 moles of ATP, after correction for the 2-mole equivalents of ATP invested at the thiokinase reaction. The overall production of ATP per gram of palmitate is about twice that per gram of glucose, because glucose is already partially oxidized in comparison with palmitate. For this reason, the caloric value of fats is about twice that of sugars (Table 15.2).

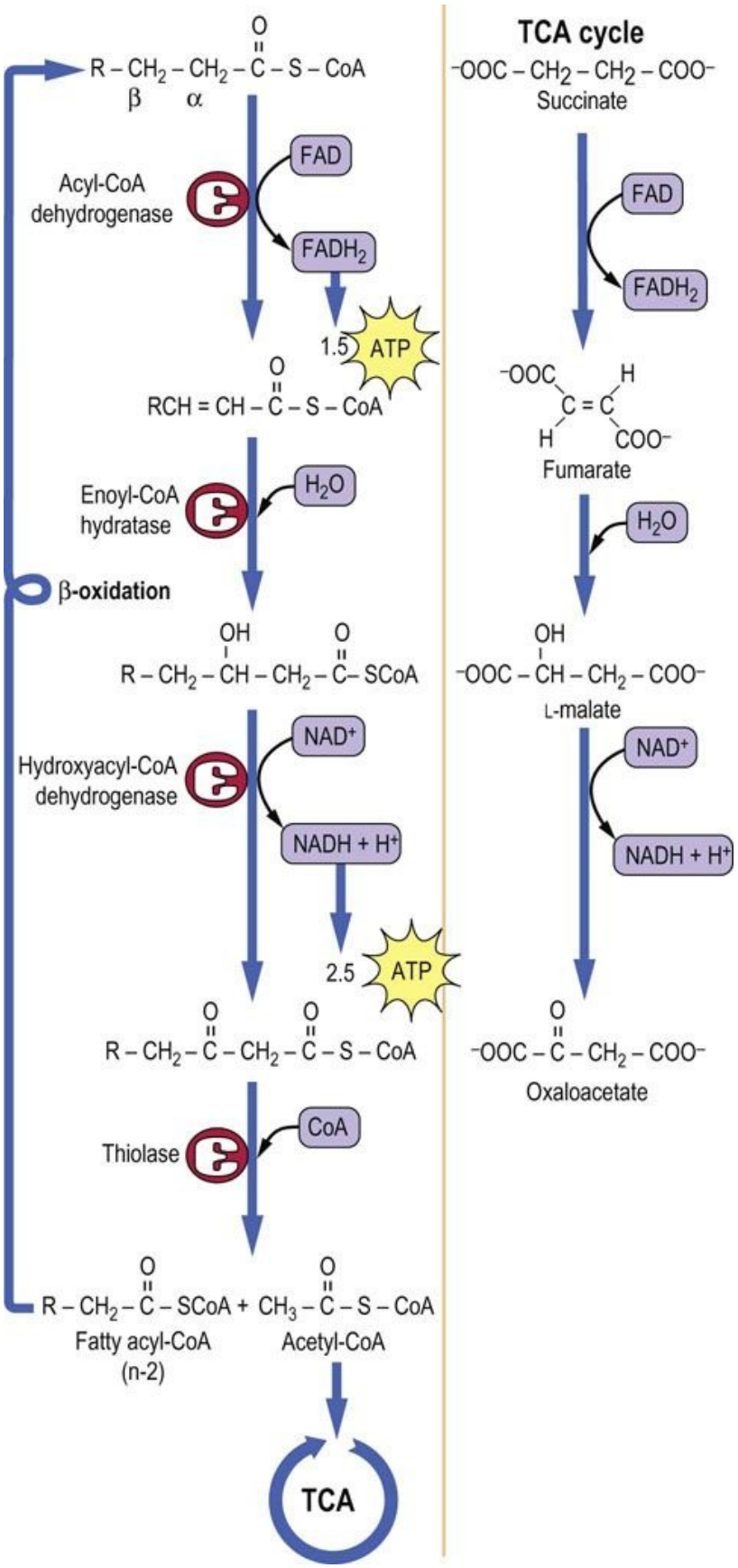

TCA cycle
Succinate
FAD
FADH2
Acyl-CoA dehydrogenase
1.5 ATP
Fumarate
H2O
Enoyl-CoA hydratase
β-oxidation
L-malate
NAD+
NADH + H+
Hydroxyacyl-CoA dehydrogenase
2.5 ATP
Oxaloacetate
CoA
Thiolase
Fatty acyl-CoA (n-2)
Acetyl-CoA
TCA

FIG. 15.4 β-Oxidation of fatty acids.
Oxidation occurs in a series of steps at the carbon that is β to the keto group. Thiolase cleaves the resultant β-ketoacyl-CoA derivative to give acetyl-CoA and a fatty acid with two fewer carbon atoms, which then re-enters the β-oxidation cascade. Note the similarity between these reactions and those of the TCA cycle, shown on the right.

The four steps in the cycle of β-oxidation are shown in detail in Figure 15.4. Note the similarity between the sequence of these reactions and those from succinate to oxaloacetate in the TCA cycle. In common with succinate dehydrogenase, acyl-CoA dehydrogenase uses FAD as a coenzyme, and is an integral protein in the inner mitochondrial membrane. Even the *trans* geometry of fumarate and the stereochemical configuration of L-malate in the TCA cycle are mirrored by *trans*-enoyl-CoA and L-hydroxyacyl-CoA intermediates in β-oxidation. The last step of the β-oxidation cycle is catalyzed by thiolase, which traps the energy obtained from the carbon–carbon bond cleavage as acyl-CoA, allowing the cycle to continue without the necessity of reactivating the fatty acid. The cycle continues until all the fatty acid has been converted to acetyl-CoA, the common intermediate in the oxidation of carbohydrates and lipids.

Clinical BoxImpaired oxidation of medium-chain fatty acids fatty acyl coa dehydrogenase deficiency

Fatty acyl CoA dehydrogenase is not a single enzyme, but a family of enzymes with chain-length specificity for oxidation of short-, medium-and long-chain fatty acids; fatty acids are transferred from one enzyme to the other during chain-shortening β-oxidation reactions. Medium-chain fatty acyl CoA dehydrogenase (MCAD) deficiency is an autosomal recessive disease characterized by hypoketotic hypoglycemia. It presents in infancy, and is characterized by high concentrations of medium-chain carboxylic acids, acyl carnitines, and acyl glycines in plasma and urine. Hyperammonemia may also be present, as a result of liver damage. Concentrations of hepatic mitochondrial medium-chain acyl CoA derivatives are also increased, limiting β-oxidation and recycling of CoA during ketogenesis. The inability to metabolize fats during fasting is life threatening because it limits gluconeogenesis and causes hypoglycemia. MCAD deficiency is treated by frequent

feeding, avoidance of fasting, and carnitine supplementation. Deficiencies in short-and long-chain fatty acid dehydrogenases have similar clinical features.

Peroxisomal catabolism of fatty acids

Peroxisomes are subcellular organelles found in all nucleated cells. They are involved in the oxidation of a number of substrates, including urate, and long-, very long-and branched-chain fatty acids. They are also the principal sites of production of hydrogen peroxide (H_2O_2) in the cell, and account for nearly 20% of oxygen consumption in hepatocytes. Peroxisomes have a carnitine shuttle and conduct β-oxidation by a pathway similar to the mitochondrial pathway, except that their acyl-CoA dehydrogenase is an oxidase, rather than a dehydrogenase. $FADH_2$ produced in this and other oxidation reactions, including α-and ω-oxidation, is oxidized by molecular oxygen to produce H_2O_2. This pathway is energetically less efficient than β-oxidation in the mitochondrion where ATP is produced by oxidative phosphorylation. Peroxisomal enzymes cannot oxidize short-chain fatty acids, so products such as butanoyl-, hexanoyl-and octanoyl-carnitine are exported or diffuse from peroxisomes for further catabolism in the mitochondrion.

Zellweger syndrome, resulting from defects in import of enzymes into peroxisomes, is a severe multiorgan disorder, leading to death usually at about 6 months of age; it is characterized by accumulation of long-chain fatty acids in neuronal tissue, most likely because of the inability to turn over neuronal fatty acids. Peroxisomes also have anabolic functions. They are thought to have a role in production of acetyl-CoA for biosynthesis of cholesterol and polyisoprenoids (Chapter 17), and they contain the dihydroxyacetone-phosphate acyltransferase required for synthesis of plasmalogens (Chapter 28). The fibrates are a class of hypolipidemic drugs that act by inducing peroxisomal proliferation in liver.

Alternative pathways of oxidation of fatty acids

Unsaturated fatty acids yield less $FADH_2$ when they are oxidized

Unsaturated fatty acids are already partially oxidized, so less $FADH_2$, and correspondingly less ATP, is produced by their oxidation. The double bonds in polyunsaturated fatty acids have *cis* geometry and occur at three-carbon intervals, whereas the intermediates in β-oxidation have *trans* geometry and the reactions proceed in two-carbon steps. The metabolism of unsaturated fatty acids therefore requires additional isomerase and oxidoreductase enzymes, both to shift the position and to change the geometry of the double bonds.

Odd-chain fatty acids produce succinyl-CoA from propionyl-CoA

The oxidation of fatty acids with an odd number of carbons proceeds from the carboxyl end, like that of normal fatty acids, except that propionyl-CoA is formed by the last thiolase cleavage reaction. The propionyl-CoA is converted to succinyl-CoA by a multistep process involving three enzymes and the vitamins **biotin and cobalamin** (Fig. 15.5). The succinyl-CoA enters directly into the TCA cycle.

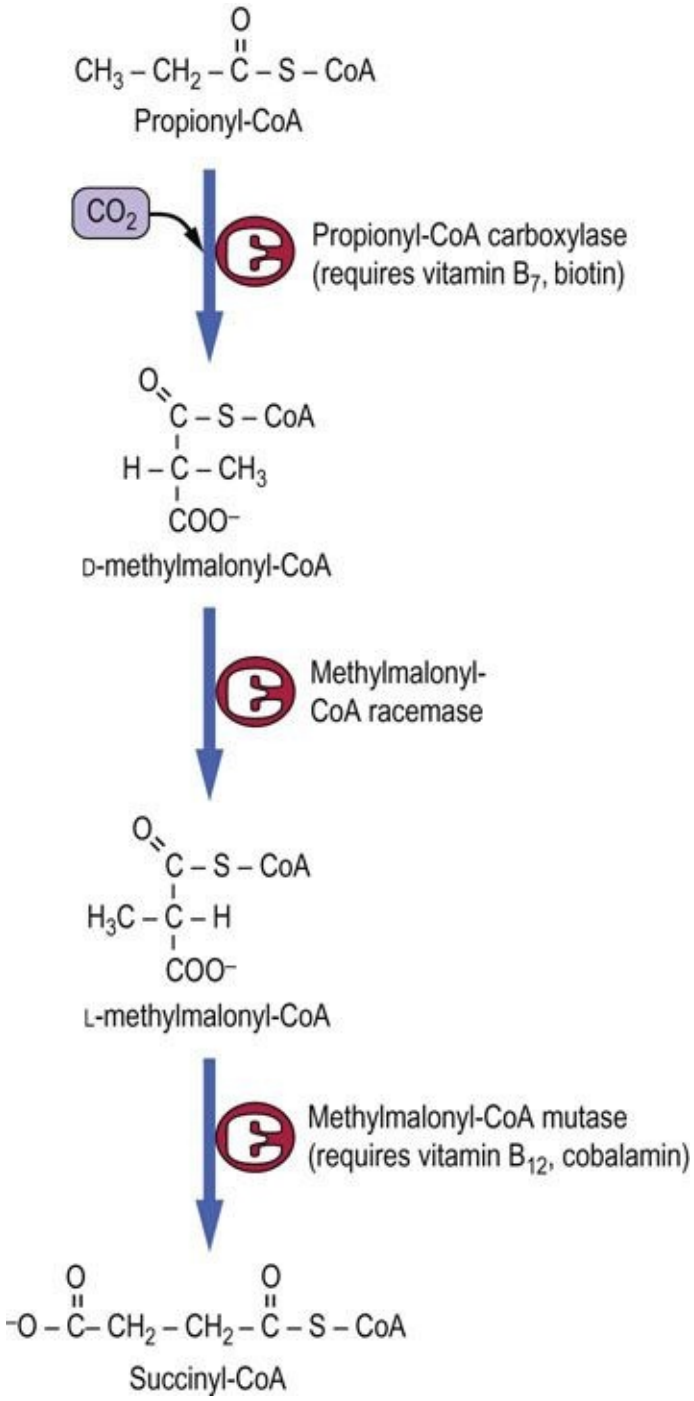


FIG. 15.5 Metabolism of propionyl-CoA to succinyl-CoA.
Propionyl-CoA from odd-chain fatty acids is a minor source of carbons for gluconeogenesis. The intermediate, methylmalonyl-CoA, is also produced during catabolism of branched-chain amino acids. Defects in methylmalonyl-CoA mutase or deficiencies in vitamin B_{12} lead to methylmalonic aciduria.

α-Oxidation initiates oxidation of branched-chain fatty acids to acetyl-CoA and propionyl-CoA

Phytanic acids are branched-chain polyisoprenoid lipids found in plant chlorophylls. Because the β-carbon of phytanic acids is at a branch point, it is not possible to oxidize this carbon to a ketone. The first and essential step in catabolism of phytanic acids is α-oxidation to a pristanic acid, releasing the α-carbon as carbon dioxide. Thereafter, as shown in Figure 15.6, acetyl-CoA and propionyl-CoA are released alternately and in equal amounts. Refsum's disease is a rare neurologic disorder, characterized by accumulation of phytanic acid deposits in nerve tissues as a result of a genetic defect in α-oxidation.

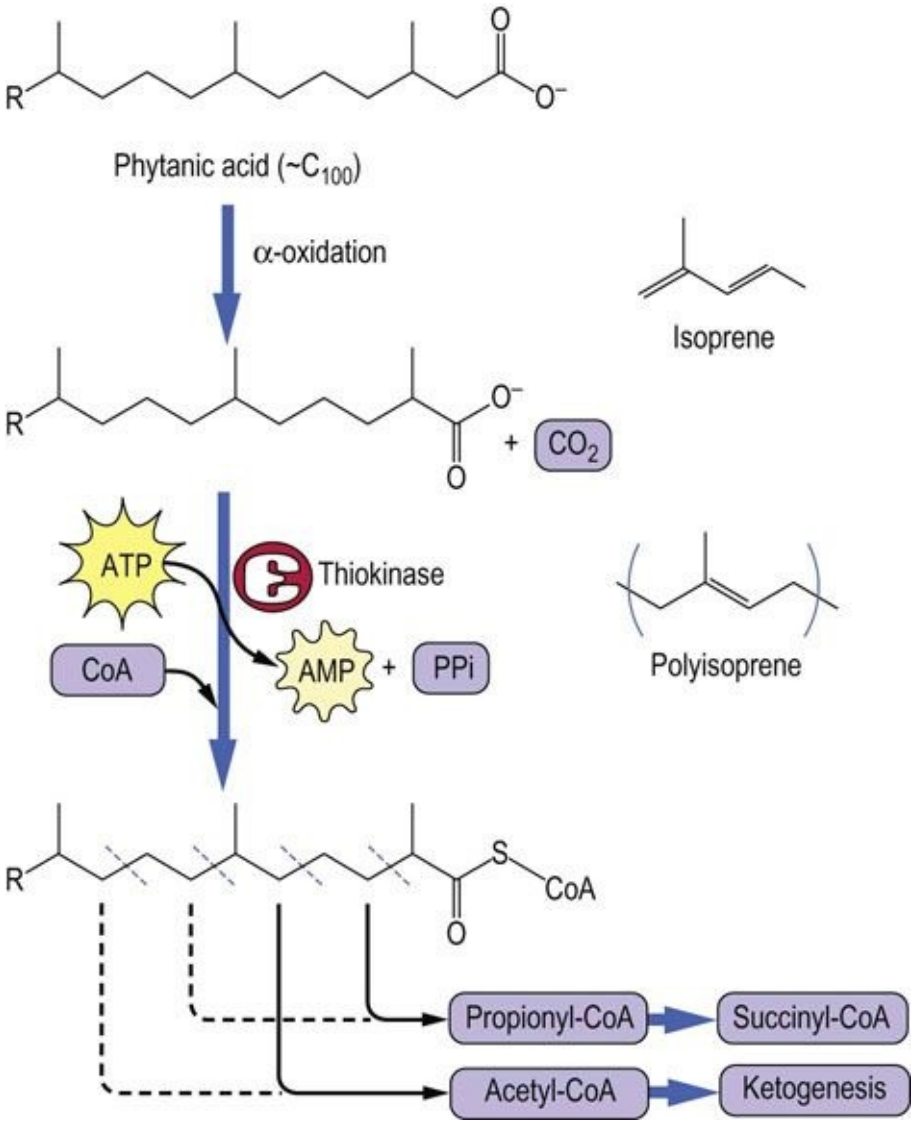


FIG. 15.6 α-Oxidation of branched-chain phytanic acids.
The first carbon of phytanic acids is removed as carbon dioxide. In subsequent cycles of β-oxidation, acetyl-CoA and propionyl-CoA are released alternately.

Ketogenesis – a metabolic pathway unique to liver

Ketogenesis in fasting and starvation

Ketogenesis is a pathway for regenerating CoA from excess acetyl-CoA

The liver uses fatty acids as its source of energy for gluconeogenesis during fasting and starvation. Fats are a rich source of energy and, under conditions of fasting or starvation, liver mitochondrial concentrations of fat-derived ATP and NADH are high, inhibiting isocitrate dehydrogenase and shifting the oxaloacetate–malate equilibrium toward malate. TCA cycle intermediates that are formed from amino acids released from muscle as part of the response to fasting and starvation (see Chapter 21) are also converted to malate in the TCA cycle. The malate exits the mitochondrion to take part in gluconeogenesis (Chapter 13). The resulting low level of oxaloacetate in hepatic mitochondria limits the activity of the TCA cycle, resulting in an inability to metabolize acetyl-CoA efficiently in the TCA cycle. Although the liver could obtain sufficient energy to support gluconeogenesis simply by the enzymes of β-oxidation, which generate both $FADH_2$ and NADH, the accumulation of acetyl-CoA, with concomitant depletion of CoA, eventually limits β-oxidation.

Clinical BoxDefects in β-oxidation

Dicarboxylic Aciduria and β-Oxidation of Fatty Acids

Several disorders of lipid catabolism, including alterations in the carnitine shuttle, acyl-CoA dehydrogenase deficiencies, and Zellweger syndrome (a defect in peroxisome biogenesis) are associated with the appearance of medium-chain dicarboxylic acids in urine. When β-oxidation of fatty acids is impaired, fatty acids are oxidized, one carbon at a time, by α-oxidation or from the ω-carbon by microsomal cytochrome P_{450}-dependent hydroxylases and dehydrogenases. These dicarboxylic acids are substrates for peroxisomal β-oxidation, which continues to the level of short-

chain dicarboxylic acids, which are then excreted from the peroxisome and eventually appear in urine.

What does the liver do with the excess acetyl-CoA that accumulates in fasting or starvation?

The problem of dealing with excess acetyl-CoA is a critical one because CoA is present in only catalytic amounts in tissues, and free CoA is required to initiate and continue the cycle of β-oxidation which is the primary source of ATP in liver during gluconeogenesis. To recycle the acetyl-CoA, the liver uses a unique pathway known as ketogenesis, in which free CoA is regenerated and the acetate group appears in blood in the form of three water-soluble lipid-derived products: **acetoacetate, β-hydroxybutyrate, and acetone**. The pathway of formation of these 'ketone bodies' (Fig. 15.7) involves the synthesis and decomposition of **hydroxymethylglutaryl (HMG)-CoA** in the mitochondrion. The liver is unique in its content of HMG-CoA synthase and lyase, but is deficient in enzymes required for metabolism of ketone bodies, which explains their export into blood.

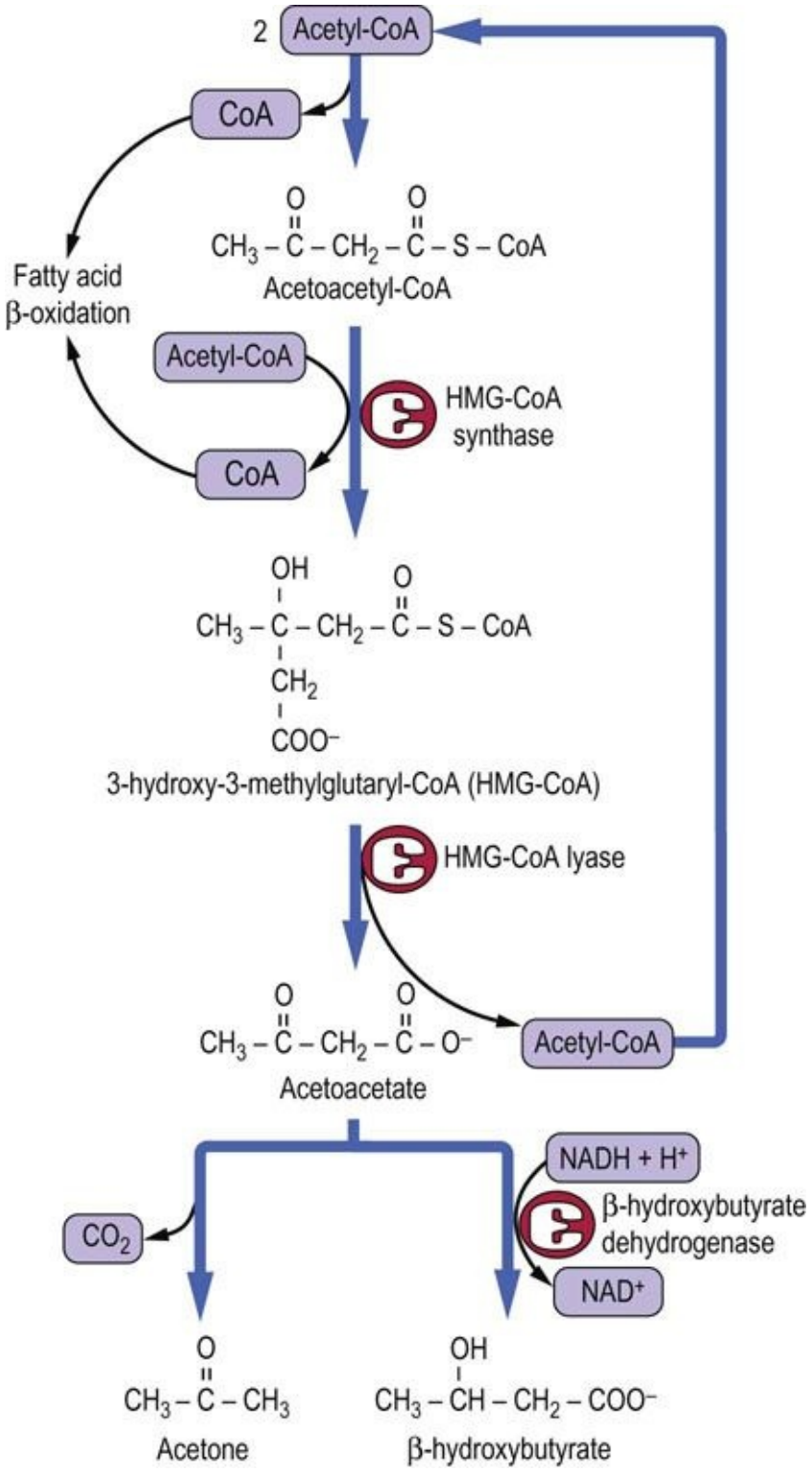


FIG. 15.7 Pathway of ketogenesis from acetyl-CoA.
Ketogenesis generates ketone bodies from acetyl-CoA, releasing the CoA to participate in β-oxidation. The enzymes involved, HMG-CoA synthase and lyase, are unique to hepatocytes; mitochondrial HMG-CoA is an essential intermediate. The initial product is

acetoacetic acid, which may be enzymatically reduced to β-hydroxybutyrate by β-hydroxybutyrate dehydrogenase, or may spontaneously (nonenzymatically) decompose to acetone, which is excreted in urine or expired by the lungs.

Ketone bodies are taken up in extrahepatic tissues, including skeletal and cardiac muscle, where they are converted to CoA derivatives for metabolism (Fig. 15.8). Ketone bodies increase in plasma during fasting and starvation (Table 15.3) and are a rich source of energy. They are used in cardiac and skeletal muscle in proportion to their plasma concentration. During starvation, the brain also converts to the use of ketone bodies for more than 50% of its energy metabolism, sparing glucose and reducing the demand on degradation of muscle protein for gluconeogenesis (see Chapters 13 and 21).

Table 15.3

Plasma concentrations of fatty acids and ketone bodies in different nutritional states

Substrate	Plasma concentration (mmol/L)		
	Normal	Fasting	Starvation
Fatty acids	0.6	1.0	1.5
Acetoacetate	<0.1	0.2	1–2
β-Hydroxybutyrate	<0.1	1	5–10

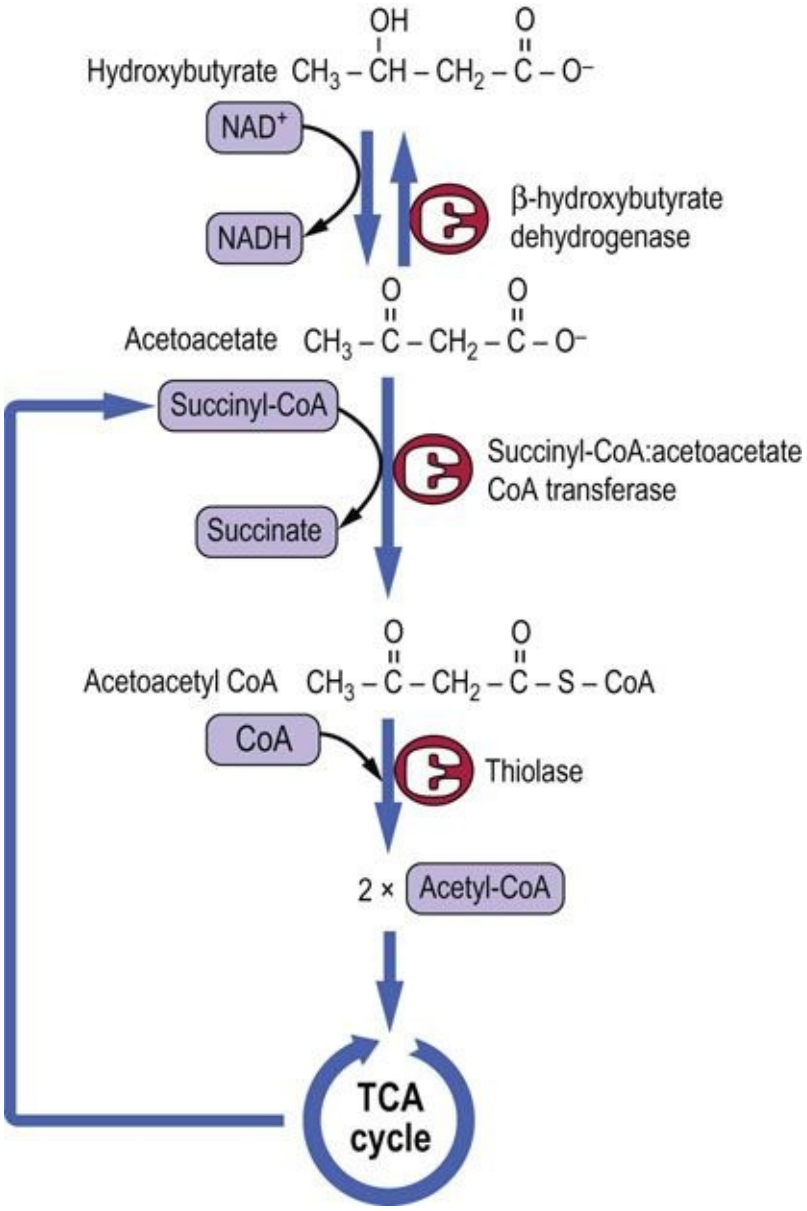


FIG. 15.8 Catabolism of ketone bodies in peripheral tissues.
Succinyl-CoA:acetoacetate CoA transferase catalyzes the conversion of acetoacetate to acetoacetyl-CoA. A thiokinase-type enzyme may also directly activate acetoacetate in some tissues.

Clinical BoxKetone bodies in urine (ketonuria) and weight-loss programs

The appearance of ketone bodies in the urine is an indication of

active fat metabolism and gluconeogenesis. Ketonuria may also occur normally in association with a high-fat, low-carbohydrate diet. Some weight-loss programs encourage gradual reduction in carbohydrate and total caloric intake until ketone bodies appear in urine (measured with Ketostix). Dieters are urged to maintain this level of caloric intake, checking urinary ketones regularly to confirm the consumption of body fat.

Comment.

Ketostix and similar ‘dry chemistry’ tests are convenient test strips for estimating urinary ketone bodies. They strips contain a chemical reagent, such as nitroprusside, which reacts with acetoacetate in urine to form a lavender color, graded on a scale with a maximum of ‘4+’ (see Chapter 23 and Fig. 23.13). A reaction of ‘1+’ (representing 5–10 mg ketone bodies/100 mL) or ‘2+’ (10–20 mg/100 mL) on the test strip was established as a goal to assure continued fat metabolism, and therefore weight loss. This type of diet is discouraged today, because the appearance of ketone bodies in the urine indicates greater concentrations in the plasma, and may cause metabolic acidosis.

Clinical BoxDefective ketogenesis

Ketogenesis as a Result of a Deficiency in Carnitine Metabolism

The clinical presentation of deficiencies in carnitine metabolism occurs in infancy and is often life threatening. Characteristic features include hypoketotic hypoglycemia, hyperammonemia, and altered plasma free carnitine concentration. Hepatic damage, cardiomyopathy, and muscle weakness are common.

Comment.

Carnitine is synthesized from lysine and α-ketoglutarate, primarily in liver and kidney, and is normally present in plasma in a

concentration of about 50 µmol/L (8 mg/dL). There are high-affinity uptake systems for carnitine in most tissues, including the kidney, which resorbs carnitine from the glomerular filtrate, limiting its excretion in urine. Homozygous deficiencies in carnitine transporters, CPT-I and -II, and the translocase result in defects in long-chain fatty acid oxidation. Plasma and tissue carnitine concentrations decrease to <1 µmol/L in carnitine transport deficiency, because of both defective uptake into tissues and excessive loss in urine. On the other hand, plasma free carnitine may exceed 100 µmol/L (20 mg/dL) in CPT-I deficiency. In both translocase and CPT-II deficiency, total plasma carnitine may be normal, but is mostly in the form of acyl carnitine esters of long-chain fatty acids – in the former case, because they cannot be transported into the mitochondrion, and in the latter because of backflow out from mitochondria. These diseases are treated by carnitine supplementation, by frequent high-carbohydrate feeding, and by avoidance of fasting.

Mobilization of lipids during gluconeogenesis

Carbohydrate and lipid metabolism are coordinately regulated by hormone action during the feed-fast cycle

Insulin, glucagon, epinephrine, and cortisol control the direction and rate of glycogen and glucose metabolism in liver. During fasting and starvation, hepatic gluconeogenesis is activated by glucagon and requires the coordinated degradation of proteins and release of amino acids from muscle, and the degradation of triglycerides and release of fatty acids from adipose tissue. The latter process, known as lipolysis, is controlled by the adipocyte enzyme **hormone-sensitive lipase**, which is activated by phosphorylation by cAMP-dependent protein kinase A in response to increasing plasma concentrations of glucagon (Chapters 13 and 21). Like gluconeogenesis, lipolysis is inhibited by insulin.

The activation of hormone-sensitive lipase has predictable effects – increasing the concentration of free fatty acids and glycerol in plasma during fasting and

starvation (Fig. 15.9); similar effects are observed in response to epinephrine during the stress response. Epinephrine activates both glycogenolysis in the liver and lipolysis in adipose tissue so that both fuels, glucose and fatty acids, increase in blood during stress. Cortisol exerts a more chronic effect on lipolysis and also causes insulin resistance. Cushing's syndrome (Chapter 39), in which there are high blood concentrations of cortisol, is characterized by hyperglycemia, muscle wastage, and redistribution of fat from glucagon-sensitive adipose depots to atypical sites, such as the cheeks, upper back, and trunk.

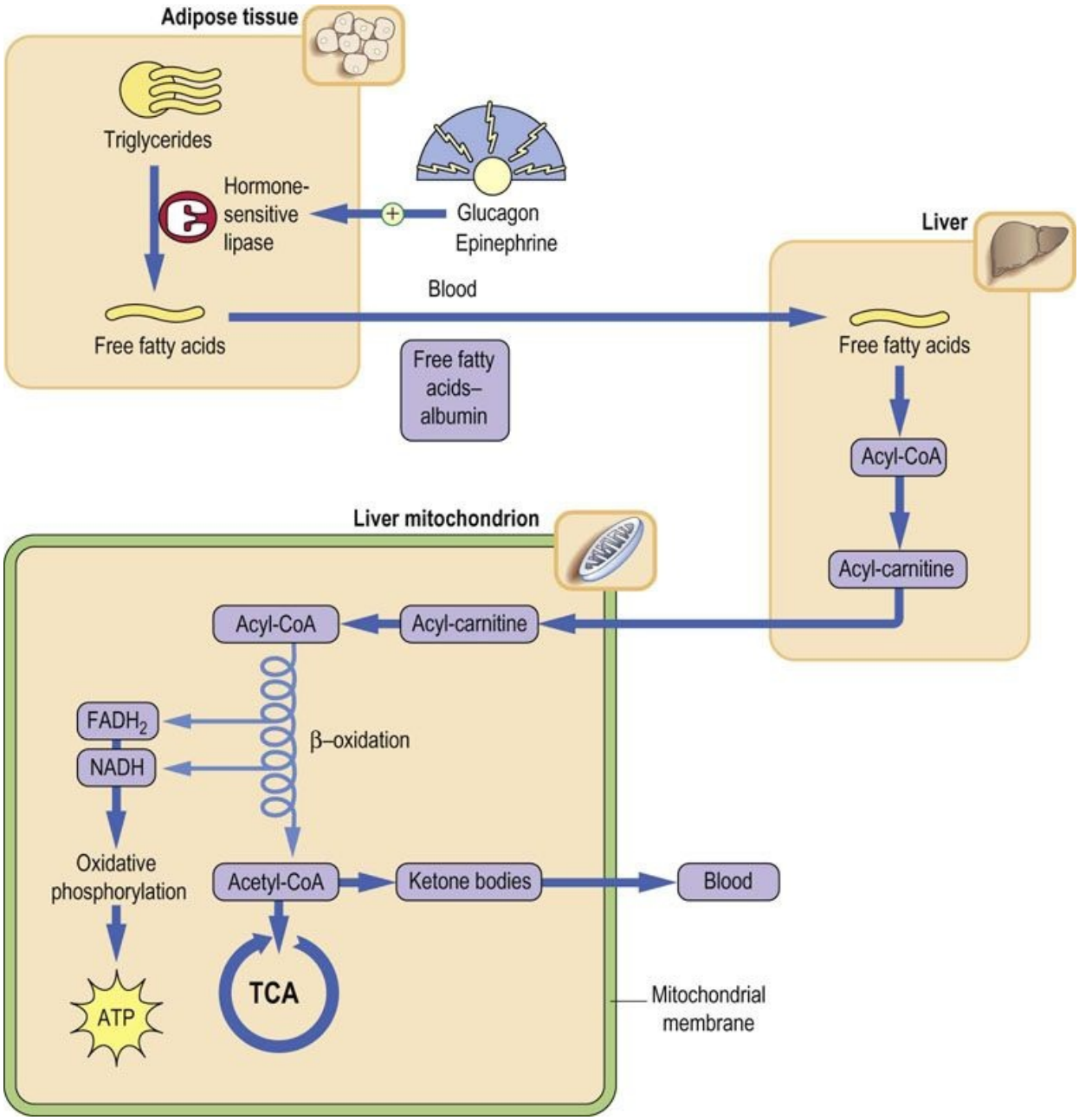


FIG. 15.9 Regulation of lipid metabolism by glucagon and epinephrine. Glucagon and epinephrine activate hormone-sensitive lipase in adipose tissue, in coordination with activation of proteolysis in muscle and gluconeogenesis in liver. Metabolism of fatty acids through β-oxidation in liver yields ATP for gluconeogenesis. The acetyl-CoA is converted to and released to blood as ketone bodies. These effects are reversed by insulin following a meal.

Regulation of ketogenesis

Ketogenesis is activated in concert with gluconeogenesis during fasting and starvation

Ketogenesis increases when hormone-sensitive lipase is activated by glucagon in adipose tissue during fasting and starvation, and in diabetes. Under these conditions, plasma fatty acid concentration increases, and the liver uses these fatty acids to support gluconeogenesis. The energy is derived primarily from β-oxidation, and the product, acetyl-CoA, is metabolized by ketogenesis. Why isn't the acetyl-CoA used in the tricarboxylic acid cycle?

During gluconeogenesis, glucagon activation of the cAMP cascade in liver inhibits glycolysis (Fig. 13.9), limiting pyruvate flux from carbohydrates. Any pyruvate that is formed, mostly from lactate and alanine, is converted to oxaloacetate by pyruvate carboxylase, which is activated by acetyl-CoA (Fig. 13.8). The oxaloacetate is then converted to malate for gluconeogenesis and, because of the low level of oxaloacetate, a substrate for citrate synthase, the acetyl-CoA is directed to ketogenesis, rather than used for energy metabolism in the tricarboxylic acid cycle. The orientation toward ketogenesis is controlled by the energy charge of liver. The high ATP concentration, produced by fat metabolism, inhibits the tricarboxylic acid cycle at the isocitrate dehydrogenase step (Chapter 14). Further, by respiratory control (Chapter 9), high ATP leads to an increase in the mitochondrial membrane potential, which inhibits the electron transport chain. The resulting increase in the NADH/NAD^+ ratio favors reduction of oxaloacetate to malate, which exits the mitochondrion for gluconeogenesis, rather than consumption in the tricarboxylic acid cycle. In summary, during gluconeogenesis, acetyl-CoA derived from fatty acid metabolism is converted to ketone bodies; it has nowhere else to go! The increase in ketone bodies in plasma (**ketonemia**) leads to their appearance in urine (**ketonuria**). In type 1 diabetes, the high rate of ketogenesis may lead to excessive ketonemia, and possibly to life-threatening diabetic **ketoacidosis** (Chapter 21).

Clinical BoxHellp and aflp syndromes in mothers of children born with lchad (incidence 1 in 200,000)

Long-chain L-3-hydroxyacyl-CoA dehydrogenase (LCAD) deficiency can present in a wide variety of ways. Those affected are prone to episodes of nonketotic hypoglycaemia, but may

develop fulminant hepatic failure, cardiomyopathy, rhabdomyolysis, and occasionally neuropathy and retinopathy. As with deficiencies in MCAD, treatment involves avoidance of fasting and diets enriched in medium-chain fatty acids.

Perhaps the most striking feature of this rare defect in fatty acid metabolism is the association with maternal HELLP (**h**emolysis, **e**levated **l**iver enzymes and **l**ow **p**latelets) and AFLP (**a**cute **f**atty **l**iver of **p**regnancy). These potentially fatal obstetric emergencies may occur in mothers who are heterozygotes for LCHAD, especially if the child has LCHAD. These syndromes are also associated with another recessive fatty acid defect, carnitine palmitoyl-transferase-I deficiency.

Summary

■ Unlike carbohydrate fuels, which enter the body primarily as glucose or sugars that are converted to glucose, lipid fuels are heterogeneous with respect to chain length, branching, and unsaturation.
■ The catabolism of fats is primarily a mitochondrial process, but also occurs in peroxisomes.
■ Using a variety of chain length-specific transport processes and catabolic enzymes, the primary pathways of catabolism of fatty acids involve their oxidative degradation in two-carbon units, a process known as β-oxidation, which produces acetyl-CoA.
■ In most tissues, the acetyl-CoA units are used for ATP production in the mitochondrion.
■ In liver, acetyl-CoA is catabolized to ketone bodies, primarily acetoacetate and β-hydroxybutyrate, by a mitochondrial pathway termed ketogenesis. The ketone bodies are exported from liver for energy metabolism in peripheral tissue.
■ Ketonemia and ketonuria develop gradually during fasting, while ketoacidosis may develop during poorly controlled diabetes when fat metabolism is increased to high levels for support of gluconeogenesis.

Active learning

1. Compare the metabolism of acetyl-CoA in liver and muscle. Explain why the liver produces ketone bodies during gluconeogenesis. What prevents hepatic oxidation of acetyl-CoA?
2. Review the merits of carnitine usage as a performance enhancer during exercise and as a supplement for geriatric patients.
3. Review the current use and mechanism of action of peroxisome proliferator drugs for treatment of dyslipidemia and diabetes.
4. Compare the mechanisms underlying development of ketoacidotic hyperglycemia and nonketotic hypoglycemia.

Further reading

Cahill, GF, Jr. Fuel metabolism in starvation. *Annu Rev Nutr*. 2006; 26:1–22.

Islinger, M, Grille, S, Fahimi, HD, *et al.* The peroxisome: an update on mysteries. *Histochem Cell Biol*. 2012; 137:547–574.

Kossoff, EH, Hartman, AL. Ketogenic diets: new advances for metabolism-based therapies. *Curr Opin Neurol*. 2012; 25:173–178.

Longo, N, Amat di San Filippo, C, Pasquali, M. Disorders of carnitine transport and the carnitine cycle. *Am J Med Genet C Semin Med Genet*. 2006; 142:77–85.

McPherson, PA, McEneny, J. The biochemistry of ketogenesis and its role in weight management, neurological disease and oxidative stress. *J Physiol Biochem*. 2012; 68:141–151.

Rector, RS, Ibdah, JA. Fatty acid oxidation disorders: maternal health and neonatal outcomes. *Semin Fetal Neonatal Med*. 2010; 15:122–128.

Sass, JO. Inborn errors of ketogenesis and ketone body utilization. *J Inherit Metab Dis*. 2012; 35:23–28.

Solis, JO, Singh, RH. Management of fatty acid oxidation disorders: a survey of current treatment strategies. *J Am Diet Assoc*. 2002; 102:1800–1803.

Stojanovic, V, Ihle, S. Role of beta-hydroxybutyric acid in diabetic ketoacidosis: a review. *Can Vet J*. 2011; 52:426–430.

Wanders, RJ, Waterham, HR. Peroxisomal disorders: the single peroxisomal enzyme deficiencies. *Biochim Biophys Acta*. 2006; 1763:1707–1720.

Wanders, RJ, Ferdinandusse, S, Brites, P, *et al.* Peroxisomes, lipid metabolism and lipotoxicity. *Biochim Biophys Acta*. 2010; 1801:272–280.

Wilcken, B. Fatty acid oxidation disorders: outcome and long-term prognosis. *J Inherit Metab Dis*. 2010; 33:501–506.

Wolfsdorf, J, Craig, ME, Daneman, D, *et al.* Diabetic ketoacidosis in children and adolescents with diabetes. *Pediatr Diabetes*. 2009; 10(Suppl 12):118–133.

Websites

Carnitine. http://lpi.oregonstate.edu/infocenter/othernuts/carnitine/.

Lipids OnLine – slide library. www.lipidsonline.org/slides/.

Peroxisomal disorders. www.emedicine.com/neuro/topic309.htm.

CHAPTER 16

Biosynthesis and Storage of Fatty Acids

Fredrik Karpe and John I. Broom

Learning objectives

After reading this chapter you should be able to:

- Describe the pathway of fatty acid synthesis, and in particular the roles of acetyl-CoA carboxylase and the multifunctional enzyme fatty acid synthase.
- Outline short-term and long-term regulation of fatty acid synthesis.
- Explain the concepts of elongation and desaturation of the fatty acid chain.
- Describe the synthesis of triglycerides.
- Discuss the endocrine function of adipose tissue.

Introduction

Most fatty acids required by humans are supplied in the diet; however, the pathway for their de novo synthesis (lipogenesis) from two-carbon compounds is present in many tissues such as liver, brain, kidney, mammary gland and adipose tissue. It is also highly active in many cancers. In general, the pathway of de novo synthesis is primarily active in situations of excess energy intake, particularly in the form of excess carbohydrate. In this situation, carbohydrate, and to a lesser extent amino acid precursors, are converted to fatty acids in the liver and stored as triacylglycerol (TAG, also known as triglyceride) in cellular lipid droplets. This may happen in many cell types but the adipocytes (and adipose tissue) are dedicated to store large quantities of TAG. In humans, adipose tissue may not be quantitatively the most important place of fatty acid synthesis: the main lipogenic organ is the liver.

The pathway for lipogenesis is not simply the reverse of oxidation of fatty acids (Chapter 15). Lipogenesis requires a completely different set of enzymes and is located in a different cellular compartment, the cytosol. Furthermore, it uses reduced nicotinamide adenine dinucleotide phosphate (NADPH) as a source of reductive power, as opposed to nicotinamide adenine dinucleotide (NAD^+) required for β-oxidation.

The sterol regulatory element-binding proteins-1 (mainly SREBP 1c ; also SREBP1a) provides a master regulation of de novo lipogenesis through transcriptional control. The SREBP is an endoplasmic-reticulum-bound and membrane-sensing protein that undergoes proteolytic cleavage enabling its transport to the nucleus. In the nucleus, SREBP binds to specific DNA sequences (the sterol regulatory elements or SREs) located in the control regions of the genes that encode enzymes needed for lipogenesis.

Fatty acid synthesis

Fatty acids are synthesized from acetyl-CoA

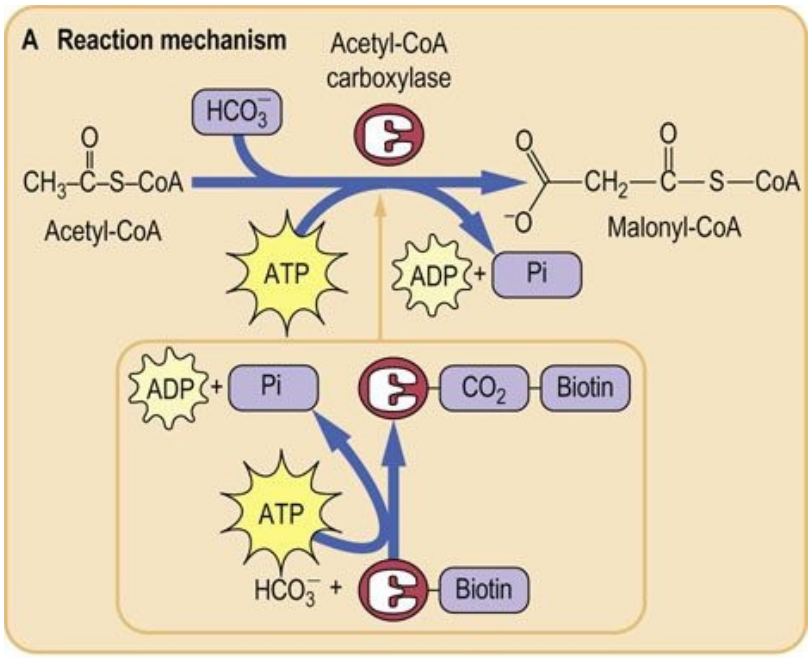
The synthesis of fatty acids in mammalian systems can be considered as a two-stage process, both stages requiring acetyl-CoA units and both employing multifunctional proteins in multienzyme complexes.

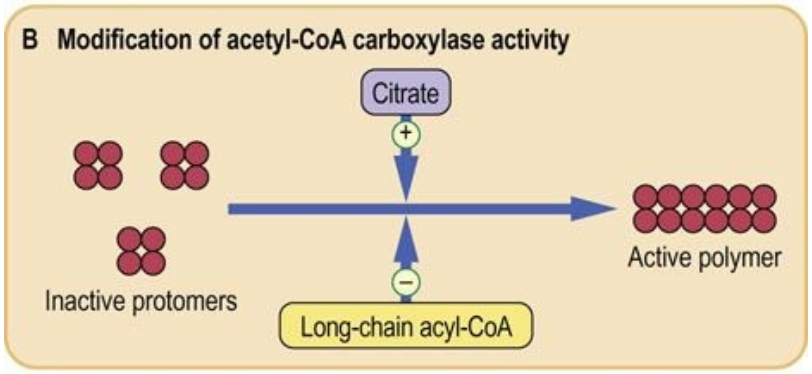
- Stage 1 is formation of the key precursor malonyl-CoA from acetyl-CoA by acetyl-CoA carboxylase.
- Stage 2 is elongation of the fatty acid chain in two-carbon increments by fatty acid synthase.

The preparatory stage: acetyl-CoA carboxylase

Carboxylation of acetyl-CoA to malonyl-CoA is the committed step of fatty acid synthesis

In the first stage of fatty acid biosynthesis, acetyl-CoA, mostly derived from carbohydrate metabolism, is converted to malonyl-CoA by the action of the enzyme acetyl-CoA carboxylase (Fig. 16.1). There are two forms of acetyl-CoA carboxylase (ACC1 and ACC2). ACC1 is located in the cytoplasm and committed to fatty acid synthesis, whereas ACC2 is in mitochondria where it regulates fatty acid oxidation (ACC2 inhibition results in increased lipid oxidation). ACC1 is a biotin-dependent enzyme with distinct enzymatic and a carrier protein function: its subunits serve as a biotin carboxylase, a transcarboxylase and biotin carboxyl carrier protein. The enzyme is synthesized in an inactive protomer form, each protomer containing all the above subunits, a molecule of biotin, and a regulatory allosteric site for the binding of citrate (a Krebs cycle metabolite) or palmitoyl-CoA (the end product of the fatty acid biosynthetic pathway). The reaction itself takes place in stages: first, there is the carboxylation of biotin, involving adenosine triphosphate (ATP), followed by the transfer of this carboxyl group to acetyl-CoA to produce the malonyl-CoA. At this stage, the free enzyme–biotin complex is released.


A Reaction mechanism
Acetyl-CoA carboxylase
HCO_3^-
$CH_3–C(=O)–S–CoA$
Acetyl-CoA
$^-O–C(=O)–CH_2–C(=O)–S–CoA$
Malonyl-CoA
ATP
ADP + Pi
ADP + Pi
CO_2
Biotin
ATP
HCO_3^- +
Biotin


B Modification of acetyl-CoA carboxylase activity
Citrate
+
Inactive protomers
Active polymer
−
Long-chain acyl-CoA

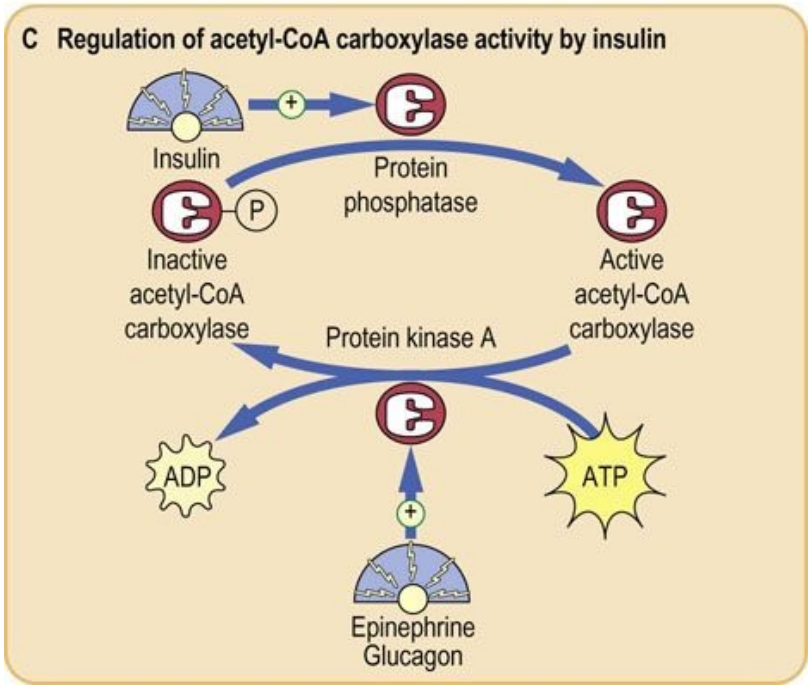

C Regulation of acetyl-CoA carboxylase activity by insulin
+
Insulin
Protein phosphatase
P
Inactive acetyl-CoA carboxylase
Active acetyl-CoA carboxylase
Protein kinase A
ADP
ATP
+
Epinephrine
Glucagon

FIG. 16.1 Conversion of acetyl-CoA to malonyl-CoA.
(A) Reaction catalyzed by the acetyl-CoA carboxylase. The enzyme has the covalently attached biotin, which is carboxylated using a molecule of ATP. **(B)** Acetyl-CoA carboxylase requires the presence of citrate for polymerization to its active form. **(C)** The activity of acetyl-CoA carboxylase is regulated by a phosphorylation–dephosphorylation mechanism. This in turn is controlled by hormones that regulate fuel metabolism: insulin, glucagon and epinephrine.

This process allows the building up of fatty acid molecules with even numbers of carbon atoms. Propionyl-CoA is a substrate for the synthesis of fatty acids with an odd number of carbon atoms, but this is not seen in humans.

Acetyl-CoA carboxylase is subject to strict regulation

The protomers of acetyl-CoA carboxylase polymerize in the presence of citrate or isocitrate, producing the active form of the enzyme. The polymerization is inhibited by palmitoyl-CoA binding to the same allosteric site. The respective stimulatory and inhibitory effects of citrate and palmitoyl-CoA are entirely logical: under conditions of high citrate concentration, energy storage is desirable but when palmitoyl-CoA, the product of the pathway, accumulates, a decrease in the synthesis of fatty acids is appropriate. There is an additional control mechanism, independent of the citrate or palmitoyl-CoA, involving phosphorylation and dephosphorylation of the enzyme molecule. This involves hormone-dependent protein phosphatase/kinase (Fig. 16.1). Phosphorylation inhibits the enzyme and dephosphorylation activates it. Phosphorylation of the enzyme is promoted by glucagon or epinephrine and the dephosphorylation is promoted by the insulin, which is a lipogenic hormone. Phosphorylation is also dependent on the activation of the AMP-activated protein kinase (AMPK).

Dietary carbohydrate and fat intake also controls acetyl-CoA carboxylase

The carboxylation of acetyl-CoA to malonyl-CoA commits the pathway to fatty acid synthesis. This is why this enzyme is under such strict short-term control. Longer-term control is exerted by the induction or repression of enzyme synthesis affected by diet: synthesis of acetyl-CoA carboxylase is upregulated under conditions of high-carbohydrate/low-fat intake, whilst starvation or high-fat/low-carbohydrate intake leads to downregulation of synthesis of the enzyme.

Synthesizing a fatty acid chain: fatty acid synthase

The second major step in fatty acid synthesis also involves a multienzyme complex, the fatty acid synthase. This enzyme system is much more complex than acetyl-CoA carboxylase. The protein contains seven distinct enzyme activities and an acyl carrier protein (ACP). ACP, a highly conserved protein, replaces CoA as the entity that binds to the elongating fatty acid chain. The structure of this molecule is shown in Figure 16.2 and consists of a dimer of large identical polypeptides arranged head to tail. Each monomer contains all seven enzyme activities and the ACP. It also contains a long pantetheine group which acts as a flexible 'arm', making the molecule being synthesized available to different enzymes in the fatty acid synthetase complex. The function in fatty acid synthesis is shared between the two polypeptide chains.

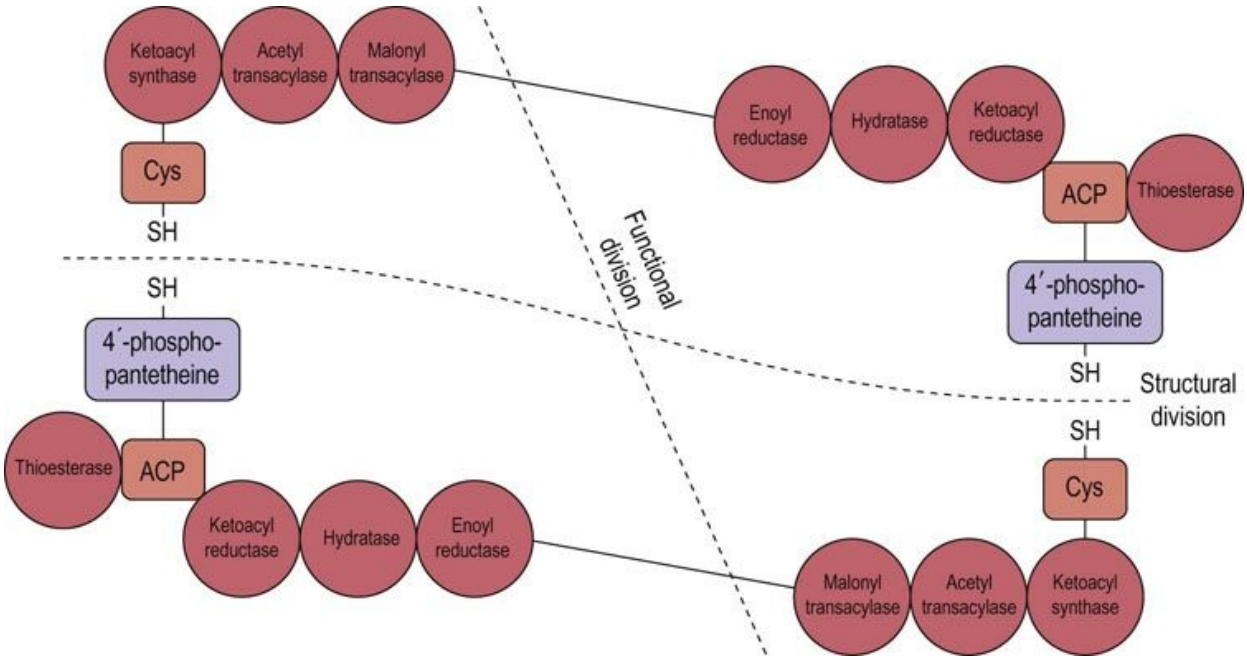


FIG. 16.2 Structure of fatty acid synthase.
Fatty acid synthase is a dimer consisting of two large subunits arranged head to tail. It contains seven distinct enzyme activities and an acylcarrier protein (ACP). Cys, cysteine.

Fatty acid synthase builds the fatty acid molecule up to 16-

carbon length

The reaction proceeds after an initial priming of the cysteine (Cys-SH) group with acetyl-CoA, a reaction catalyzed by **acetyl transacylase** (Fig. 16.3). Then malonyl-CoA is transferred by **malonyl transacylase** to the -SH residue of the pantetheine group attached to the ACP in the other subunit. Next, **3-ketoacyl synthase** (the condensing enzyme) catalyzes the reaction between the previously attached acetyl group and the malonyl residue, liberating CO_2 and forming the 3-ketoacyl enzyme complex. This frees the cysteine residue on chain 1 that had been occupied by the acetyl-CoA. The 3-ketoacyl group subsequently undergoes sequential reduction, dehydration and again reduction to form a saturated acyl–enzyme complex. The next molecule of malonyl-CoA displaces the acyl group from the pantetheine-SH group to the now free cysteine group, and the reaction sequence is repeated through six more cycles (seven cycles altogether). Once the 16-carbon chain (palmitate) is formed, the saturated acyl–enzyme complex activates the **thioesterase,** releasing the molecule of palmitate from the enzyme complex. The two -SH sites are now free, allowing another cycle of palmitate synthesis to be initiated.

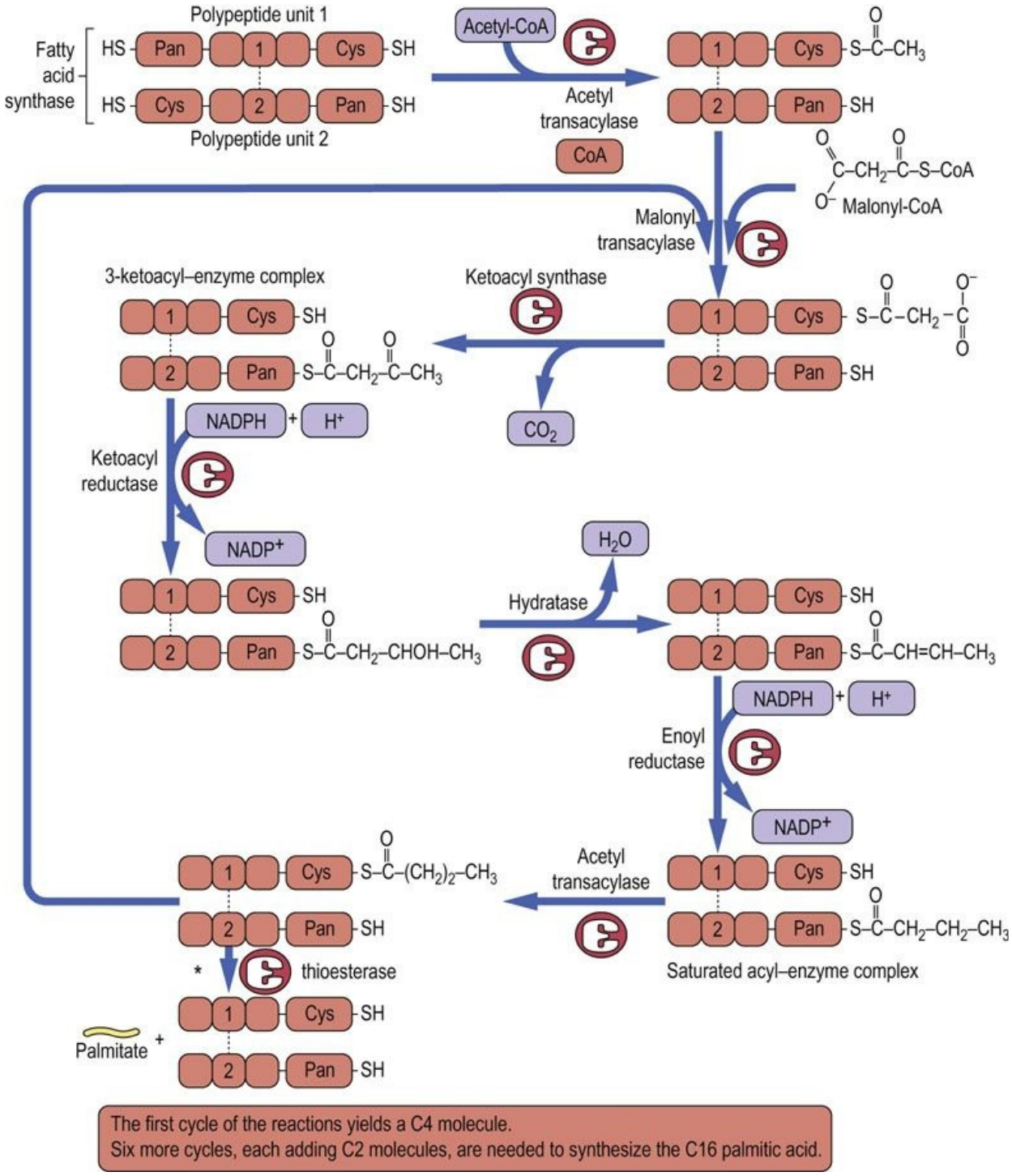


FIG. 16.3 Reactions catalyzed by fatty acid synthase.
The synthesis of a fatty acid chain is initiated by a molecule of malonyl-CoA (C3) which reacts with the first molecule of acetyl-CoA (C2); this produces a C4 molecule (1 carbon is lost as CO_2 during condensation of malonyl-CoA and acetyl-CoA). There are six more cycles, each adding 2C to the fatty acid chain (seven cycles altogether), and the result is a 16-carbon molecule of palmitate. NADPH, reduced nicotinamide adenine dinucleotide phosphate; Pan, pantetheine.

*This reaction occurs once 16-carbon fatty acyl chain has been formed.

The synthesis of one palmitate molecule requires 8 molecules of acetyl-CoA, 7 ATP, 14 NADPH and 14 H^+:

$$8\text{Ac-CoA} + 7\text{ATP} + 14\text{NADPH} + 14\text{H}^+ \rightarrow \text{CH}_3(\text{CH}_2)_{14}\text{COO}^-(\text{palmitate}) + 14\text{NADP}^+ + 8\text{CoA} + 6\text{H}_2\text{O} + 7\text{ADP} + 7\text{Pi} + 7\text{CO}_2$$

In common with the acetyl-CoA carboxylase system, fatty acid synthase is also regulated by the presence of phosphorylated sugars via an allosteric effect, and also by induction and repression of the enzyme.

Alteration in the amount of enzyme protein is affected by the nutritional state

Rates of fatty acid synthesis are greatest when an individual follows a hypercaloric, high-carbohydrate/low-fat diet and are low during fasting/starvation or when eating a high-fat diet.

The malate shuttle

Malate shuttle allows recruitment of two-carbon units from the mitochondrion to the cytoplasm

The primary molecule required for the synthesis of fatty acids is acetyl-CoA. However, acetyl-CoA is generated in the mitochondria and cannot freely cross the inner mitochondrial membrane. As said above, fatty acid biosynthesis occurs in the cytosol. The malate shuttle is a mechanism allowing the transfer of two-carbon units from the mitochondria to the cytosol: it involves the malate–citrate antiporter (Fig. 16.4). Pyruvate derived from glycolysis is decarboxylated to acetyl-CoA in the mitochondria; it subsequently reacts with oxaloacetate in the tricarboxylic acid (TCA) cycle (Chapter 14) to form citrate. Translocation of a molecule of citrate to the cytosol via the antiporter is accompanied by transfer of a molecule of malate to the mitochondrion. In the cytosol, citrate, in the presence of ATP and CoA, undergoes cleavage to acetyl-CoA and oxaloacetate by citrate

lyase. This makes acetyl-CoA available for carboxylation to malonyl-CoA and for the synthesis of fatty acids. The synthesis of fatty acids is also linked to glucose metabolism through the pentose phosphate pathway, which is the main provider of NADPH required for lipogenesis. Fructose is specifically channeled through this pathway and is very lipogenic. Some NADPH is also generated by the $NADP^+$-linked decarboxylation of malate to pyruvate by the malic enzyme.

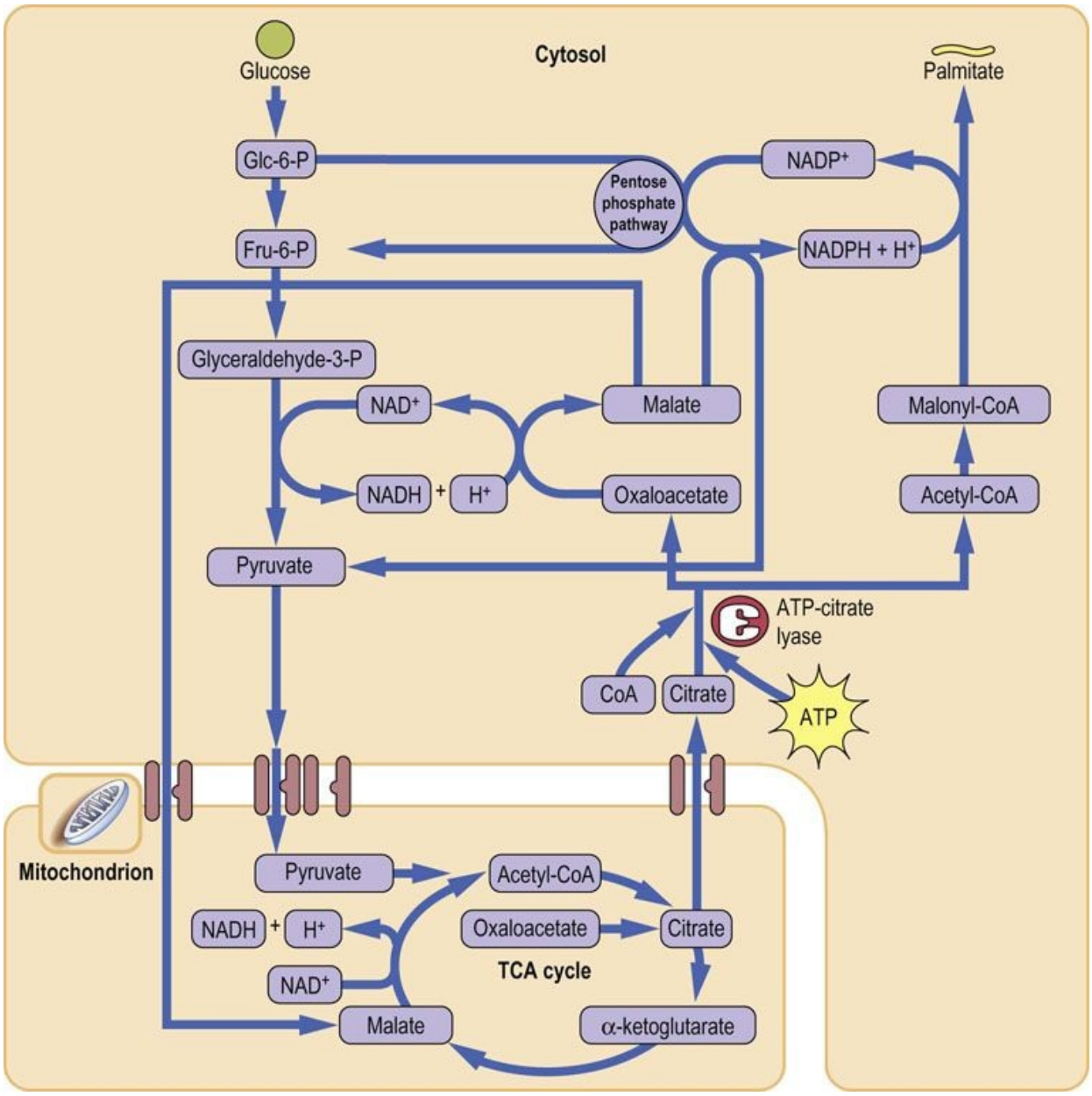


FIG. 16.4 The malate shuttle.

Acetyl-CoA is generated in the mitochondria and cannot cross the mitochondrial membrane. The malate shuttle facilitates the transport of two-carbon units from the mitochondria to cytoplasm. Citrate, synthesized from Acetyl-CoA and oxalate is transported out of the mitochondria. In the cytosol it is split back into Acetyl-CoA and oxalate. Oxalate is then converted to malate, which returns to the mitochondrion – thus, the 'shuttle'. Acetyl-CoA is resynthesized in the cytoplasm and enters lipogenesis. Fru-6-P, fructose-6-phosphate; Glc-6-P, glucose-6-phosphate; NADH, reduced nicotinamide adenine dinucleotide. (See also Fig. 9.7.)

Advanced concept box Changes in enzyme expression in response to food intake regulate storage of energy substrates

The fed state is associated with the **induction of enzymes that increase fatty acid synthesis** in the liver. Several enzymes are induced, including those involved in glycolysis, *e.g.* glucokinase (the hepatic form of hexokinase) and pyruvate kinase, as well as enzymes linked to increased production of NADPH (glucose-6-P dehydrogenase, 6-phosphogluconate dehydrogenase, and malic enzyme). Further, there is an increased expression of citrate lyase, acetyl-CoA carboxylase, fatty acid synthase, and Δ^9 desaturase.

Also, in the fed state, there is a **concomitant repression of the key enzymes involved in gluconeogenesis**. Phosphoenolpyruvate carboxykinase, glucose-6-phosphatase and some aminotransferases are reduced in amount, either by reduction in synthesis or by increased degradation (see Chapter 21).

Fatty acid elongation

Elongation of a fatty acid chain beyond 16-carbon length requires another set of enzymes

Palmitate released from fatty acid synthase becomes a substrate for the synthesis of longer-chain fatty acids, with the exception of certain essential fatty acids (see below). Chain elongation occurs by the addition of further two-carbon fragments derived from malonyl-CoA (Fig. 16.5). This process occurs on the endoplasmic reticulum by the action of yet another multienzyme complex – **fatty acid elongase**. The reactions occurring during chain elongation are similar to those involved in fatty acid synthesis, except that the fatty acid is attached to CoA, rather than to the ACP. In fact, there are seven discrete fatty acid elongases with different tissue expressions and substrate specificities (ELOVL9-7; ELOVL stands for ‘elongation of very long chain fatty acids’).

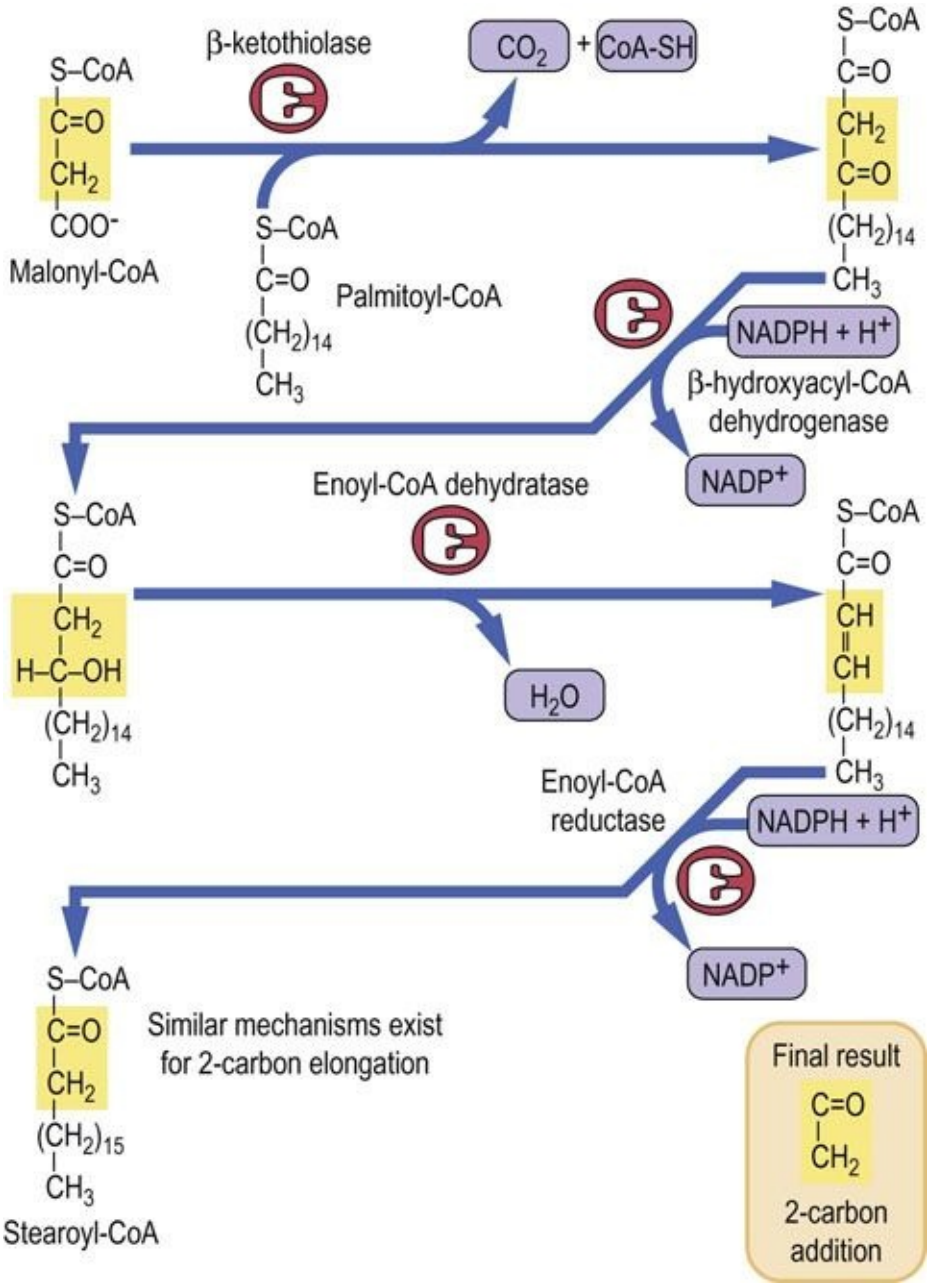


FIG. 16.5 Elongation of the fatty acids.
Fatty acid elongation occurs on the endoplasmic reticulum and is carried out by a multienzyme complex, fatty acid elongase.

The substrates for the cytosolic fatty acid elongase include saturated fatty acids with a chain length from 10-carbon upwards, and also unsaturated fatty acids. Very long-chain (22–24-carbon) fatty acids are produced in the brain, and elongation of stearoyl-CoA (C_{18}) in the brain increases rapidly during myelination, producing fatty acids required for the synthesis of sphingolipids.

Fatty acids can also be elongated in the mitochondria, where yet another system is used: it is NADH-dependent and uses acetyl-CoA as a source of two-carbon fragments. It is simply the reverse of β-oxidation (Chapter 15) and the substrates for chain elongation are short-and medium-chain fatty acids containing fewer than 16 carbon atoms. During fasting and starvation, elongation of fatty acids is greatly reduced.

Desaturation of fatty acids

Desaturation reactions require molecular oxygen

The body has a requirement for mono-and polyunsaturated fatty acids, in addition to saturated fatty acids. Some of these need to be supplied in the diet; these two unsaturated fatty acids, linoleic and linolenic, are known as the essential fatty acids (EFAs). The desaturation system requires molecular oxygen, NADH, and cytochrome b_5. The process of desaturation, like that of chain elongation, occurs on the endoplasmic reticulum and results in the oxidation of both the fatty acid and NADH (Fig. 16.6).

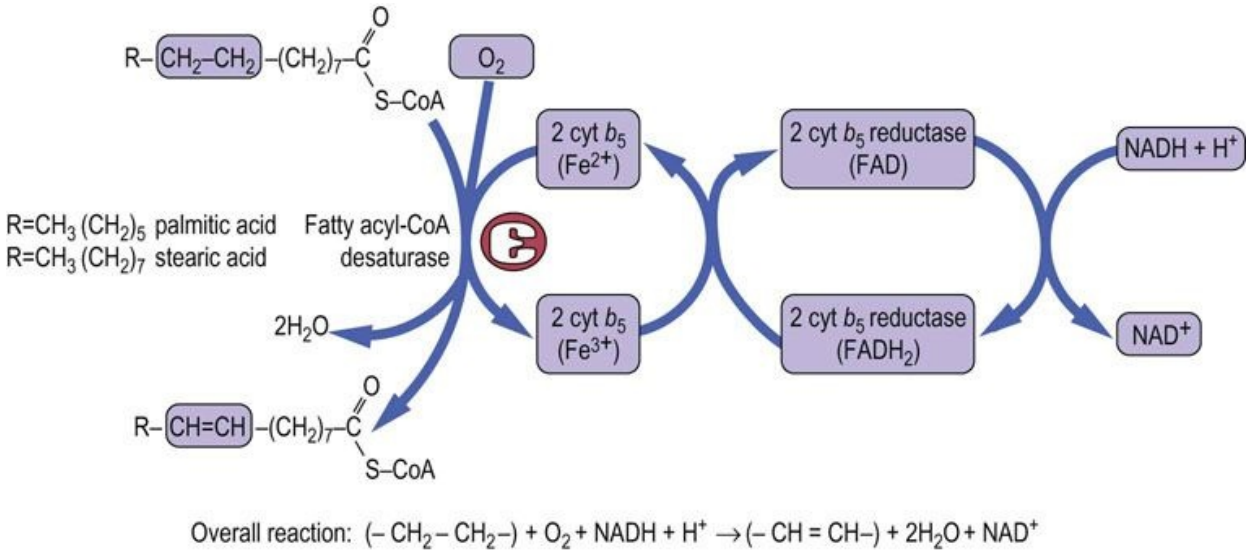


FIG. 16.6 Desaturation of fatty acids.
Desaturation of the fatty acids takes place in the endoplasmic reticulum. The reaction requires molecular oxygen, $NADH_2$, $FADH_2$ and cytochrome b_5. cyt b_5, cytochrome b_5; FAD, flavin adenine dinucleotide; $FADH_2$, reduced flavin adenine dinucleotide.

In man, the desaturase system is unable to introduce double bonds between carbon atoms beyond carbon-9 and the ω-(terminal methyl) carbon atom. Most desaturations occur between carbon atoms 9 and 10 (annotated as Δ^9 desaturations), *e.g.* those with palmitic acid producing palmitoleic acid (C-16:1, Δ^9), and those with stearic acid producing oleic acid (C-18:1, Δ^9). This step is

catalyzed by stearoyl-CoA desaturase (SCD).

Essential fatty acids

The ω-3 and ω-6 fatty acids (or their precursors) must be supplied with diet

As discussed above, the human desaturase is unable to introduce double bonds beyond C-9. On the other hand, two types of fatty acids – those having double bonds 3 carbons from the methyl end (**ω-3 fatty acids**) and 6 carbons from the methyl end (**ω-6 fatty acids**) – are required for the synthesis of eicosanoids (C-20 fatty acids), precursors of important molecules such as prostaglandins, thromboxanes and leukotrienes. Therefore, the ω-3 and ω-6 fatty acids (or their precursors) must be supplied in the diet. As it happens, they are obtained from dietary vegetable oils, and meat which contain the ω-6 fatty acid, linoleic acid (C-18:2, $\Delta^{9,12}$) and the ω-3 fatty acid, linolenic acid (C-18:3, $\Delta^{9,12,15}$). Linoleic acid is converted in a series of elongation and desaturation reactions to arachidonic acid (C-20:4, $\Delta^{5,8,11,14}$), the precursor for the synthesis of other eicosanoids in man. Elongation and desaturation of linolenic acid produce eicosapentaenoic acid (EPA; C-20:5, $\Delta^{5,8,11,14,17}$), which is a precursor of yet another series of eicosanoids (see Table 3.2). However, the elongation/desaturation of C-18:3, $\Delta^{9,12,15}$ to EPA occurs at a low rate; most of EPA in the human body derives from fish consumption.

Storage and transport of fatty acids: synthesis of triacylglycerols

Fatty acids derived from endogenous synthesis or from the diet are stored and transported as triacylglycerols

In both liver and adipose tissue, triacylglycerols (TAG) are produced by a pathway involving phosphatidic acid as an intermediate (Fig. 16.7). The source of glycerol phosphate is, however, different in the two tissues. **Glycerol** is the source of phosphatidic acid in the liver. However, in the adipose tissue, due to the lack of glycerol kinase, **glucose** is the indirect source of glycerol, with the glycolytic metabolite dihydroxyacetone phosphate being its immediate precursor.

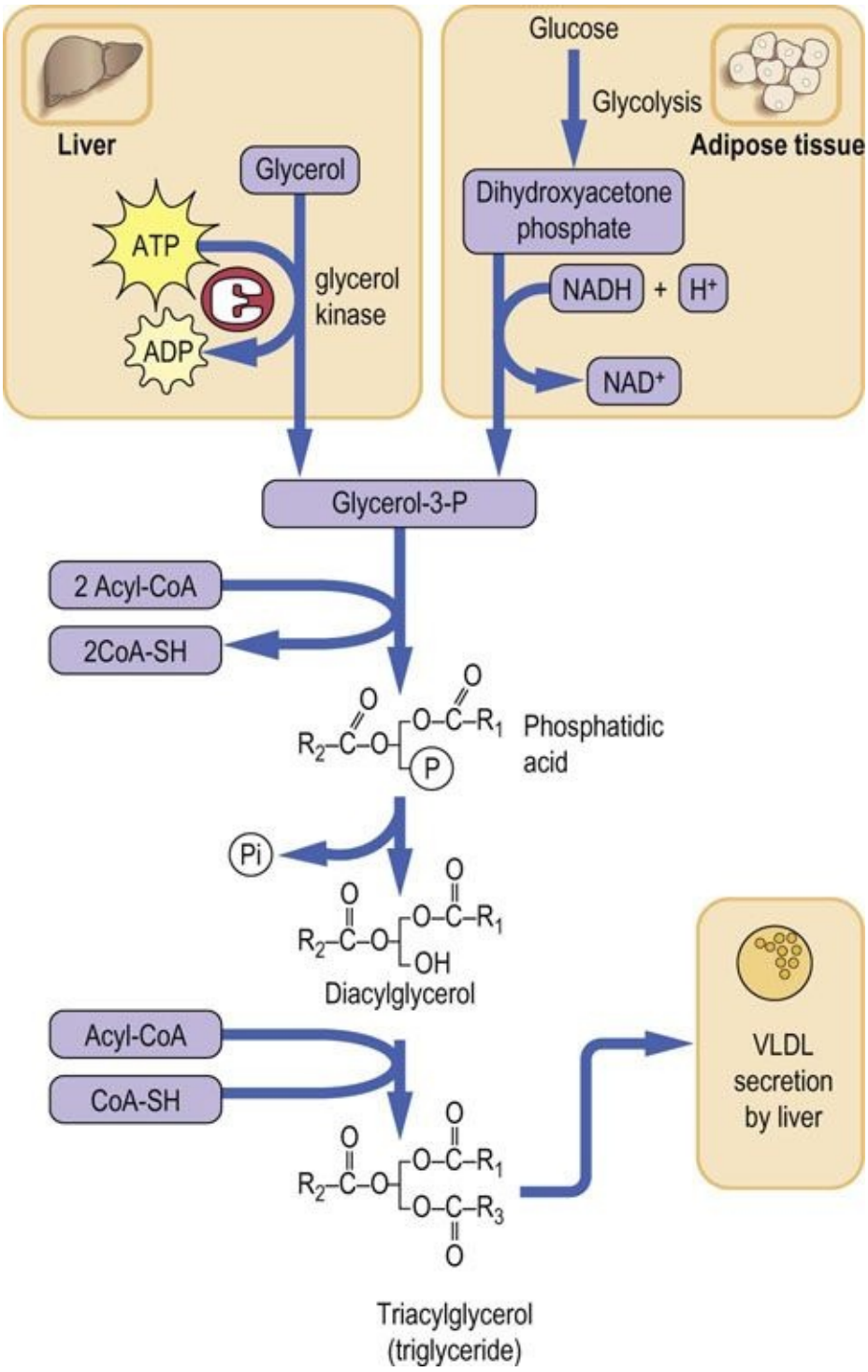


FIG. 16.7 Triacylglycerol synthesis.
Triacylglycerols (triglycerides) are synthesized in the liver and in adipose tissue. The source of glycerol-3-P is different in the two tissues. In the liver it is glycerol but adipose tissue has no glycerol kinase activity. There, glycerol-3-P is generated from the glycolytic intermediate, dihydroxyacetone phosphate. The central 'backbone' of the phosphatidic acid, diacylglycerol and triacylglycerol molecule shown in the figure consists of three

carbon atoms saturated with hydrogens (compare Fig. 10.8)

Triacylglycerols produced in the liver on the smooth endoplasmic reticulum can only be transiently stored

The liver has the unique capacity to off-load stored TAG by producing lipoprotein complexes with cholesterol, phospholipids and apolipoproteins (the latter also synthesized on the endoplasmic reticulum) for export in the form of very low-density lipoprotein (VLDL). The VLDL is then processed in the Golgi apparatus and released into the bloodstream for uptake by other tissues. In order to mobilize the transiently stored TAG, a lipolytic reaction occurs. (It results in the formation of diacylglycerols (DAG), which can then again enter the TAG synthetic pathway of VLDL assembly). The nature of this lipase is as yet not known, but is of significant medical interest due to the complications of fatty liver disease.

VLDL, once released into the bloodstream, is acted upon by lipoprotein lipase (LPL). This enzyme is found attached to the basement membrane glycoproteins of capillary endothelial cells and is active against both VLDL and chylomicrons (Chapter 18).

In the fed state, when adipose tissue is actively taking up fatty acids from the lipoproteins and storing them as TAG, the adipocytes synthesize LPL and secrete it into the capillaries of the adipose tissue. This increased synthesis and secretion of LPL is stimulated by insulin. Increased insulin levels also stimulate the uptake of glucose by adipose tissue and promote glycolysis. This has the net effect of producing increasing amounts of α-glycerophosphate, and it facilitates the synthesis of TAG within adipocytes. The skeletal muscle capillary bed also has LPL but it is inhibited by insulin. Instead, LPL is activated in skeletal muscle by its contractions.

Insulin is an important hormone in relation to fatty acid synthesis and storage. It promotes glucose uptake in both the liver and adipose tissue. In the liver, by increasing fructose-2,6-bisphosphate levels, it stimulates glycolysis, thus increasing pyruvate production. By stimulating dephosphorylation of pyruvate dehydrogenase complex and thus activating this enzyme, insulin promotes production of acetyl-CoA, stimulating the TCA cycle and increasing citrate levels, which, in turn, through stimulation of the acetyl-CoA carboxylase, increase the rate of fatty acid synthesis (see also Chapter 21).

Clinical box Lipid abnormalities in alcoholism

A 36-year-old woman attending a well-woman clinic was found to have serum concentrations of triglyceride 73.0 mmol/L (6388 mg/dL) and cholesterol 13 mmol/L (503 mg/dL). After some initial prevarication she admitted to drinking three bottles of vodka and six bottles of wine per week. When she discontinued alcohol, her triglyceride concentrations decreased to 2 mmol/L (175 mg/dL) and her cholesterol concentration decreased to 5.0 mmol/L (193 mg/dL). Three years later, the woman presented again with an enlarged liver and return of the lipid abnormality. Liver biopsy indicated alcoholic liver disease with steatosis (infiltration of the liver cells with fat).

Comment.

In alcoholic individuals, the metabolism of alcohol produces increased amounts of reduced hepatic nicotinamide adenine dinucleotide (NADH). Increased NADH + H^+/NAD^+ ratio inhibits the oxidation of fatty acids. Fatty acids reaching the liver either from dietary sources or by mobilization from adipose tissue are therefore re-esterified with glycerol to form triglycerides. In the initial stages of alcoholism, these are packaged with apolipoproteins and exported as very low-density lipoproteins (VLDLs). **Increased concentration of VLDL,** and hence of serum triglycerides, is often present in the early stages of alcoholic liver disease. As the liver disease progresses, there is a failure to produce the apolipoproteins and export the fat as VLDL; **accumulation of triglycerides in the liver cells** ensues.

Regulation of total body fat stores

Adipose tissue is an active endocrine organ

It has long been understood that increased energy intake without appropriate increase in energy expenditure is associated with obesity, which is characterized by increased adiposity, meaning both the number of adipocytes and their fat content. In this sense, the quantity of stored triacylglycerol is merely a consequence of energy balance. However, it is now clear that adipose tissue, far from being an inert storage reservoir, is hormonally active. Adipocytes produce hormones such as leptin, adiponectin and resistin (collectively known as adipokines), growth factors such as vascular endothelial growth factor, and proinflammatory cytokines such as tumor necrosis factor α (TNF-α) and interleukin 6 (IL-6). Such hormonal signals, in particular leptin, may alter energy balance. This is further discussed in Chapter 22.

Clinical box Lifestyle and obesity

A 48-year-old ex-infantryman (height 1.91 m) presented with the problem of increasing weight over the previous 8 years since leaving the Army. At the time of his retirement from active service, he had weighed 95 kg (209 lb) but at presentation weighed 193 kg (424.6 lb). His current occupation was that of truck driver. He denied any change in food intake since leaving the Army, but admitted to taking little or no exercise. Detailed enquiry indicated that his daily dietary intake provided between 12,600 and 16,800 kJ (3000 and 4000 kcal), with a fat intake approaching 40%. The patient was initially placed on a healthy eating plan, with fat intake reduced to 35% of total calories. He was advised to exercise and proceeded to swim three or four times per week. His weight immediately began to decrease, rapidly at first and then at 3–4 kg (6.6–8.8 lb) each month until it stabilized at 145–150 kg (319–330 lb). He was then placed on a high-protein/low-carbohydrate/low-fat diet, which induced a return of weight loss that continued for a further year, resulting in a final weight of 93 kg (204.6 lb).

Comment.

Obesity is increasingly prevalent in many parts of the world. **Clinical obesity is now clearly defined in terms of height and weight through the body mass index (BMI),** which is calculated as the weight in kilograms divided by the height in meters^2 (see Chapter 22 for details).

BMI 25–30 kg/m^2 is classified as overweight or grade I obesity, BMI >30 kg/m^2 is clinical or grade II obesity, and BMI >40 kg/m^2 is classified as morbid or grade III obesity. Our patient had a BMI of 53 at presentation, falling to 26 after prolonged diet. If energy input exceeds output over time, then weight will increase. **Obesity predisposes to several diseases.** The most important is type 2 diabetes mellitus: 80% of this type of diabetes is associated with the obese state. Other associated illnesses include coronary heart disease, hypertension, stroke, arthritis, and gall bladder disease.

Summary

■ Fatty acid synthesis and storage are essential components of body energy homeostasis.
■ Fatty acid synthesis takes place in the cytosol. Its committed step is the reaction catalyzed by acetyl-CoA carboxylase.
■ Elongation of the fatty acid chain (up to the length of 16 carbon atoms) is carried out by the dimeric fatty acid synthase, which possesses several enzyme activities. Both acetyl-CoA carboxylase and fatty acid synthase are subject to a complex regulation.
■ The malate shuttle facilitates the transfer of two-carbon units from the mitochondria to cytoplasm for use in fatty acid synthesis.
■ The reducing power for fatty acid synthesis in the form of NADPH is supplied by the pentose phosphate pathway and also by the malate shuttle.
■ The essential unsaturated fatty acids are linoleic and linolenic acid. Linoleic acid is converted to arachidonic acid, which in turn serves as the precursor of prostaglandins.
■ Adiposity signals are provided by adipokines, particularly leptin. Insulin is also important in the regulation of food intake.

Active learning

1. Describe how a growing fatty acid chain is transferred between the subunits of fatty acid synthase.
2. How are the eicosanoids synthesized?
3. Explain why the rate of lipolysis in the fed state is low.
4. Describe the committed step of fatty acid synthesis and its regulation.
5. What are the sources of acetyl-CoA for fatty acid synthesis?
6. Compare and contrast fatty acid synthesis and their oxidation.

Further reading

Brown, MS, Ye, J, Rawson, RB, Goldstein, JL. Regulated intramembrane proteolysis: a control mechanism conserved from bacteria to humans. *Cell*. 2000; 100:391–398.

Guillou, H, Zadravec, D, Martin, PG, Jacobsson, A. The key roles of elongases and desaturases in mammalian fatty acid metabolism: Insights from transgenic mice. *Prog Lipid Res*. 2010; 49:186–199.

Gurr MI, Harwood JLK, Frayn KN, eds. Lipid biochemistry: an introduction. Oxford: Blackwell Science, 2008.

CHAPTER 17

Biosynthesis of Cholesterol and Steroids

Marek H. Dominiczak, Graham Beastall and †A. Michael Wallace

Learning objectives

After reading this chapter you should be able to:

- List the main steps involved in the synthesis of the cholesterol molecule.
- Discuss the regulation of intracellular cholesterol concentration.
- Explain mechanisms governing cholesterol metabolism and excretion.
- Describe bile acids and their enterohepatic circulation.
- Outline the main pathways of synthesis of steroid hormones.

Introduction

Cholesterol is essential for cellular structure and function

Cholesterol is an essential component of mammalian **cell membranes**. It is also a precursor of important compounds such as the **steroid hormones**, **vitamin D**, and the **bile acids**. Humans synthesize approximately 1 g of cholesterol each day. The rate of endogenous cholesterol synthesis and its dietary intake determine its plasma concentration. Excess of dietary cholesterol reduces its endogenous synthesis.

A typical Western diet contains approximately 500 mg (1.2 mmol) of cholesterol daily, mainly in meat, eggs, and dairy products (Chapter 22). Under normal circumstances, 30–60% of this is absorbed during passage through the gut. Following intestinal absorption, cholesterol is transported to the liver and to peripheral tissues as a component of lipoprotein particles, the chylomicrons.

In the cell, all cholesterol is present within membranes. The regulatory events occur either within membranes or in their close vicinity

Cellular cholesterol synthesis, and cholesterol uptake from the extracellular fluid (primarily as a component of the low-density lipoprotein [LDL] particles) are precisely regulated at both translational and post-translational level – and the membrane concentration of the free cholesterol is the main regulator of cholesterol synthesis. The pathway of cholesterol synthesis in its early stages, provides substrates for the synthesis of compounds important for cell proliferation the electron transport and for combating oxidative stress. The key intermediates in cholesterol synthesis are mevalonate, farnesyl pyrophosphate, squalene and lanosterol (Fig. 17.1). The oxysterols generated at the later stages of the pathway also contribute to the regulation of cholesterol homeostasis. Finally, range of genetically controlled transporters regulate cholesterol absorption and its intestinal and biliary transport.

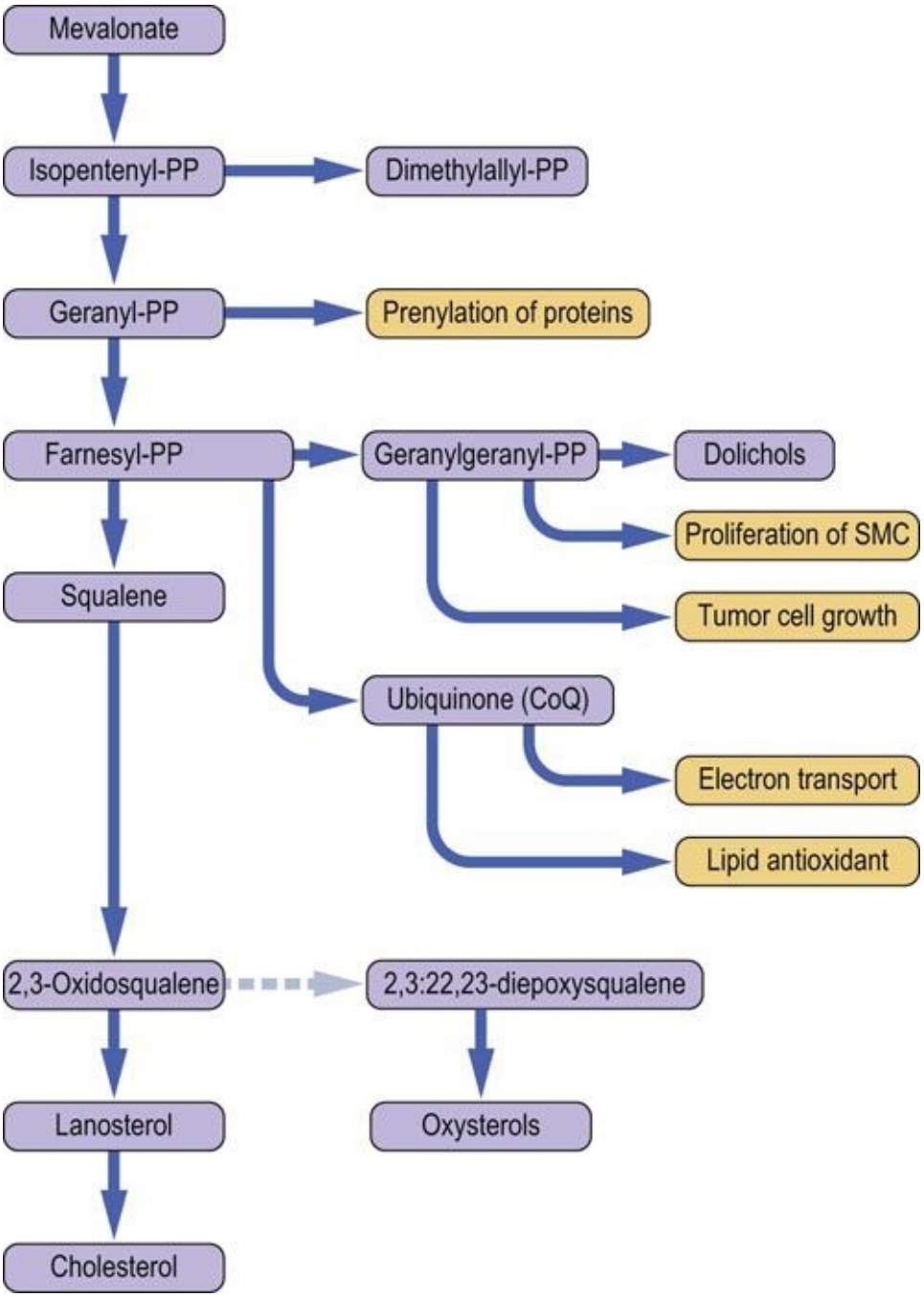


FIG. 17.1 Cholesterol synthesis and related pathways. The pathway of cholesterol synthesis is also a source of compounds that participate in a variety of cell functions. These are shown in orange-colored boxes. (modified from Charlton-Menys V, Durrington PN. Exp Physiol 2007; 93:27–42, with permission). SMC, smooth muscle cells; CoQ, coenzyme Q; PP, pyrophosphate.

The liver repackages cholesterol and triglycerides into very low-density lipoproteins (VLDL). After peripheral hydrolysis of VLDL by lipoprotein lipase, the VLDL remnants and LDL deliver cholesterol back to the liver. Chapter 18

describes lipoprotein metabolism in detail.

Humans cannot metabolize the sterol ring of cholesterol

Cholesterol is excreted by the liver in bile either in the form of bile acids or as free cholesterol. The so called primary bile acids are synthesized in the liver, and another ‘set’ known as the secondary bile acids is produced from them by the intestinal bacteria. Most bile acids are reabsorbed in the terminal ileum and recycled back to the liver.

Clinical significance

Excess supply of cholesterol is important in the etiology of atherosclerosis (Chapter 18). Cholesterol is also a major component of gallstones. Another cause of clinical disorders are the inherited deficiencies of enzymes that participate in the synthesis of steroid hormones from cholesterol. The latter are important in neonatal medicine.

The Cholesterol Molecule

The structure of cholesterol is shown in Figure 17.2. It has a molecular weight of 386 Da and contains 27 carbon atoms, of which 17 are incorporated into four fused rings (the cyclopentanoperhydrophenanthrene nucleus). Two further carbons are in the methyl groups at the junctions of rings AB and CD, and eight are in the side chain. Cholesterol is almost entirely composed of carbon and hydrogen atoms; there is a solitary hydroxyl group is attached to carbon 3. Cholesterol is also almost completely saturated structure, having just one double bond between carbon atoms 5 and 6.

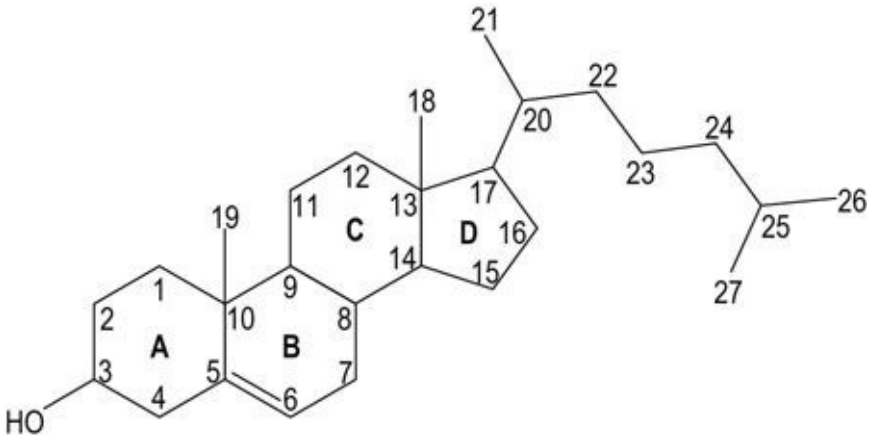


FIG. 17.2 Structure of cholesterol.
A–D is the conventional notation used to describe the four rings. Numbers 1–27 describe the carbon atoms.

Cholesterol changes membrane fluidity

Cholesterol is an essential component of cell membranes. It is held in the lipid bilayer by physical interactions between the planar steroid ring and the fatty acid chains. The absence of covalent bonding means that it may easily transfer in and out of the membrane. Membranes are fluid structures where both the lipid and protein molecules move and undergo conformational change (Chapter 8). The more fluid the phospholipid bilayer becomes, the more permeable is the membrane. At body temperature, the long hydrocarbon chains of the lipid

bilayer are capable of considerable motion. Cholesterol is located between these hydrocarbon chains. **Cholesterol increases fluidity of membranes** rich in phospholipids and sphingolipids that contain saturated fatty acids.

Cholesterol is clustered in regions within the lipid bilayer. In a such a cluster, there may be 1 mole of cholesterol per mole of phospholipid, while in adjacent areas there may be no cholesterol. Thus, the membrane contains cholesterol-rich impermeable patches and more permeable cholesterol-free regions. Different cell organelles may differ in cholesterol content by the factor of 10. It is, for instance, virtually absent from the inner mitochondrial membrane.

Free and esterified cholesterol

Cholesterol is poorly soluble in water. Only about 30% of circulating cholesterol occurs in the free form, the majority forms esters with long-chain fatty acids such as oleic and linoleic acids. Cholesteryl esters (CE) are even less soluble in water than free cholesterol.

Cholesterol is esterified in the plasma by the enzyme **cholesterol-lecithin acyltransferase** and in the cells by the **acyl-CoA : cholesterol acyltransferase (ACAT)**. In the plasma, it is incorporated into a range of lipoproteins (Chapter 18) where it is present mostly as CE. CE are stored in lipid droplets in the endoplasmic reticulum.

Intestinal absorption of cholesterol

Cholesterol is absorbed (and secreted) in the intestine by specific transporters

Dietary cholesterol is absorbed from the intestine via a membrane transporter known as the Nieman–Pick C1-like (NPC1L1) protein. Another transporter present in the apical (luminal) side of enterocytes is the ATP-binding cassette G5/G8, comprising two half-transporters – ABCG5 and ABCG8. These transport cholesterol in the 'reverse' direction, back to the intestine, and are also involved in secretion of other sterols into the bile. These transporters are upregulated by the nuclear transcription factor, the liver X receptor (see below). The genes coding for these transporters have a sterol response element in their promoter regions. Mutations in these genes result in tissue accumulation of plant sterols (**sitosterolemia**). The drug **ezetimibe** suppresses the NPC1L1-mediated cholesterol transport and has been used in the treatment of hypercholesterolemia.

Biosynthesis of cholesterol

Cholesterol is synthesized from acetyl coenzyme A

Virtually all human cells have the capacity to make cholesterol. Liver is the major site of cholesterol synthesis and smaller amounts are synthesized in the intestine, adrenal cortex and gonads. The synthesis of cholesterol molecule requires a source of carbon atoms, a source of reducing power and significant amounts of energy. The acetyl-coenzyme A (acetyl-CoA) provides a high-energy starting point. Acetyl-CoA may be provided from the β-oxidation of long-chain fatty acids, the dehydrogenation of pyruvate and the oxidation of ketogenic amino acids such as leucine and isoleucine. The reducing power is provided by reduced nicotinamide a dinucleotide phosphate (NADPH) generated in the pentose phosphate pathway (Chapter 12).

Energy is provided by the breakdown of adenosine triphosphate (ATP). Overall, the synthesis of 1 mole of cholesterol requires 18 moles of acetyl-CoA, 36 moles of ATP and 16 moles of NADPH. All the biosynthetic reactions occur within the cytoplasm, although some of the required enzymes are bound to the ER membranes.

Mevalonic acid is the first unique compound in the pathway of cholesterol synthesis

Three molecules of acetyl-CoA are converted into the 6-carbon mevalonic acid (Fig. 17.3). The first two steps occur in the cytoplasm and are condensation reactions leading to the formation of the 3-hydroxy-3-methylglutaryl-CoA (HMG-CoA). These reactions, catalyzed by acetoacetyl-CoA thiolase and HMG-CoA synthase, are the same as the ones in the synthesis of ketone bodies, although the latter process occurs within mitochondria.

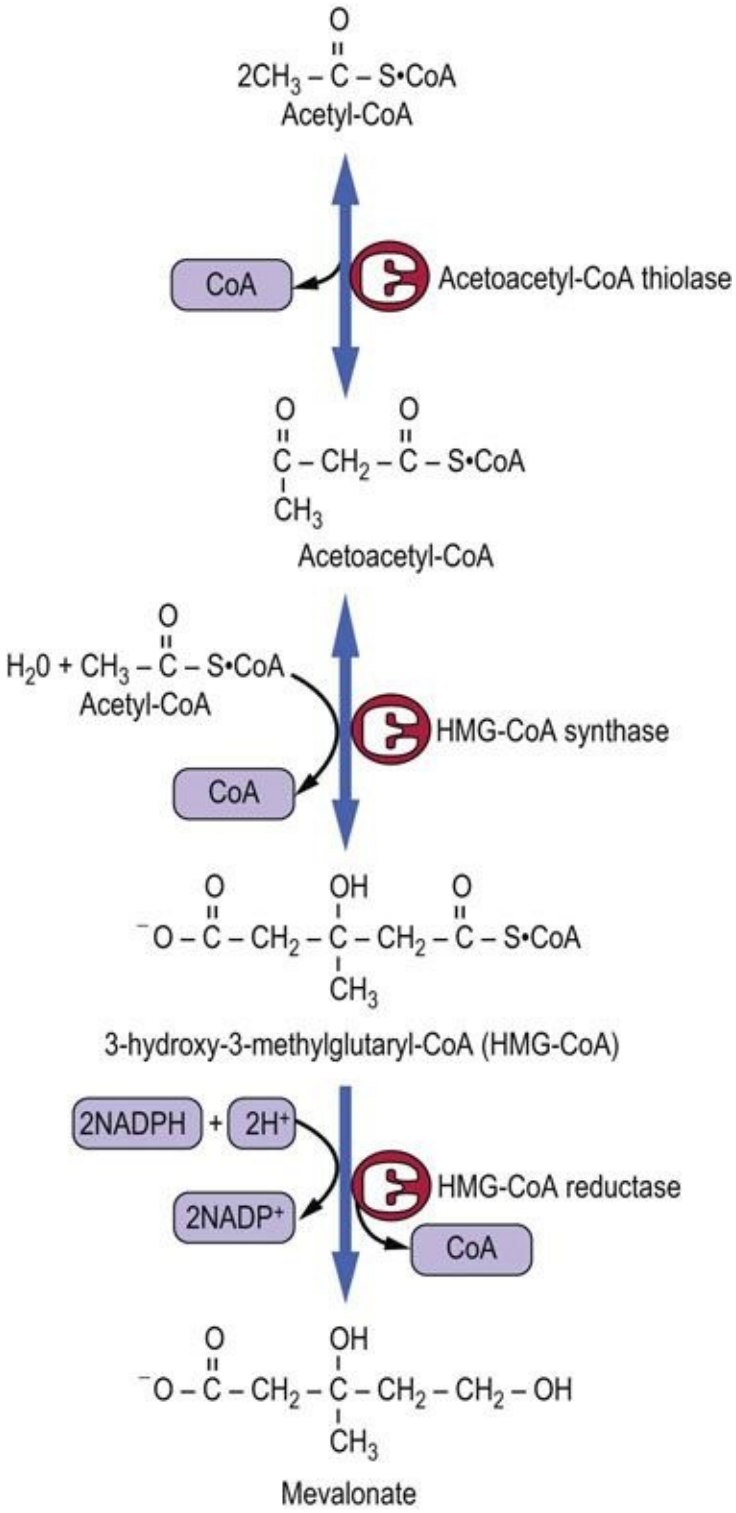


FIG. 17.3 Biosynthesis of the mevalonic acid.
Mevalonic acid contains 6 carbon atoms derived from the three molecules of acetyl-CoA.

HMG-CoA reductase is the rate-limiting enzyme in the pathway

The rate-limiting reaction in the pathway of cholesterol synthesis is that catalyzed by the HMG-CoA reductase (HMGR), which leads to the irreversible formation of mevalonic acid. The reaction uses two molecules of NADPH.

HMGR is embedded in the ER. It is controlled at multiple levels: by feedback inhibition, by the rate of its degradation, by phosphorylation (it is active in a nonphosphorylated state) and by changes in gene expression. It is also affected by several hormones: insulin and tri-iodothyronine increase its activity, while glucagon and cortisol inhibit it. HMGR can also be phosphorylated (and thus inhibited) by the 'energy sensor' enzyme, the AMP-dependent kinase (AMPK, Chapter 22). Oxysterols also regulate the HMGR.

Farnesyl pyrophosphate is made up of three isoprene units

Three molecules of mevalonic acid are phosphorylated in two reactions requiring kinases and ATP. Subsequent decarboxylation yields the **isomeric 5-carbon isoprene units**, isopentenyl pyrophosphate and dimethylallyl pyrophosphate, which condense together to form geranyl pyrophosphate. The geranyl pyrophosphate is enlarged to geranyl-geranyl pyrophosphate. Further condensation with isopentenyl pyrophosphate produces the 15-carbon atom molecule, farnesyl pyrophosphate (Fig. 17.4). As well as being an intermediate in cholesterol biosynthesis, farnesyl pyrophosphate is the branching point for the synthesis of dolichol (a substrate in glycoprotein synthesis) and ubiquinone (Fig. 17.1).

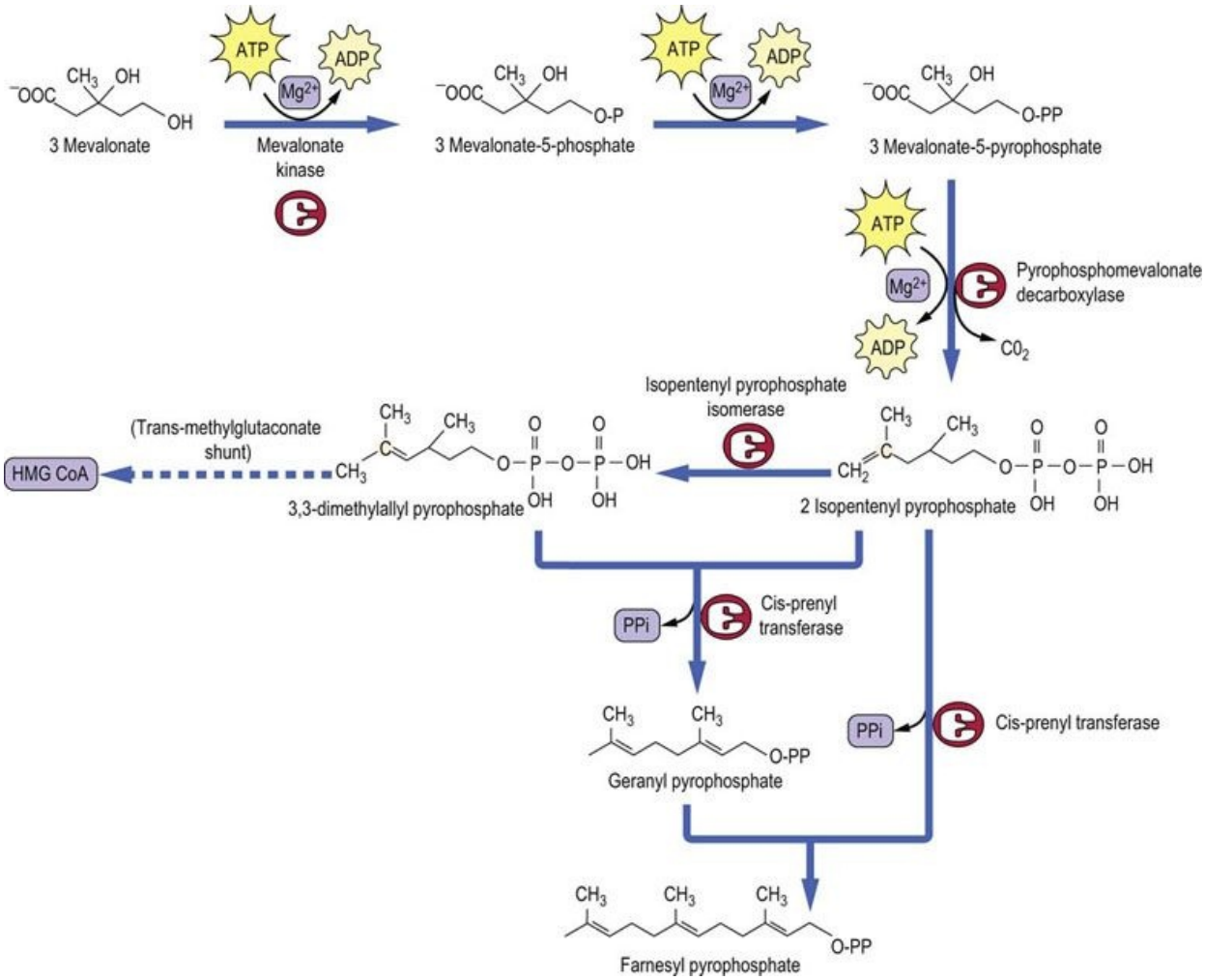


FIG. 17.4 Biosynthesis of farnesyl pyrophosphate.
Farnesyl pyrophosphate is made up of three isoprene units. Note that the isoprene units can be channelled into the trans-methylglutaconate shunt pathway, which 'recovers' the HMG-CoA (see texst for details). ADP, adenosine diphosphate.

Squalene is a linear molecule capable of a ring formation

Squalene synthase condenses two molecules of farnesyl pyrophosphate to form **squalene**, a 30-carbon hydrocarbon containing six double bonds (Fig. 17.5), which later enable it to fold into a ring similar to the steroid nucleus. Several intermediates are involved at this stage.

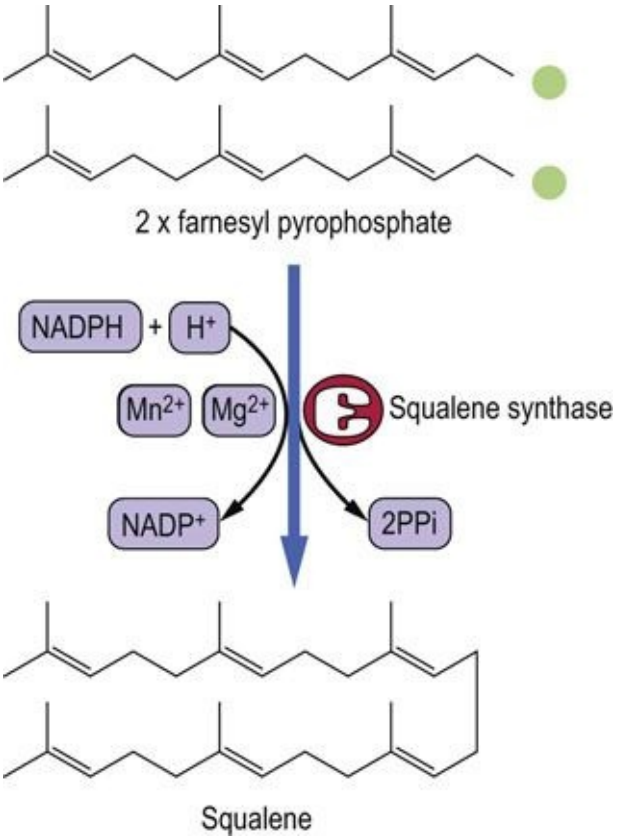


FIG. 17.5 Biosynthesis of squalene.
Squalene, a still linear molecule, results from the condensation of two molecules of farnesyl pyrophosphate. The six double bonds enable the squalene structure to fold into a ring later.

Squalene cyclizes to lanosterol

Before the closure of the ring, squalene is converted to squalene 2,3-oxide by squalene monooxygenase. This NADPH-dependent monooxygenase inserts oxygen molecule into the structure. Thereafter, cyclization is catalyzed by oxidosqualene cyclase, yielding lanosterol (Fig. 17.6).

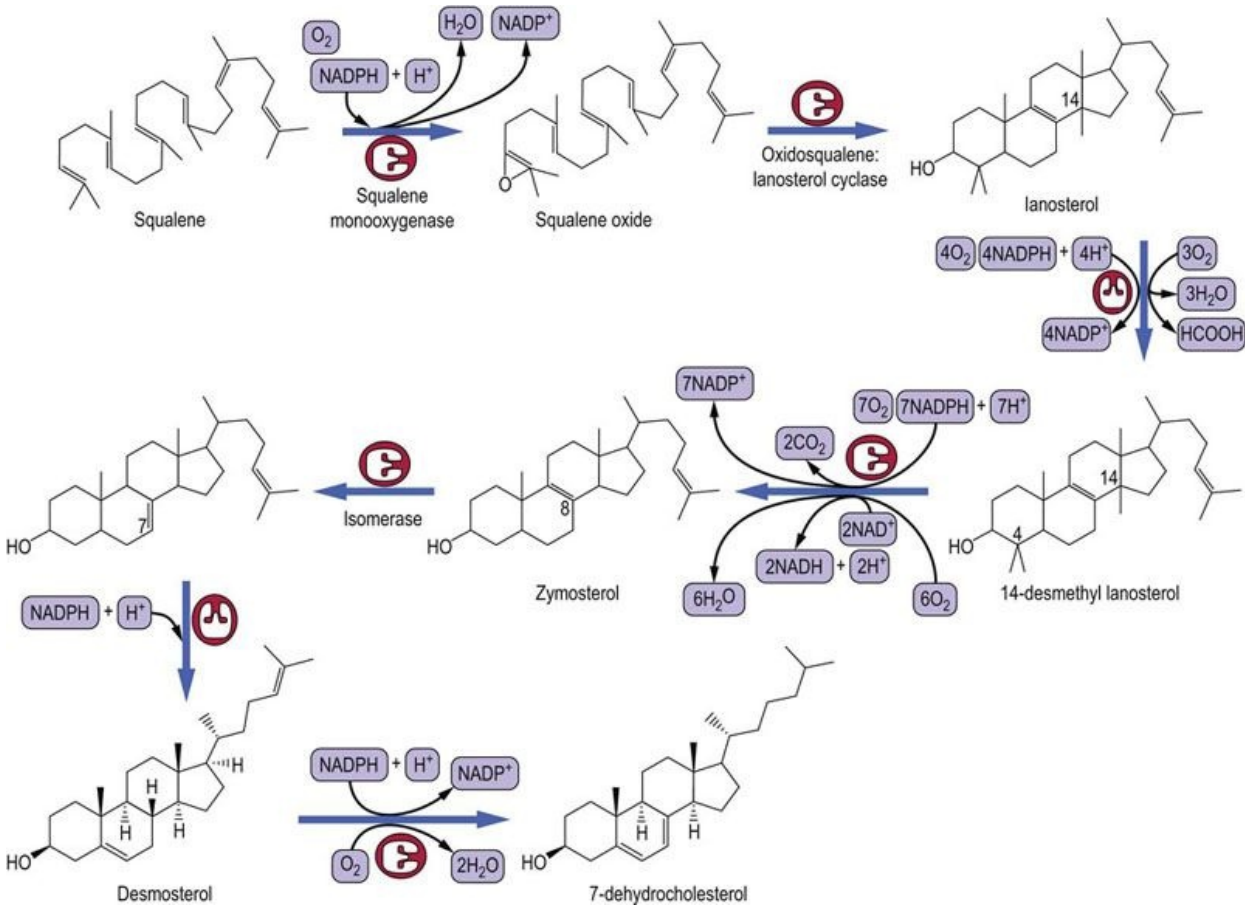


FIG. 17.6 Later stages of cholesterol synthesis.
These reactions occur while bound to a squalene-and sterol-binding proteins. FAD, flavin adenine dinucleotide; NADH, reduced nicotinamide adenine dinucleotide.

In plants, there is a different product of squalene cyclization, known as cycloartenol, which is further metabolized to a range of phytosterols, including sitosterol, rather than to cholesterol.

Final stages of cholesterol biosynthesis occur on a carrier protein

Squalene, lanosterol and all the subsequent intermediates in cholesterol synthesis are hydrophobic molecules. In order for the final steps of the pathway to occur in an aqueous medium, these intermediates react while bound to a squalene-and sterol-binding protein. The conversion from the 30-carbon lanosterol to the 27-carbon cholesterol involves decarboxylation reactions, isomerization and a reduction (Fig. 17.6).

Cholesterol can be converted to oxysterols by oxidation of its side chain. Oxysterols are important in cholesterol metabolism in the brain

This is performed by a cytochrome P450 enzyme, cholesterol 24-hydroxylase (CYP46A1, present in the brain) and 27-hydroxylase (CYP27A1, present in other tissues). The 27-hydroxycholesterol can cross the blood–brain barrier without the need for an energy-requiring transporter. The 25-hydroxycholesterol regulates the liver X receptors (LXRs). In the brain, also through LXR, it regulates the expression of apolipoprotein E (an important transporter of cholesterol in the brain) and the expression of transporters ABCA1, ABCG1 and ABCG4 present on astrocyte membranes.

Measurements of plant sterols and cholesterol precursors are used as markers of cholesterol absorption and metabolism

In the studies of cholesterol metabolism, the plant sterols campesterol, sitosterol and biliary sterol 5α-cholestanol have been used as markers of cholesterol absorption. On the other hand, the measurements of cholesterol precursors such as mevalonic acid, squalene or lanosterol have been used as markers of cholesterol synthesis.

Advanced concept box PCSK9 protease regulates degradation of LDL receptors

Serine protease PCSK9 (proprotein convertase subtilisin/kexin type 9) is a regulator of LDL receptors. PCSK9 is secreted from the liver, is present in plasma, and binds to the extracellular domain of the LDL receptor. After the LDL-receptor complex is internalized, PCSK9 prevents it from recycling to the membrane and channels it towards degradation. PSK9 overexpression in transgenic mice lowers the LDL receptor levels. In **hypercholesterolemic individuals** with a gain-of-function mutation, the PCSK9 has increased affinity for the LDL receptor. On the other hand, the loss-of-function mutation results in lower plasma cholesterol concentrations.

Regulation of cholesterol synthesis

At an organism level, the supply of cholesterol is either through diet or through de novo synthesis

Many factors are involved in the regulation of the intracellular concentration of cholesterol (Table 17.1). Under normal circumstances, there is an inverse relationship between dietary cholesterol intake and the rate of cholesterol biosynthesis. This ensures a relatively constant supply of cholesterol. It also explains why dietary restriction is only likely to achieve a moderate reduction in the plasma cholesterol concentration.

Table 17.1

Regulation of intracellular cholesterol concentration

Processes that increase the free cholesterol concentration
De novo synthesis
Hydrolysis of intracellular cholesteryl esters by cholesterol ester hydrolase
Dietary intake of cholesterol
Receptor-mediated uptake of LDL: upregulation of the LDL receptors
Processes that decrease intracellular free cholesterol concentration
Inhibition of de novo cholesterol synthesis
Downregulation of the LDL receptors
Esterification of cholesterol by acyl-coenzyme A : cholesterol acyl transferase
Release of cholesterol from the cell to high-density lipoproteins (HDL)
Conversion of cholesterol to bile acids or to steroid hormones
Factors that affect the activity of HMG-CoA reductase
Intracellular concentration of HMG-CoA
Membrane concentration of cholesterol
Hormones: insulin, tri-iodothyronine, glucagon, cortisol

The cell acquires cholesterol from both de novo synthesis

and the external supply

Note that the ‘external supply’ in the case of a cell does not necessarily equal dietary source. The exogenous cholesterol reaches cells predominantly within lipoproteins: as a component of chylomicron remnants, VLDL remnants, or LDL. These lipoproteins bind to the apoB/E receptors present on the plasma membranes and the lipoprotein/receptor complexes are taken up into the cell (Chapter 18). In the cytoplasm, vesicles carrying the internalized complexes are acted upon by lysosomal enzymes, which separate the LDL from the receptor and hydrolyze cholesterol esters. Free cholesterol is released to the membrane. The LDL apoprotein B is degraded.

Regulation of intracellular cholesterol concentration involves HMG-CoA reductase, LDL receptor, 7α-hydroxylase and a network of nuclear receptors

The two sources of cholesterol supply, synthesis de novo and external delivery by lipoproteins, are **reciprocally related**. The intracellular (intramembrane) cholesterol concentration is a key factor regulating both cellular cholesterol synthesis and the expression of LDL receptors. Thus, an increase in the free cholesterol concentration results in the following (Fig. 17.7):

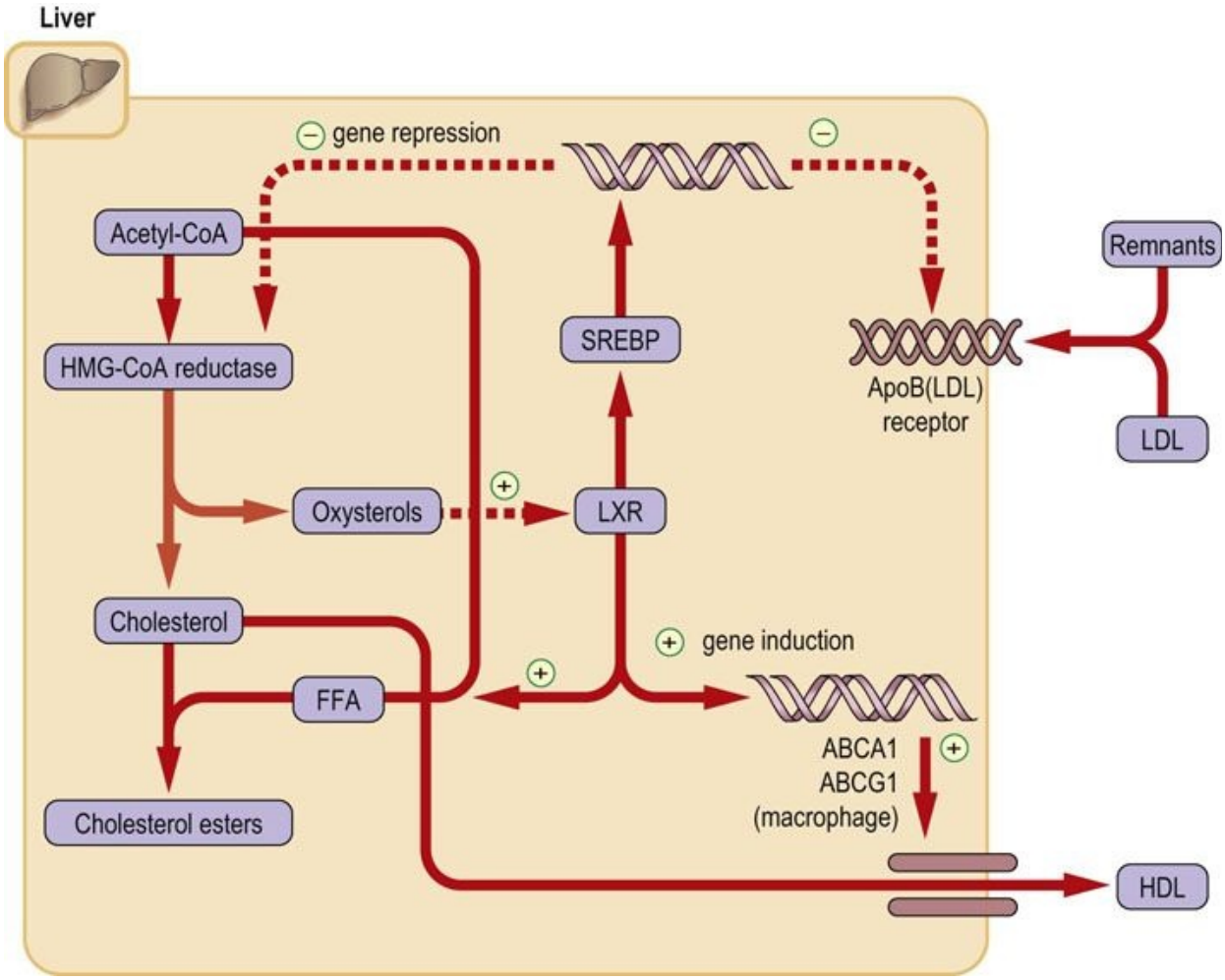


FIG. 17.7 Regulation of intracellular cholesterol concentration.
Free membrane cholesterol (and oxysterols) regulate intracellular cholesterol concentration by inducing or suppressing gene expression. Note that increase in intracellular cholesterol concentration will suppress synthesis of HMG-CoA reductase and ApoB/E receptor and at the same time increase cholesterol esterification and its transport from cells. See text for details. FFA: free fatty acid, Acetyl-CoA: acetyl coenzyme A, LXR: liver X receptor, SREBP: sterol regulatory element-binding protein.

■ A reduction in both the activity and expression of HMG-CoA reductase. This limits cholesterol synthesis.
■ Downregulation of LDL receptors. This limits the cellular uptake of cholesterol.
■ An increase in cholesterol and phospholipid efflux from cell to apoproteins A (i.e. the high-density lipoprotein [HDL]).
■ An increase in the rate of conversion of cholesterol to bile acids. This increases cholesterol excretion.

Advanced concept box Trans-methylglutaconate shunt

The dimethylallyl pyrophosphate, one of the **isoprene units** formed from mevalonate (Fig. 17.4), can be dephosphorylated and broken down into **acetoacetate** and **acetyl-CoA**, which may then be diverted into other pathways, such as fatty acid biosynthesis. This mechanism is known as the trans-methylglutaconate shunt. Thus, high-energy compounds once destined to be converted into cholesterol may be redeployed to meet a higher priority need. Note that an increase in fatty acid synthesis will increase the amount of substrate for cholesterol esterification.

Sterol regulatory element-binding proteins (SREBPs) regulate the genes coding for enzymes involved in cholesterol synthesis

SREBPs are synthesized as 120 kDa inactive precursors, which are integral part of the ER membrane. They bind to the ER protein known as the SREBP cleavage-activating protein (SCAP). The SCAP/SREBP complex transfers from the ER to the Golgi apparatus, where SREBPs are cleaved by a protease, releasing the active transcription factors, which in turn translocate to the nucleus and activate all the genes in the cholesterol synthetic pathway (Fig. 17.7). This process is subject to ingenuous regulatory mechanism.

Cholesterol molecules bind to sterol-sensing intermembrane domains (cholesterol ‘receptors’) present on the SCAP protein. This allows binding of the SCAP/SREBP complex to another ER protein, Insig-1 (Insig stands for ‘insulin-induced gene’). The stability of the SCAP/SREBP/Insig-1 complex is a key regulatory event. This is how it works:

When cholesterol is depleted, the SCAP/SREBP complex dissociates from Insig-1 and travels to the Golgi apparatus. However, **when cholesterol concentration in the membrane is high**, cholesterol binding to SCAP induces conformational change which stabilizes the SCAP/SREBP/Insig-1 complex, blocking its movement to the Golgi. Consequently, there is a decrease in nuclear

SREBP and the transcription of genes associated with cholesterol synthesis remains repressed. The synthesis of cholesterol is inhibited. Cholesterol excess also decreases the level of Insig-1 mRNA (Fig. 17.8).

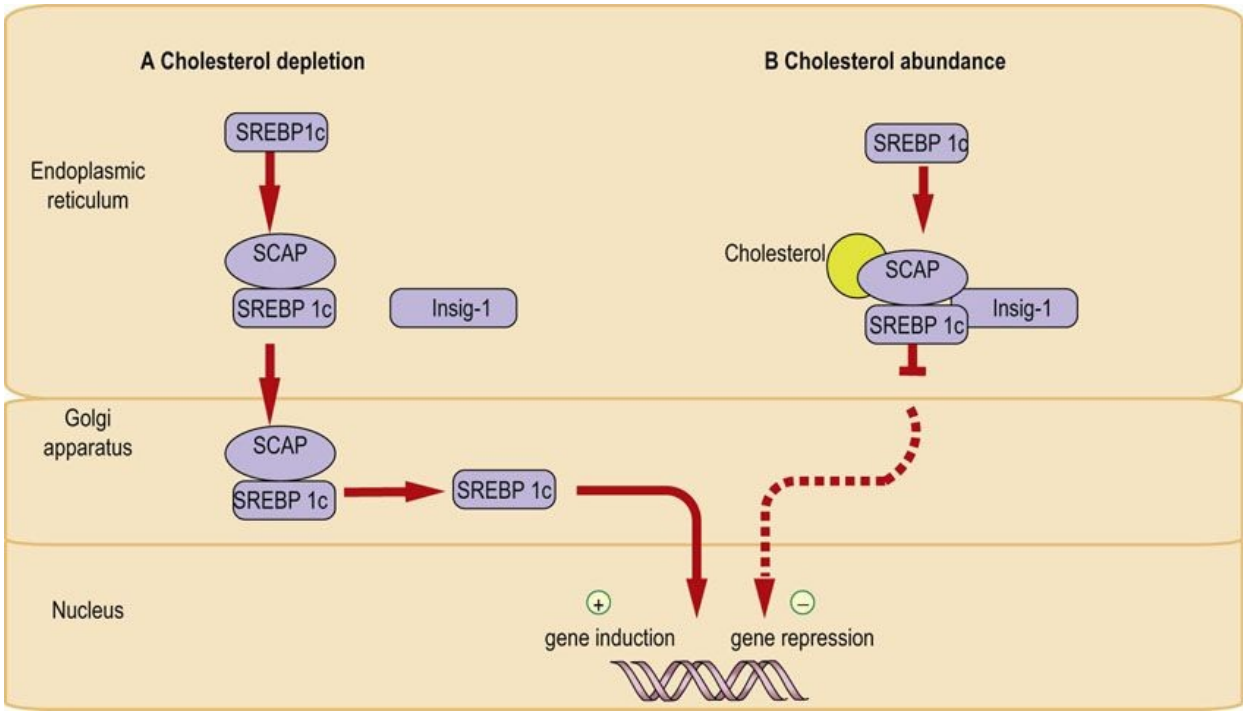


FIG. 17.8 Control of gene transcription by sterol-regulated transcription factors (SREBPs).
(**A**) When the membrane concentration of free cholesterol is low, the concentration SCAP/SREBP complex transfers from the ER to Golgi apparatus. Proteolysis takes place and the active transcription factor enters nucleus and initiates gene transcription. (**B**) When the membrane concentration of free cholesterol is high, cholesterol-binding-induced conformational change in the SCAP protein stabilizes its binding to the Insig-1. The complex remains in the ER with SREBP in inactive form, and gene transcription is repressed. See text for details. ER, endoplasmic reticulum; SREBP, sterol regulatory element binding protein; SCAP, SREBP-cleavage-activating protein.

When cells become again deprived of cholesterol, SCAP/SREBP dissociates from Insig-1, and the complex is again free to move to the Golgi apparatus, restoring nuclear SREBP level and restarting cholesterol synthesis. The SREBPs also induce Insig-1 synthesis.

HMG-CoA reductase is regulated by cholesterol in a mechanism involving its degradation

HMGR also possesses the sterol-sensing domain. When cholesterol level is high, HMGR, similarly to the SCAP/SREBP complex, binds to the Insig-1 protein. Here, however, the effect of such binding is different: it increases the ubiquination of the enzyme and channels it towards degradation. The overall effect is inhibition of cholesterol synthesis.

SREBPs have wide-ranging effects on the synthesis of cholesterol and fatty acids

In addition to the effect on cholesterol synthesis, SREBPs **increase the expression of LDL receptor gene** and affect **fatty acid synthesis**. In mammals, there are two closely related SREBPs: SREBP 1a and 1c. They are produced by the same gene, by alternative splicing. SREBP2 regulates cholesterol synthesis and LDL-receptor gene expression, while SREBP1c controls fatty acid synthesis. SREBP 1a induces all SREBP responsive genes.

SREBP1c can be activated by liver X receptors

SREBP 1c is upregulated by the liver X receptors (LXRs). LXRs are **ligand-activated transcription factors** that are members of the nuclear receptor superfamily (Chapter 40). They form **heterodimers** with other similar molecules, such as the **retinoid X receptors** (RXRs) and the **farnesyl X receptors** (FXRs). Resultant complexes bind to the LXR response elements on the DNA, regulating gene expression. LXRs also sense intracellular cholesterol concentration and contribute to regulating both its synthesis and its efflux from cells. However, it is not the cholesterol that binds to the LXR but oxysterols, such as 25-hydroxycholesterol or 27-hydroxycholesterol (Fig. 17.1).

SREBP 1c regulates cholesterol efflux from cells

High concentration of cholesterol in the hepatocyte induces, also through LXR-SREBP1c mechanism, genes coding for cholesterol transporters that control its efflux from cells to HDL particles: the expression of ABCA1 (the transporter that controls efflux of cholesterol from cells to nascent HDL) and ABCG1 (the transporter that stimulates efflux of cholesterol to more mature HDL2 and HDL3). Note that another transcription factor, PPARα, also regulates cholesterol efflux, acting through LXR (Chapter 18). PPARa is affected by a group of lipid-lowering drugs known as **fibrates** (**derivatives of** the fibric acid).

SREBP 1c regulates fatty acid synthesis

High intracellular cholesterol concentration also induces (through SREBP1c) genes coding for all the enzymes catalyzing fatty acid synthesis. The increased supply of fatty acids provides substrates for cholesterol esterification. Thus an increase in cholesterol concentration increases the concentration of oxysterols, and this in turn stimulates fatty acid synthesis, providing the substrate for cholesterol esterification. SREBP1c also induces genes involved in NADPH generation and the synthesis of triglycerides and phospholipids.

Statins are drugs that inhibit HMG-CoA reductase

The HMGR inhibitors, known as **statins**, lower cholesterol by binding to the site of HMG-CoA binding on the enzyme and competitively inhibiting its activity. This results in a decrease in the intracellular cholesterol concentration. The decrease in free cholesterol stimulates the expression of LDL receptors. LDL clearance increases and the plasma LDL-cholesterol decreases. The hepatic HMGR exhibits a diurnal rhythm: its activity is at a peak about 6 hours after dark and at a minimum some 6 hours after exposure to light. Therefore, statins are usually taken at night to ensure maximum effect.

Clinical box A 50-year-old man with hypercholesterolemia treated with statin

Despite a strict low cholesterol diet, a 50-year-old man, who had a family history of early cardiovascular disease, had a serum cholesterol concentration of 8.0 mmol/L (309 mg/dL); desirable concentration is 4.0 mmol/L (<155 mg/dL). He also smoked 15 cigarettes per day. He was given smoking cessation advice and was prescribed a statin. He tolerated the therapy well and 3 months later his cholesterol was 5.5 mmol/L (212 mg/dL). The dose of the statin was increased and after a further 3 months his plasma cholesterol concentration was 4.1 mmol/L (158 mg/dL).

Comment.

Partial inhibition of the HMGR brings about a lowering of total plasma cholesterol by 30–50% and LDL-cholesterol by 30–60%. A range of statins is now available: this follows the original

discovery that compactin (later renamed mevastatin), a fungal metabolite isolated from *Penicillium citrinum*, had HMGR-inhibiting properties. The inhibition of HMGR activity leads to a decrease in intracellular free cholesterol concentration, and to consequent increased expression of the cell membrane LDL receptor (Chapter 18). The end result is the lowering of the plasma total cholesterol and LDL-cholesterol.

Bile acids

The liver removes cholesterol either in a free form or as bile acids

The bile acids are quantitatively the most abundant metabolic products of cholesterol. In man, there are four main bile acids (Fig. 17.9). They all possess 24 carbon atoms: the terminal three carbons of the cholesterol side chain are removed during synthesis. They also have a saturated steroid nucleus and differ only in the number and position of the additional hydroxyl groups. All these hydroxyl groups have the α-configuration (i.e. are located below the plane of the nucleus), This means that the isomerization of the 3β-hydroxyl group of cholesterol must occur.

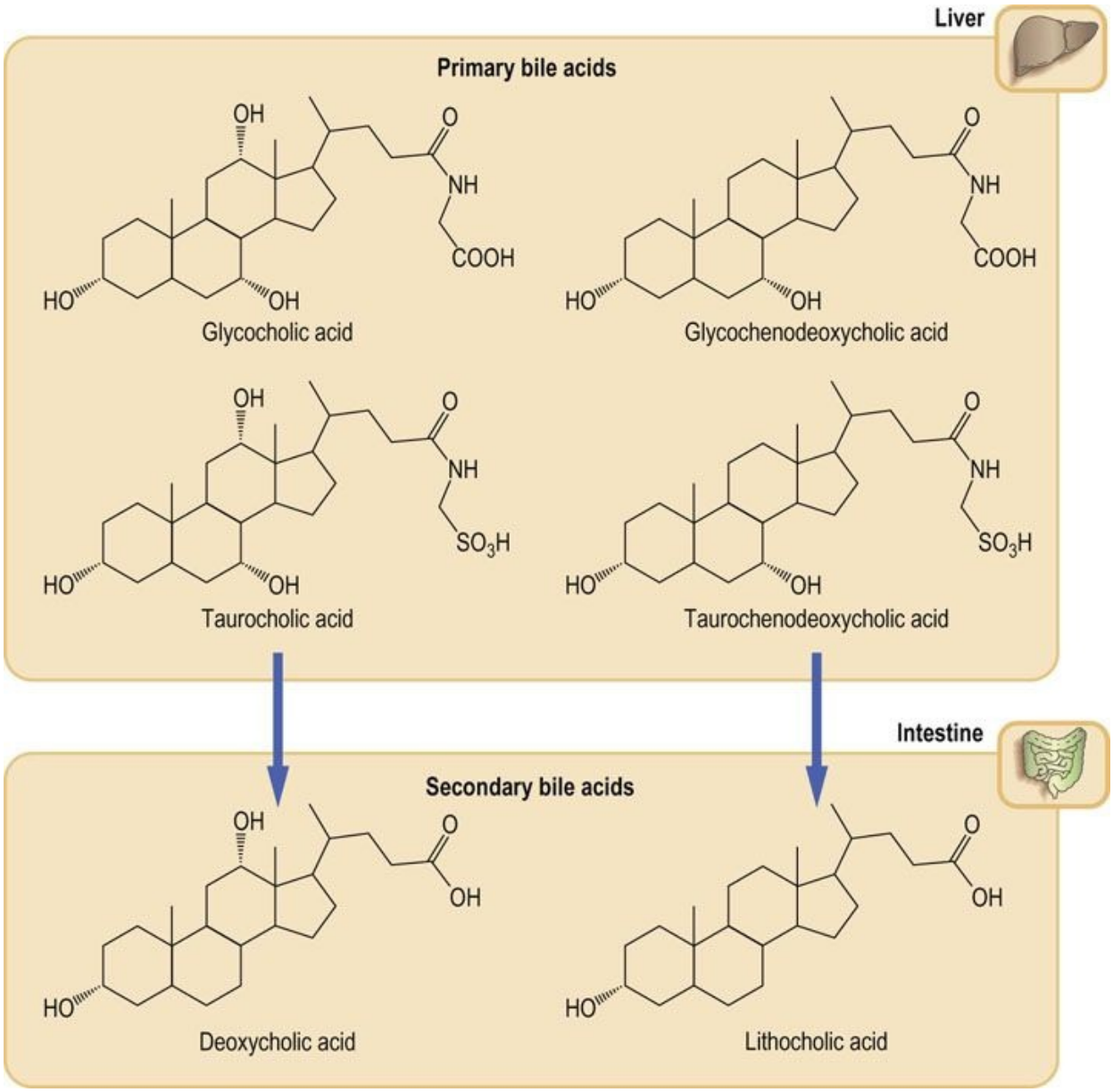


FIG. 17.9 Structure of the bile acids.
Primary bile acids are synthesized in the liver. They are converted by the intestinal bacteria to secondary bile acids.

Primary bile acids are synthesized in the liver

Biosynthesis of the bile acids occurs in liver parenchymal cells, where **cholic** and **chenodeoxycholic acids** are produced. They are known as the primary bile acids. The rate-limiting step in the biosynthesis is a reaction catalyzed by the the microsomal **7α-hydroxylase** (designated also CYP7A1), which introduces a

hydroxyl group at 7α position of cholesterol ring. It is another microsomal monooxygenase (compare the ones involved in cholesterol synthesis).

Prior to their secretion, the primary bile acids are **conjugated** through the carboxyl group, forming amide linkages with either **glycine** or **taurine**. In man, there is a 3 : 1 ratio in favor of glycine conjugates. The secreted products are thus principally glycocholic, glycochenodeoxycholic, taurocholic and taurochenodeoxycholic acids. At physiologic pH, the bile acids are mainly ionized, therefore they are present as sodium or potassium salts. The terms 'bile acids' and 'bile salts' are used interchangeably. As their name suggests, these compounds are either directly secreted into the duodenum or are stored in the gall bladder. They form an important component of bile, together with water, phospholipids, cholesterol and excretory products such as bilirubin.

Liver X receptors participate in bile synthesis and secretion

Cholesterol is pumped into bile by **ABCG5 and ABCG8 transport proteins**, the expression of which is regulated by the LXR. Importantly, bile supersaturated with cholesterol facilitates formation of cholesterol gallstones. The LXR coordinate expression of several genes relevant to cholesterol excretion, including the cholesterol 7α-hydroxylase. Cholesterol excretion into the bile is also regulated by other nuclear receptors: the farnesyl X receptor (FXR) heterodimerizes with retinoic X receptor and binds to the bile acid response elements on the DNA. FXR acts as the cellular bile acid sensor by binding bile acids and suppressing their synthesis. FXR also induces bile acid export pump ABCB11, which removes bile acids from the hepatocyte into the bile.

Secondary bile acids are synthesized in the intestine

Secondary bile acids form within the intestine through the action of the anaerobic bacteria (principally *Bacteroides*) on the primary bile acids. They are deoxycholic and lithocholic acids (Fig. 17.9). Only a proportion of primary bile acids is converted into secondary bile acids. This requires hydrolysis of the amide link to glycine or taurine prior to removal of the 7α-hydroxyl group.

Bile acids assist the digestion of dietary fat

The secretion of bile from the liver and the emptying of the gallbladder are

controlled by the gastrointestinal hormones hepatocrinin and cholecystokinin, respectively. They are released when partially digested food passes from the stomach to the duodenum. Once secreted into the intestine, the bile acids act as detergents (they possess polar carboxyl and hydroxyl groups), assisting the emulsification of ingested lipids; this aids the enzymatic digestion and absorption of dietary fat (Chapter 10).

Bile acids recirculate via the enterohepatic circulation

Up to 30 g of bile acids pass from the bile duct into the intestine each day but only 2% of this (approximately 0.5 g) is lost with the feces. Most are deconjugated and reabsorbed. Their passive reabsorption occurs in the jejunum and colon but they are mostly taken up by active transport in the ileum. Reabsorbed bile acids are transported back to the liver via the portal vein, being noncovalently bound to albumin, and are re-secreted into the bile. The process is known as the enterohepatic circulation. This recirculation also explains why bile contains both primary and secondary bile acids. The total bile acid pool is only 3 g and therefore they have to recirculate 5–10 times a day.

The bile acid flux also contributes to the control of bile acid synthesis; 7α-hydroxylase is subject to feedback inhibition by the bile acids returning to the liver through the portal vein. Dietary bile acids also decrease the expression of 7α-hydroxylase. Bile acid metabolism is summarized in Figure 17.10.

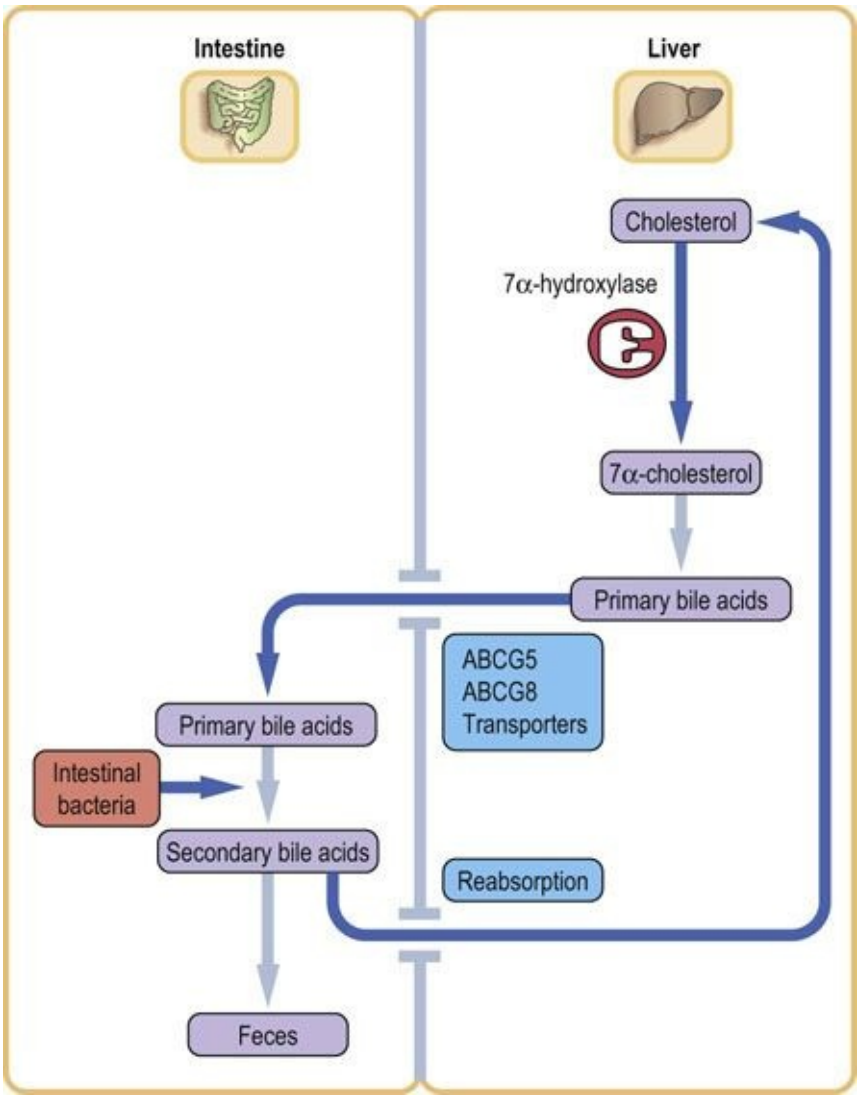


FIG. 17.10 The enterohepatic circulation of bile acids.

Cholesterol is excreted in the feces

Approximately 1 g of cholesterol is eliminated from the body each day through the feces: 50% of this is excreted as bile acids and the remainder as the isomeric saturated neutral sterols coprostanol (5β-) and cholestanol (5α-) produced by bacterial reduction of the cholesterol molecule.

Cholestyramine is a bile-acid binding resin which has been used to lower plasma cholesterol

Cholestyramine is a drug that interrupts enterohepatic circulation of the bile

acids. It leads to an increase in 7α-hydroxylase activity, and thus to increased bile acid synthesis, and their increased excretion. This in turn leads to an increased cellular cholesterol synthesis and to increased expression of LDL receptors. Cholestyramine was one of the first cholesterol-lowering drugs but now has been superseded by the statins.

Clinical box A 45-year-old woman admitted with abdominal pain and vomiting: gallstones

A 50 years' old woman woman complained of right upper quadrant abdominal pain and vomiting after eating fatty food. The only biochemical abnormality was a modestly raised alkaline phosphatase at 400 U/L (the upper reference limit is 260 U/L). An abdominal ultrasound was performed and it showed that the gall bladder contained gallstones. She was referred to the surgeons.

Comment.

Gallstones occur in up to 20% of the population of Western countries. The condition results from the formation of cholesterol-rich stones within the gall bladder. Cholesterol is present in high concentrations in bile, being solubilized in micelles that also contain phospholipids and bile acids. When the liver secretes bile with a cholesterol-to-phospholipid ratio greater than 1 : 1, it is difficult to solubilize all the cholesterol in the micelles; thus there is a tendency for the excess to crystallize around any insoluble nuclei. This is compounded by further concentration of the bile in the gall bladder, which occurs as a result of reabsorption of water and electrolytes.

The condition may be managed conservatively by reducing dietary cholesterol, and by increasing availability of bile acids that will assist with cholesterol solubilization in the bile and its excretion via the gut. Alternative treatment includes disintegration of stones by shock waves (lithotripsy) and surgery. **Elevated alkaline phosphatase is a marker of cholestasis** (Chapter 30).

Steroid hormones

Cholesterol is the precursor of all steroid hormones

Mammals produce a wide range of steroid hormones, some of which differ only by a double bond or by the orientation of a hydroxyl group. Consequently, it has been necessary to develop a systematic nomenclature to detail their exact structures. There are three groups of steroid hormones (Fig. 17.11). The **corticosteroids** have 21 carbon atoms in the basic pregnane ring. The loss of the two carbon atoms from the cholesterol side chain produces the androstane ring and the hormones known as the **androgens**. Finally, the loss of the methyl group at carbon atom 19 as part of the aromatization of the A ring results in the estrane structure found in the **estrogens**. The presence and position of double bonds, and the position and orientation of the functional groups on the basic nucleus, are characteristics of individual hormones.

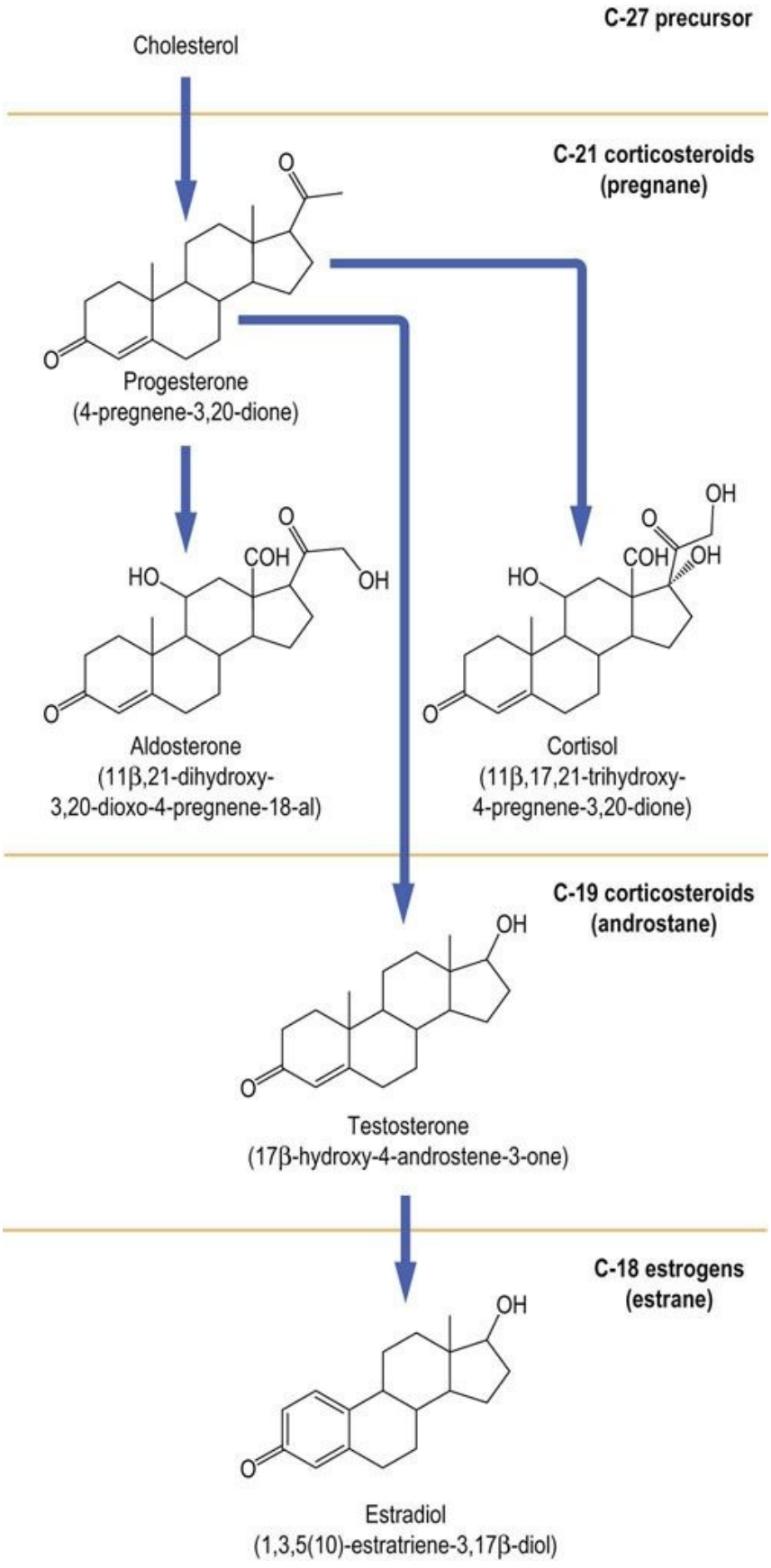

C-27 precursor
Cholesterol
C-21 corticosteroids
(pregnane)
O
O
Progesterone
(4-pregnene-3,20-dione)
O
COH
HO
OH
O
Aldosterone
(11β,21-dihydroxy-
3,20-dioxo-4-pregnene-18-al)
OH
O
COH
OH
HO
O
Cortisol
(11β,17,21-trihydroxy-
4-pregnene-3,20-dione)
C-19 corticosteroids
(androstane)
OH
O
Testosterone
(17β-hydroxy-4-androstene-3-one)
C-18 estrogens
(estrane)
OH
O
Estradiol
(1,3,5(10)-estratriene-3,17β-diol)

FIG. 17.11 Structure and nomenclature of the most important human steroid hormones. Their trivial and systematic names (in parentheses) are shown. For numbering of the atoms in a steroid molecule, see Fig. 17.1; see also Chapter 39.

Biosynthesis of the steroid hormones

Conversion of cholesterol into steroid hormones occurs in only three organs: the adrenal cortex, the testis in men, and the ovary in women

A simplification used in practice is to consider the corticosteroids as the products of the adrenal cortex, the androgens as the products of the testis and the estrogens as the products of the ovary. Such simplified pathway of steroid synthesis is shown in Figure 17.12 (see also Chapter 39). However, this is not absolute, and all three organs are capable of secreting small amounts of steroids belonging to other groups. In pathologic situations, such as a defect in steroidogenesis or a steroid-secreting tumor, a very abnormal pattern of steroid secretion may emerge.

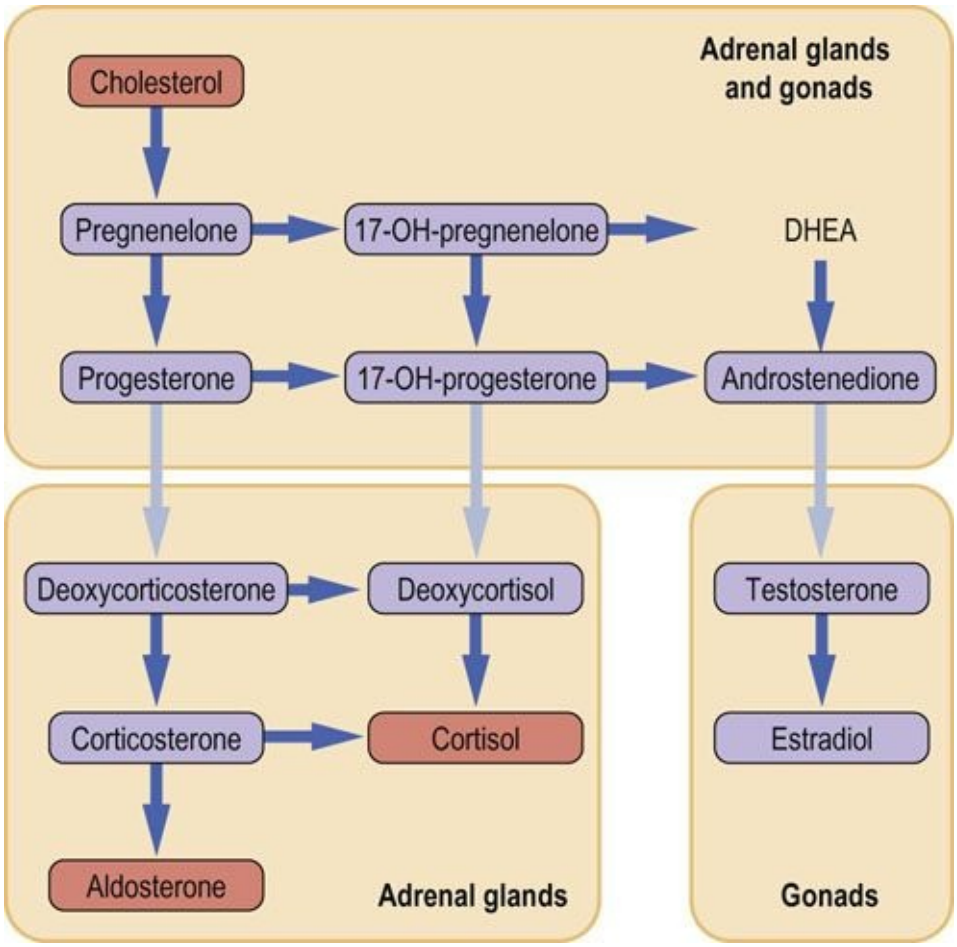


FIG. 17.12 Steroid biosynthetic pathway.
Note how the pathway branches from cholesterol, eventually leading to synthesis of mineralocorticoids (e.g. aldosterone), glucocorticoids (cortisol), androgens (testosterone) and estrogens (estradiol). DHEA, dehydroepiandrosterone.

Cytochrome P450 monooxygenases control steroidogenesis

Most of the enzymes involved in converting cholesterol into steroid hormones are cytochrome P450 proteins that require oxygen and NADPH. In its simplest form, such an enzyme complex catalyzes the replacement of a carbon–hydrogen bond with a carbon–hydroxyl bond; hence, the term monooxygenase (compare cholesterol synthesis pathway above). Hydroxylation of the adjacent carbon atoms is the forerunner to cleavage of the carbon–carbon bond. Comparison of the structure of cholesterol (Fig. 17.2) with those of the steroid hormones (Fig. 17.11) demonstrates that the biosynthetic pathway largely consists of cleavage of

carbon–carbon bonds and hydroxylation reactions. The enzymes involved have their own nomenclature in which the symbol CYP is followed by a specific suffix. Thus, CYP21A2 refers to the enzyme that hydroxylates carbon atom 21.

Clinical box Smith–lemli–opitz syndrome: a defect in 7-dehydrocholesterol reductase

The syndrome presents at birth with microencephaly, short nasal root, small chin, high arched palate, and often with midline cleft. There are often accompanying central nervous system (CNS) defects, polydactyly, and in males, ambiguous genitalia.

Despite the pathway of cholesterol synthesis and metabolism being well understood, a defect in 7-dehydrocholesterol reductase was only identified in 1993. The pathophysiology involves incomplete processing of embryonic signaling proteins (HH proteins), resulting in variable defects in different tissues.

While some of these children die in infancy, the rest, if assisted in feeding, survive with severe mental retardation (IQ 20–40). Most also develop growth retardation. Treatment involves giving additional cholesterol to the child. This improves growth but it appears to have no CNS benefits.

Corticosteroids

In the adrenal glands, zona fasciculata and zona reticularis are places of synthesis of the cortisol and the adrenal androgens. The aldosterone is synthesized in the outer layer (zona glomerulosa)

Cellular substructure of the adrenal cortex consists of three layers. The inner two layers (*zona fasciculata* and *zona reticularis*) are places of synthesis of cortisol, the main **glucocorticoid,** and the adrenal androgens. The cells of the outer layer (*zona glomerulosa*) synthesize aldosterone, the main **mineralocorticoid**

(Chapter 23). Although many of the steps are similar, they are controlled by very different mechanisms.

Biosynthesis of the cortisol depends on stimulation by pituitary adrenocorticotropic hormone (ACTH) which binds to its plasma membrane receptor and triggers a range of intracellular events, which cause hydrolysis of cholesteryl esters stored in lipid droplets and the activation of the cholesterol 20,22-desmolase enzyme, which converts C-27 cholesterol into pregnenolone, the first of the C-21 pregnane family of corticosteroids. This is the rate-limiting step of steroidogenesis. Thereafter, conversion to cortisol requires a dehydrogenation – isomerization and three sequential hydroxylation reactions at C-17, C-21 and C-11, catalyzed by the CYP enzymes. The pathway is regulated by negative feedback control of ACTH secretion by cortisol (Chapter 39).

In case of **aldosterone**, the main stimulus to its synthesis is not ACTH but angiotensin II (Chapter 24), and potassium is an important secondary stimulus. Angiotensin II, by binding to its receptor, and potassium, work cooperatively to activate the first step in the pathway: the conversion of cholesterol into pregnenolone. *Zona glomerulosa* lacks the 17α-hydroxylase but has abundant amounts of 18-hydroxylase, which is the first step of a two-stage reaction forming the 18-aldehyde group of aldosterone.

Androgens

Conversion of corticosteroids into androgens requires the 17–20 lyase/desmolase and a substrate that contains a 17α-hydroxyl group

The 17α-hydroxyl group is added prior to breaking the C17–C20 bond to yield the **androstane ring** structure. The enzyme is abundant in the Leydig cells of the testis and in the granulosa cells of the ovary. However, in these two tissues the same biosynthetic step is controlled by two different hormones. In the testis the rate-limiting cholesterol side-chain cleavage step is stimulated by the luteinizing hormone (LH), whereas in the ovary it is stimulated by the follicle-stimulating hormone (FSH).

Estrogens

Conversion of androgens into estrogens involves removal of the methyl group at C-19 by the 19-aromatase

The A ring undergoes two dehydrogenations, yielding the characteristic 1,3,5(10)-estratriene nucleus. This aromatase is most abundant in the granulosa cells of the ovary, although the enzyme in adipose tissue can also convert some testosterone into estradiol. Biological actions of the steroid hormones are diverse and are best considered as belonging to the trophic hormone system (Chapter 39). Many genetic defects have been identified in the structures of the CYP enzymes. These defects lead to abnormal steroid biosynthesis and to clinical disorders such as **congenital adrenal hyperplasia**.

Advanced concept box
Abnormalities in steroid synthesis are revealed by ALTERED pattern of urinary steroid metabolites

Steroid metabolites are excreted in urine mostly as water-soluble sulfate or glucuronic acid conjugates. The procedure used for their identification is gas chromatography – mass spectrometry (GCMS); it is very similar to methods adopted for the identification of anabolic steroids in sport. The first step in the analysis involves enzymatic release of the steroids from these conjugates; this is followed by chemical derivatization to increase their stability and improve separation, which is carried out by gas chromatography on capillary columns at high temperatures. Final detection is by mass fragmentation: for each steroid metabolite, a unique ion fragmentation ‘fingerprint’ is achieved, which allows positive identification and quantitation.

CLINICAL BOX A neonate born with ambiguous genitalia

Congenital adrenal hyperplasia

A neonate is born with ambiguous genitalia. Within 48 hours the infant becomes distressed and hypotensive. Biochemical investigations reveal:

- Na^+ 115 mmol/L (135–145 mmol/L)
- K^+ 7.0 mmol/L (3.5–5.0 mmol/L)
- 17-hydroxyprogesterone 550 nmol/L (upper reference limit 50 nmol/L)

Comment.

This baby has a severe form of steroid 21-hydroxylase deficiency, the commonest of a range of conditions known as congenital adrenal hyperplasia, which are characterized by defects in activity of one of the enzymes in the steroidogenic pathway. The condition has a genetic basis, and it leads to a failure to produce cortisol (and also possibly aldosterone). This results in reduced negative feedback inhibition of the pituitary production of ACTH. ACTH thus continues to stimulate the adrenal gland to produce steroids upstream of the enzyme block. The accumulating steroids include **17-hydroxyprogesterone**, which is further metabolized to testosterone (Figs 17.12 and 17.13). This results in androgenization of a female neonate. Mineralocorticoid deficiency causes renal salt wasting and requires urgent treatment with steroids and fluids. Long-term maintenance therapy with hydrocortisone and a mineralocorticoid suppresses ACTH and androgen production.

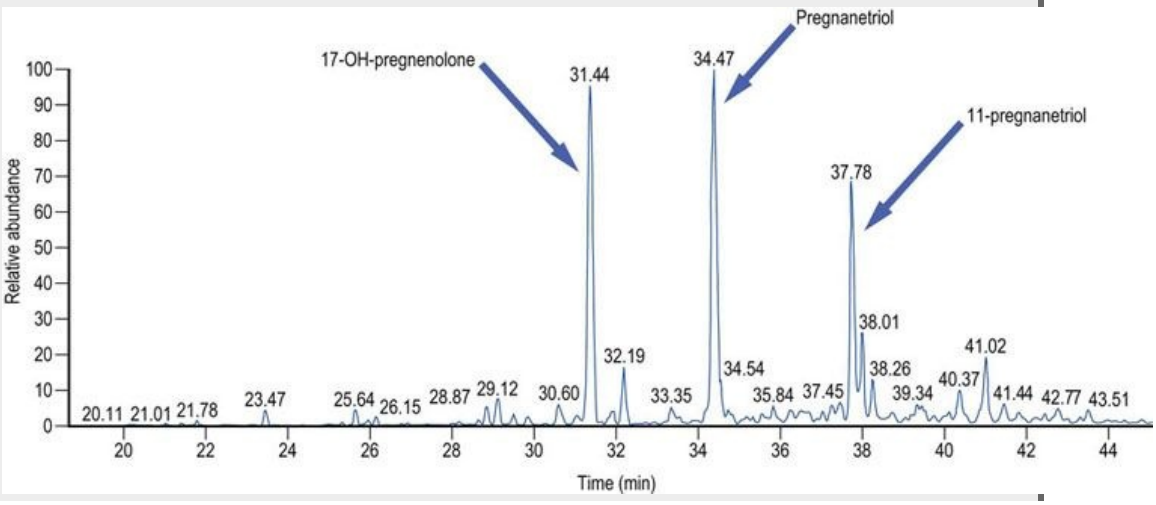


FIG. 17.13 Separation of urinary steroids performed by mass spectrometry.
In the clinical laboratory, the measurement of urinary steroid metabolites aids the diagnosis of a number of inherited disorders of the synthesis and metabolism of adrenal steroids, and steroid-producing tumors. It is particularly valuable in identifying the site of the defect in congenital adrenal hyperplasia. These investigations are most often performed in neonates with ambiguous genitalia, children with precocious puberty and in patients with suspected Cushing's syndrome (Chapter 39). This is a urinary steroid metabolite pattern from a patient with 21-hydroxylase deficiency variant of congenital adrenal hyperplasia. The most prominent steroid metabolites are 17-hydroxypregnenolone, pregnanetriol and 11-oxo-pregnanetriol.
Total ion chromatogram of a urinary steroid metabolite pattern from a patient with 21-hydroxylase deficiency variant of congenital adrenal hyperplasia. In this condition the most prominent steroid metabolites are 17-hydroxypregnanolone, pregnanetriol and 11-oxo-preganetriol.
X-axis: time at which the chromatographically separated steroid metabolites are detected by the mass spectrometer. Y-axis: relative abundance (quantity of ions).

A less severe form of this condition is a result of partial enzyme deficiency. It occurs in young women who present with menstrual irregularity and hirsutism as a consequence of excess of adrenal androgens.

Mechanism of action of steroid hormones

Steroid hormones act via nuclear receptors

All steroid hormones act by binding to ligand-activated nuclear receptors. The superfamily of hormone receptors also includes receptors for the thyroid hormone T_3 and the active forms of vitamins A and D (Chapter 40). Adjacent to the hormone-binding domain is a highly conserved DNA-binding domain, which is characterized by the presence of two zinc fingers (Fig. 34.3). Binding of the steroid ligand facilitates translocation of the activated receptor to the nucleus and its binding to a specific steroid response element in the promoter regions of target genes. This leads to gene transcription (Chapter 34). Genetic variability of steroid receptors structure may be associated with a variable degree of hormone resistance and diverse clinical presentations. See also discussion of the steroid

receptor in Chapter 35.

Elimination of steroid hormones

Most steroid hormones are excreted in urine. There are two main steps in this process. Firstly, the biological potency of the steroid must be removed and this is achieved by **a series of reduction reactions**. Secondly, the steroid structure must be rendered water soluble: this is achieved by **conjugation to a glucuronide or sulfate**, usually through the hydroxyl group at C-3. As a result, many different steroid hormone conjugates are present in urine, some of them in high concentrations. Urinary steroid profiling by gas chromatography – mass spectrometry typically identifies more than 30 such steroids. Their relative concentrations may be used to pinpoint specific defects in the steroidogenic pathway (Fig. 17.13).

Vitamin D_3

Vitamin D_3 (cholecalciferol) is derived from cholesterol and plays a key role in calcium metabolism. The actions and metabolism of vitamin D are described in Chapter 26.

Summary

■ Cholesterol is an essential constituent of cell membranes and the precursor molecule for bile acids, steroid hormones and vitamin D.
■ Cholesterol is both supplied with the diet and synthesized de novo from the acetyl-CoA.
■ Cholesterol biosynthesis is precisely regulated. The rate-limiting enzyme in the synthetic pathway is the HMG-CoA reductase.
■ The transformation of cholesterol into bile acids and steroid hormones involves several hydroxylation reactions catalyzed by cytochrome P450 monooxygenases.

Active learning

1. Describe the regulation of intracellular cholesterol concentration.
2. What are the secondary bile acids and how are they produced?
3. Discuss the enterohepatic circulation of bile acids.
4. Discuss the role of monooxygenases in steroid synthesis.

Further reading

Charlton-Menys, V, Durrington, PN. Human cholesterol metabolism and therapeutic molecules. *Exp Physiol*. 2007; 93:27–42.

Goldstein, J, DeBose Boyd, RA, Brown, MS. Protein sensors for membrane sterols. *Cell*. 2006; 124:35–46.

Janowski, BA, Willy, PJ, Devi, TR, *et al.* An oxysterol signaling pathway mediated by the nuclear receptor LXRa. *Nature*. 1996; 383:728–731.

Marcil, M, Brooks-Wilson, A, Clee, SM, *et al.* Mutations in the *ABC1* gene in familial HDL deficiency with defective cholesterol efflux. *Lancet*. 1999; 354:1341–1346.

Maxfield, FR, Van Meer, G. Cholesterol, the central lipid of mammalian cells. *Curr Opin Cell Biol*. 2010; 22:422–429.

Ory, DS. Nuclear receptor signaling in the control of cholesterol homeostasis. *Circ Res*. 2004; 95:660–670.

Sakai, J, Rawson, RB. The sterol regulatory element-binding protein pathway: control of lipid homeostasis through regulated intracellular transport. *Curr Opin Lipidol*. 2001; 12:261–266.

Vegiopoulos, A, Herzig, S. Glucocorticoids, metabolism and metabolic diseases. *Mol Cell Endocrinol*. 2007; 275:43–61.

Young, SG, Fong, LG. Lowering plasma cholesterol by raising LDL receptors – revisited. *N Engl J Med*. 2012; 366:1154–1155.

Websites and downloads

Cholesterol biosynthetic pathway – Rat Genome Database. http://rgd.mcw.edu/rgdweb/pathway/pathwayRecord.html?acc_id=PW:0000454.

Herman, GE, Disorders of cholesterol biosynthesis: prototypic metabolic malformation syndromes. *Hum Mol Genet*. 2003;12(suppl 1):R75–R8. http://hmg.oxfordjournals.org/content/12/suppl_1/R75.full

KEGG Pathway Database. Primary bile acid biosynthesis – Reference pathway. www.genome.jp/kegg/pathway/map/map00120.html.

King, M. http://themedicalbiochemistrypage.org/steroid-hormones.php

CHAPTER 18

Lipoprotein Metabolism and Atherogenesis

Marek H. Dominiczak

Learning objectives

After reading this chapter you should be able to:

■ Describe the composition and functions of lipoproteins present in plasma: the chylomicrons, very low-density lipoproteins, remnant particles, low-density lipoproteins and high-density lipoproteins.

■ Describe the fuel transport pathway and the overflow pathway of lipoprotein metabolism.

■ Describe the reverse cholesterol transport and its links with other pathways of lipoprotein metabolism.

■ Outline mechanisms and regulation of intracellular cholesterol concentration, including the role of relevant transcription factors, receptors and enzymes.

■ Comment on laboratory tests that assess lipid metabolism and cardiovascular risk.

■ Discuss the main component processes of atherogenesis – endothelial dysfunction, arterial deposition of lipids, chronic low-grade inflammation and thrombosis – as well as their relationships to atherosclerotic plaque growth, destabilization and rupture.

Introduction

Lipoproteins distribute cholesterol and triacylglycerols (esters of glycerol and fatty acids, synonymously called triglycerides; we use both terms), from the intestine and the liver, to peripheral tissues

These processes link closely with the energy metabolism. Abnormalities of lipoprotein metabolism are key factors in the development of atherosclerosis, a process affecting arterial walls and causing coronary heart disease, stroke and peripheral vascular disease. Atherosclerosis-related cardiovascular disease is presently the most frequent cause of death in the world: ischemic heart disease and cerebrovascular disease are together responsible for 23.6% of all deaths worldwide (WHO data 2011).

Free fatty acids and triacylglycerols are transported between organs and tissues

Fatty acids are, together with glucose, the key energy substrates. However, in contrast to glucose, the fatty acids can be stored long term in the adipose tissue (as triacylglycerols, Chapter 16) to provide energy during periods of fasting. They are absorbed from the gastrointestinal tract as components of ingested food but are also synthesized endogenously, primarily in the liver and intestine. Fatty acids need to be transported from their places of absorption or synthesis to peripheral tissues. While the free (nonesterified) short-and medium-chain fatty acids ‘travel’ in plasma bound to albumin, the long-chain fatty acids are too hydrophobic to be transported in this manner. Instead, they are transported as triacylglycerols packaged into particles known as lipoproteins.

Lipoproteins

Lipoproteins are composed of hydrophilic, hydrophobic and amphipathic molecules

Lipoprotein particles contain triacylglycerols, cholesterol, phospholipids, and proteins (apolipoproteins). They also transport fat-soluble vitamins such as vitamin A and vitamin E. The hydrophobic cholesteryl esters and triacylglycerols reside in the core of the lipoprotein particles and amphipathic phospholipids and free cholesterol, together with apolipoproteins, form their outer layer (Fig. 18.1). Some apolipoproteins, such as apolipoprotein B (apoB), are embedded in the particle surface while others, such as apoC, are only loosely bound and can be exchanged between different particles.

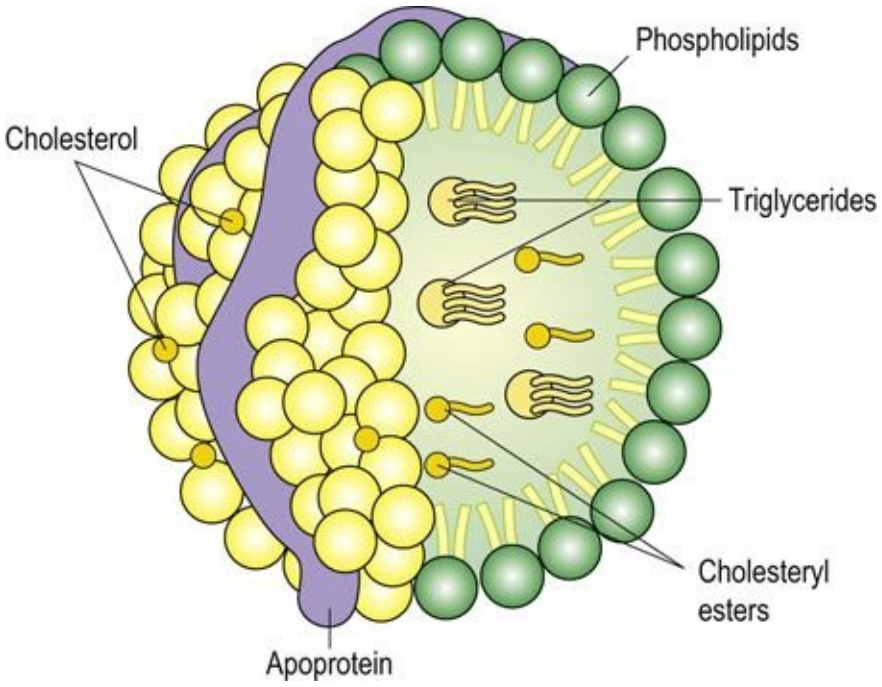


FIG. 18.1 Lipoprotein particle.
Lipoprotein particle has a hydrophilic external surface and a hydrophobic interior. The external layer contains free cholesterol, phospholipids, and apolipoproteins. Cholesteryl esters and triacylglycerols are located in the hydrophobic particle core.

Plasma lipoproteins are particles of different size and density

Lipoproteins are classified on the basis of either their density or constituent set of apolipoproteins: in plasma they form a continuum of size and density (Table 18.1). The main lipoprotein classes are the chylomicrons, the very low-density lipoproteins (VLDL), the remnant particles (which are nearly identical to the intermediate-density lipoproteins, IDL), the low-density lipoproteins (LDL), and the high-density lipoproteins (HDL). The VLDL and remnant particles are triacylglycerol-rich, whereas the LDL are triacylglycerol-poor and cholesterol-rich. With a decreasing triacylglycerol content, the density of a particle increases and its size decreases. Thus the density increases from the chylomicrons (the lightest), through VLDL, IDL, LDL, to the HDL (the heaviest).

Table 18.1

Lipoprotein classes

Particle	Density (kg/L)	Main component	Apolipoproteins*	Diameter (nm)
Chylomicrons	0.95	TG	B48 (A, C, E)	75–1200
VLDL	0.95–1.006	TG	B100 (A, C, E)	30–80
IDL	1.006–1.019	TG & cholesterol	B100, E	25–35
LDL	1.019–1.063	Cholesterol	B100	18–25
HDL	1.063–1.210	Protein	AI, AII (C, E)	5–12

TG, triacylglycerol (triglyceride); VLDL, very low-density lipoproteins; IDL, intermediate-density lipoproteins; HDL, high-density lipoproteins. When separated by electrophoresis VLDL are called pre-β-lipoproteins, LDL, β-lipoproteins and HDL α-lipoproteins. *The most abundant apoproteins present in a given lipoprotein particle are indicated first, with those that are exchanged with other particles in parentheses.

Advanced concept box

Lipoproteins can be separated by ultracentrifugation

Every clinical laboratory uses centrifuges to separate red blood cells from serum or plasma. These machines develop a moderate centrifugal force, 2000–3000 *g*. However, in the specialist lipid, protein and nucleic acid biochemistry much larger centrifugal forces (40,000–100,000 *g*) are applied to plasma to separate particles and molecules. This technique is called

ultracentrifugation and is extensively used in lipid research. **When centrifugal** force is applied to a solution, particles that are heavier than the surrounding solvent sediment, and those lighter than the solvent, float to the surface at a rate proportional to the applied centrifugal force and to the particle size. The formula below shows factors that affect particle movement:

$$v = [d^2(P_p - P_s) - g] / 18\mu$$

where v = sedimentation rate; d = diameter; P_p = particle density; P_s = solvent density; μ = viscosity of the solvent, and g = gravitational force.

In a technique known as **flotation ultracentrifugation**, the plasma is over-layered with a solution of defined density, *e.g.* 1.063 kg/L, the density of the VLDL. After several hours of centrifugation (with the rotor speeds around 40,000 rev/min), the VLDL float to the surface, where they can be harvested. Solutions of different density can be used to separate other lipoproteins. Modifications of the ultracentrifugation technique, such as density gradient centrifugation, can be applied to separate plasma into several 'bands' containing different lipoprotein fractions.

Apolipoproteins

Apolipoproteins are protein components of lipoprotein particles. Their role is both structural and metabolic

Apolipoproteins that are embedded into the surface of lipoprotein particles determine their metabolic fate through interactions with cellular receptors. They also regulate the activity of enzymes involved in lipid transport and distribution. **Each class of lipoproteins contains a characteristic set of apolipoproteins.** The main apolipoproteins are listed in Table 18.2. The most important are apoA, apoB, apoC, apoE, and apo(a).

Table 18.2

Structure and function of apolipoproteins

Apo	Genes	Examples of isoforms	Synthesis	Structure	Function	Lipoproteins	Lipoprotein metabolism pathway
AI	Chromosome 11, AI/C3/A4/A5 gene cluster	Six polymorphic isoforms Mutations: Apo AI Tangier AI Milano AI Marburg	Liver, intestine	243 AA, 28,000 Da	Structural in HDL. LCAT activator	70% of HDL protein. Most abundant protein in HDL. Chylomicrons, VLDL	RCT, fuel transport pathway
AII	Chromosome 1		Liver, intestine	77AA, 17,400 Da Mainly present as dimer (mol. mass above is that of the dimer)	Structural in HDL.	20% of HDL protein. Second most abundant after apoAI. Chylomicrons, VLDL	RCT (main marker), fuel transport pathway
AIV	Chromosome 11, A1/C3/A4/A5 gene cluster	ApoAIV 360 (common) ApoAIV-1, ApoAIV-2	Liver, intestine		Metabolism of TG-rich particles. Interacts with CII in LPL. LCAT activator	Chylomicrons, HDL, free in plasma	Fuel transport pathway, RCT
AV	Chromosome 11, A1/C3 A4/A5 gene cluster	Multiple variants	Liver		Chylomicron and VLDL assembly. LPL activator	Chylomicrons, VLDL, HDL	Fuel transport pathway, RCT
CIII	Chromosome 11, A1/C3 A4/A5 gene cluster	Variants with differing sialic acid content: CIII-0, CIII-1, CIII-2	Liver, intestine	79AA, 8800 Da	LPL inhibitor. Masks or displaces apoE from LRP	Surface of TG-rich particles: chylomicrons, VLDL remnants, HDL	Fuel transport pathway, RCT
CII	Chromosome 19		Liver, intestine	79AA, 8900 Da	LPL activator: deficiency leads to gross hypertriglyceridemia	Chylomicrons, VLDL, HDL	Fuel transport pathway, RCT
CI	Chromosome 19		Liver, intestine	57AA	LCAT activator, LPL inhibitor, CETP inhibitor. Inhibits apoE binding to LRP	Chylomicrons, VLDL, HDL	Fuel transport pathway, RCT
B100	Chromosome 2	Over 100 polymorphisms	Liver	4536 AA, 550,000 Da	Structural component of VLDL, IDL, LDL. Ligand for LDL-receptor	VLDL, IDL, LDL	Fuel transport pathway, overflow pathway. One molecule per particle – marker of particle number
B48	Chromosome 2		Intestine	2152 *N*-terminal AA of B100, 264,000 Da. 8–10% CHO	Structural component of chylomicrons and chylomicron remnants	Chylomicrons, chylomicron remnants	Fuel transport pathway

Apo	Genes	Examples of isoforms	Synthesis	Structure	Function	Lipoproteins	Lipoprotein metabolism pathway
E	Chromosome19, E/C1/C2/C4 gene cluster	Three main isoforms: E2 E3 E4. Many variants	Liver, intestine, brain, kidney, spleen, adrenals, and other	299 AA, 34,200 Da	Multifunction protein. LDL-receptor ligand for LDL and chylomicron remnants. LRP ligand. Modulates LPL, CETP, LCAT, HTGL. Antioxidant molecule. Regulator of inflammatory response	Chylomicrons, VLDL, remnants HDL	Fuel transport pathway, RCT
(a)	Chromosome 6. Linkage with plasminogen gene.	Over 20 isoforms, dependent on number of kringle 4 repeats Kringle 4 region most variable	Liver	Variable molecular mass: 187,000–800,000 Da pre-beta mobility. High sialic acid content		HDL-2, LDL	Role in fibrinolysis?

CHO, carbohydrates; AA, aminoacids; RCT, reverse cholesterol transport; LRP, LDL-receptor-like protein. LPL,lipoprotein lipase; LCAT, lecithin: cholesterol acyltransferase; CETP, cholesteryl ester transfer protein; HTGL, hepatic triglyceride lipase.

See text for references.

Reproduced, with permission, from Dominiczak MH, Caslake MJ. Apolipoproteins: metabolic role and clinical biochemistry applications. Ann Clin Biochem 2011; 48: 498–515.

Apolipoproteins A (AI and AII) are present in the HDL. Apo AI is a small protein of 243 amino acids and is synthesized in the liver and in the intestine. The APOA1 gene is part of the APOA1/C3/A4/A5 complex. Apo A1 is the main apolipoprotein in HDL particles. It activates the lecithin: cholesterol acyltransferase (LCAT; the cholesterol-esterifying enzyme) and has anti-inflammatory and antioxidant properties. Clinically, it is a marker of HDL concentration.

ApoAII is an even smaller protein (77 amino acids) and is mostly synthesized in the liver. It is also present primarily in the HDL. It is a cofactor for the LCAT and the cholesterol ester transfer protein (CETP), and inhibits lipoprotein lipase (LPL). It is not routinely measured in plasma. Apolipoprotein A binds to the scavenger receptor BI.

Apolipoprotein B exists in two common variants, apoB100 and apoB48. The apoB100 controls the metabolism of the LDL, whereas its truncated form, apoB48 (see Fig. 34.7) is present in the chylomicrons. ApoB100 is a relatively large protein with a molecular mass of 513,000 kDa, comprising 4509 amino acids. It is synthesized in the liver. There are over 100 variants of the APOB gene. Mutation at the amino acid residue 3500 decreases its binding to the LDL receptor and is the cause of a condition known as **familial defective apoB**

(FDB). Because there is only 1 molecule of apoB per lipoprotein particle, the measurement of the apoB in plasma is a good marker of the sum of the VLDL, remnants, and the LDL particles. ApoB100 binds to the LDL receptor.

ApoB48 is a truncated form of apoB100. ApoB100 and apoB48 are synthesized from the same gene. A stop codon is introduced into the gene during editing of apoB100 mRNA (the '48' designation signifies that it comprises the amino-terminal 48% of the apoB sequence). It is synthesized in the enterocytes and is secreted from the intestine in the chylomicrons. ApoB48 does not bind to the LDL receptor. Its measurement is a marker of the number of chylomicrons and chylomicron remnant particles.

Apolipoprotein E has a molecular mass of 34,200 Da and comprises 299 amino acids. It is present in all lipoprotein classes. It binds to the LDL receptor with a higher affinity than apoB100. It also binds to the LDL-receptor-related protein (LRP), driving the cellular uptake of the remnant particles. ApoE stimulates the LPL, the hepatic triglyceride lipase (HTGL) and the LCAT. ApoE exists in 3 isoforms, E2, E3 and E4. Its synthesis is controlled by three major alleles, ε2, ε3, and ε4. The **E2 isoform** results from cysteine-for-arginine substitution at position 158 (compared to E3) and has lower affinity to receptors. This, in homozygotes, slows down the uptake of the remnant particles and results in **familial dyslipidemia** (also known as type III hyperlipidemia). Isoforms E3 and E2 are also associated with higher plasma insulin and glucose concentrations than E4. In the HDL, apoE contributes, together with apoAI, to cholesterol removal from cells. Interestingly, in mice, inactivation of the LRP resulted in a reduction of body weight and a decreased lipid clearance. Thus, although apoE is atheroprotective, it may be associated with weight gain by promoting lipid transport to peripheral tissues.

ApoE is also synthesized in the brain by astrocytes and microglia: it affects growth and repair of the CNS cells and also is anti-inflammatory and antioxidant. Individuals with E4 phenotype were shown to be at an increased risk of the sporadic form of **Alzheimer disease**. ApoE is not routinely measured in plasma in clinical laboratories, but phenotyping and genotyping apoE isoforms is used in the diagnosis of familial dyslipidemia.

Apolipoproteins C (CI, CII and CIII) act as enzyme activators and inhibitors and they are extensively exchanged between different lipoprotein classes.

Apolipoprotein (a) is a component of lipoprotein (a) (Lp(a)). Lp(a) is a hybrid particle that comprises apo(a) attached to apoB100 by a disulfide linkage. It is highly polymorphic and its molecular mass may vary from 187,000 to

800,000 Da. Apo(a) possesses a protease domain and a number of repeating sequences of approximately 80–90 amino acids in length, stabilized by disulfide bridges into a triple-loop structure. These structures are called kringles (the name of a Danish pastry of a similar shape). One of the kringles, kringle IV, is repeated 35 times within the apo(a) sequence. The number of kringle IV repeats determines the size of the lipoprotein (a) isoforms.

Apo(a) is synthesized in the liver and it binds to the LDL receptor. It is structurally related to plasminogen. Lp(a) concentration in plasma is almost entirely genetically determined, and is little influenced by lifestyle factors. Lp(a) is modestly associated with cardiovascular risk. Lp(a) is measured in plasma during specialist assessments of cardiovascular risk.

Clinical studies show that the measurements of plasma apolipoproteins predict the risk of CVD better than the lipid testing *i.e.* measurements of total cholesterol and LDL-cholesterol. However, since most of the large epidemiologic studies and treatment algorithms are referenced to lipid measurements, most laboratories still measure lipids for the assessment of cardiovascular risk.

Lipoprotein receptors

LDL receptor is regulated by the intracellular cholesterol concentration

Cellular uptake of the lipoproteins is mediated by apolipoprotein binding to receptors present on cell membranes. This allows cells to acquire cholesterol and other lipids. The key lipoprotein receptor is the LDL receptor (the apoB/E receptor). It was discovered by Joseph Goldstein and Michael Brown, who jointly received the Nobel Prize for this work in 1985. The receptor can bind either apoB100 or apoE. The mature receptor protein contains 839 amino acids and spans the cell membrane (Fig. 18.2). The receptor gene is located on chromosome 19 and its expression is regulated by the intracellular cholesterol concentration.

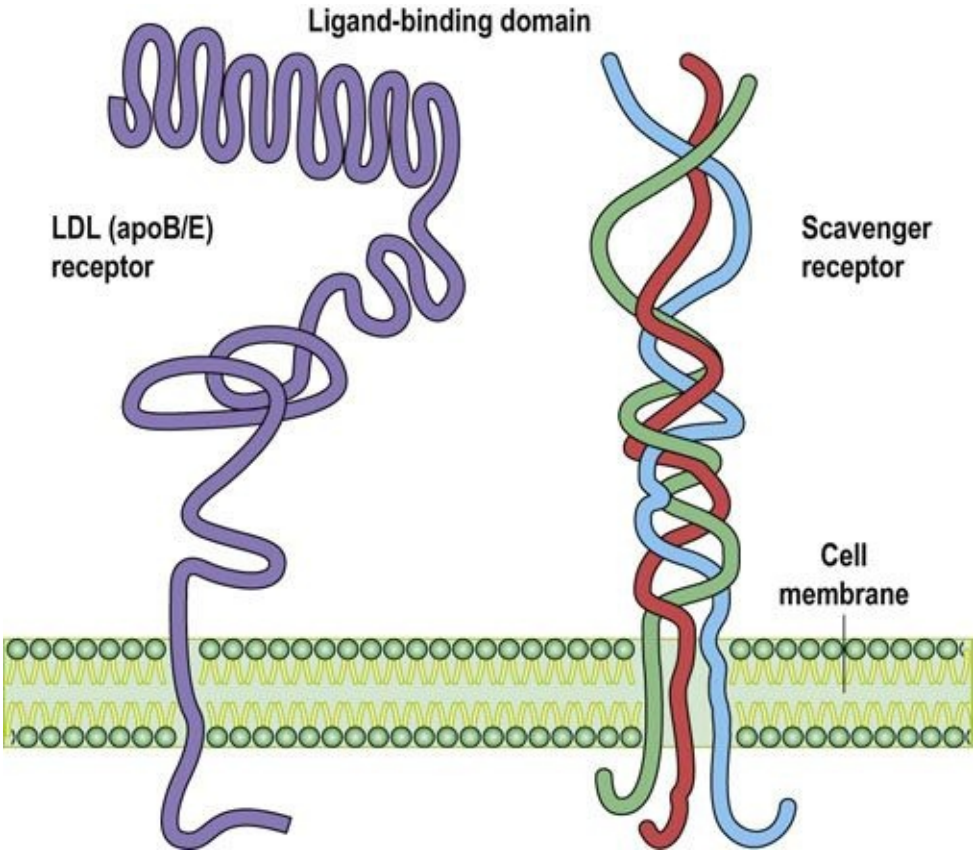


FIG. 18.2 Lipoprotein receptors.
The LDL receptor, also known as the apoB/E receptor, mediates cellular uptake of intact LDL particles. The scavenger receptor internalizes chemically modified (e.g. oxidized) LDL. Both types of receptor span cell membranes. The expression of LDL receptor is regulated by the intracellular cholesterol concentration, while the scavenger receptor remains unregulated. The scavenger receptor type A, illustrated here, is present on macrophages and has a collagen-like structure. Scavenger receptor type BI participates in the metabolism of the HDL particles.

Scavenger receptors are nonspecific and nonregulated

While the apoB/E receptor binds specific ligands, the scavenger receptors can bind many different molecules. These receptors are present on phagocytic cells such as macrophages. Importantly, they are not subject to feedback regulation, and therefore they may overload the cell with the ligand they bind. Scavenger receptors are designated as class A and class B, and CD36. Class A receptors have a collagen-like triple helical structure. They do not bind intact LDL but readily bind chemically modified (e.g. acetylated or oxidized) LDL (see Fig. 37.5). Class B receptor takes up HDL particles in the liver.

Enzymes and LIPID transfer proteins

Two hydrolases, lipoprotein lipase (LPL) and hepatic triglyceride lipase (HTGL), remove triacylglycerols from lipoprotein particles. LPL is bound to heparan sulfate proteoglycans on the surface of vascular endothelial cells, and HTGL is associated with plasma membranes in the liver. LPL digests triacylglycerols in chylomicrons and VLDL, and releases fatty acids and glycerol to cells. HTGL acts on particles already partially digested by LPL, and facilitates the conversion of IDL into LDL (see below).

Lecithin:cholesterol acyltransferase (LCAT) is a glycoprotein enzyme synthesized in the liver, which is associated with the HDL. LCAT esterifies cholesterol that HDL acquire from cells. LCAT is activated by apoAI. Within cells, however, cholesterol is esterified by a different enzyme – the acylCoA:acylcholesterol transferase (ACAT). There are two isoforms of ACAT: ACAT1 is present in macrophages and ACAT2 is present in the intestine and liver. Another protein, the cholesterol ester transfer protein (CETP), facilitates the exchange of cholesteryl esters for triacylglycerols between the HDL on the one hand, and the VLDL and IDL on the other, contributing to reverse cholesterol transport.

Pathways of lipoprotein metabolism

The fuel transport pathway and the overflow pathway reflect the fuel transport function of lipoprotein particles

The transport function of the lipoproteins is essential for the tissue distribution of two essential classes of compounds: triacylglycerols and cholesterol. Triacylglycerols and fatty acids are part of the body energy metabolism, whereas cholesterol transported by lipoproteins forms an extracellular pool available for cellular uptake, which 'backs up' the capacity of cells to synthesize cholesterol. The main stages of lipoprotein metabolism are:

- The assembly of lipoprotein particles. The chylomicrons are assembled in the intestine, the VLDL in the liver, and the HDL are synthesized both in the liver and the intestine.
- The transfer of lipoproteins to peripheral cells and the release of triacylglycerols/fatty acids from lipoproteins to cells. This is facilitated by the LPL and HTGL. As a result, chylomicrons and VLDL decrease in size and become remnant particles.
- The binding of remnant particles to the liver receptors and their uptake.
- Generation of LDL particles from remaining remnants by HTGL-mediated hydrolysis, their binding to the apoB/E receptor and cellular uptake.
- The reverse cholesterol transport, *i.e.* removal of cholesterol from cells by the HDL particles.

The pathway of metabolism of the chylomicrons and the VLDL, called here **the fuel transport pathway**, is closely related to the **feed–fast cycle** (Chapter 21) and thus to energy metabolism (Fig. 18.3). The fuel transport pathway also links closely to the **reverse cholesterol transport** through triglyceride and cholesteryl ester exchanges with the HDL particles. The completion of the fuel transport stage (i.e. the delivery of triglycerides to the peripheral cells) is linked to generation of the remnants and then the LDL particles, which form an extracellular cholesterol pool. Because the LDL generation depends on the activity of the fuel transport pathway, further LDL metabolism is termed the **overflow pathway**.

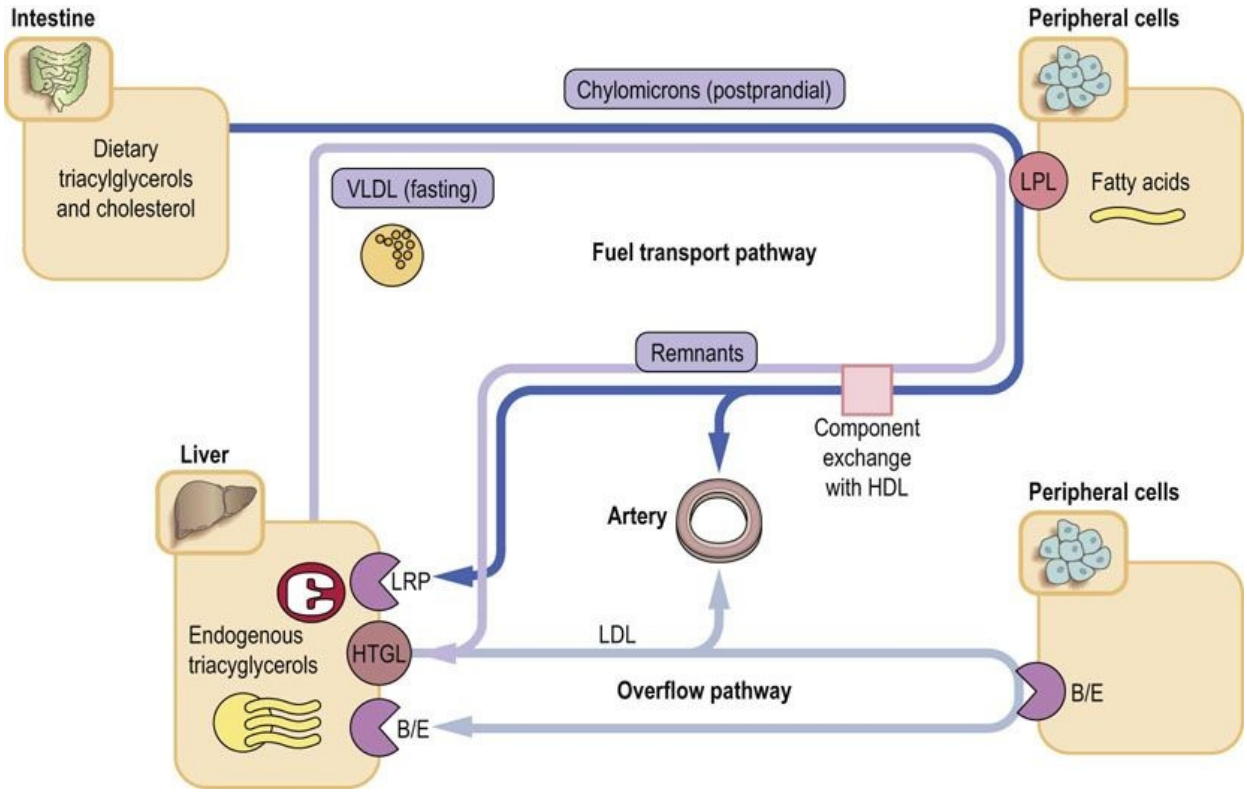


FIG. 18.3 Lipoprotein metabolism: the fuel transport pathway and the overflow pathway.
The fuel transport pathway is linked to energy metabolism and to the feed-fast cycle. In the fed state, chylomicrons transport triglycerides to the periphery, where the LPL hydrolyzes the them, liberating the fatty acids into cells. Chylomicron remnants are metabolized in the liver, after binding through the apoE to the LDL receptor, and also to the LRP. The VLDL particles transport fuel from the liver to peripheral tissues. The VLDL remnants, similarly to the chylomicron remnants, return to the liver. Approximately 65% are taken up after binding to the apoB/E receptor, and the remaining ones are hydrolyzed by HTGL, yielding the LDL particles. The triglyceride-rich particles (chylomicrons, VLDL and the remnant particles) acquire additional cholesteryl esters from the HDL in exchange for triglycerides. In clinical practice, the marker of the activity of the fuel transport pathway is the measurement of plasma triglycerides. The remnant particles can also be measured but this has been done mostly for research purposes.
The overflow pathway is the pathway of LDL metabolism. The LDL particles are generated from remnants in the fuel transport pathway, and are cholesterol-rich. The LDL are taken up by the apoB/E receptor in response to a decrease in intracellular cholesterol concentration. In clinical practice, the markers of activity of the overflow pathway are the measurements of plasma total cholesterol and LDL-C LPL, lipoprotein lipase, LRP, LDL-receptor-related protein, HTGL, hepatic triglyceride lipase.

The fuel transport pathway of lipoprotein metabolism

Chylomicrons transport dietary lipids

The pathway involving the assembly of chylomicrons is active after a fat-containing meal. Triacylglycerols present in food are acted upon by pancreatic lipases and are absorbed as monoacylglycerols, free fatty acids, and free glycerol (see Chapter 10). The cells lining the intestine (enterocytes) resynthesize triacylglycerols and, together with phospholipids and cholesterol, assemble them on the apoB48 into **chylomicron** particles. These are secreted into the lymph and reach plasma through the thoracic duct.

The main apolipoprotein of the chylomicrons is the apoB48. Chylomicrons also contain apoA, C, and E. Once they reach peripheral tissues, their triacylglycerols are hydrolyzed by the LPL, and the fatty acids enter cells. What is now left of the chylomicrons are smaller and denser particles known as the **chylomicron remnants**. The remnants acquire some cholesteryl esters from the HDL (see below). Importantly, the change in particle size uncovers the apoE on their surface, which mediates the remnant binding to the apoB/E receptor and to the LRP in the liver. The half-life of chylomicrons in plasma is less than 1 h. Chylomicrons normally appear in plasma only after fat-containing meals, giving plasma a milky appearance (Fig. 18.3).

There is now substantial evidence that the postprandial lipemia may play a role in atherogenesis. Nonfasting plasma triglyceride concentration (reflecting an increase in atherogenic lipoprotein remnants) has been linked to the risk of cardiovascular disease (CVD). The postprandial elevation of triglycerides is particularly important in persons with diabetes mellitus and insulin resistance.

VLDL particles transport triacylglycerols synthesized in the liver

In contrast to the chylomicron transport of dietary fat, triacylglycerols that are synthesized in the liver are transported by the **VLDL**. The VLDL are assembled both during fasting and after meals. They are built up around apoB100 molecules. The lipidation of apoB is facilitated by the microsomal triglyceride transfer protein (MTP). Unused apoB100 is degraded by an ubiquitin-dependent protease (Fig. 34.10). After being secreted into plasma, VLDL acquire cholesteryl esters and apolipoproteins (apoC and apoE) from the HDL. In the peripheral tissues, their triacylglycerols are hydrolyzed by the LPL in a way analogous to chylomicrons; this yields **VLDL remnants** (also called IDL). The conformations of apoB100 and apoE in the VLDL do not allow binding to the

apoB/E receptor. However, in the remnants, apoE assumes a conformation that allows such binding. The remnants are thus either taken up by the liver, or are further hydrolyzed by another enzyme, the HTGL, which, by removing practically all triacylglycerols, transforms them into LDL. Because of the previous triacylglycerol loss, the remnant particles are relatively cholesterol-rich. Their small size facilitates penetration of the vascular endothelium, and both these characteristics make them atherogenic.

VLDL enriched with cholesteryl esters give rise to highly atherogenic small-dense LDL particles

The VLDL are enriched by CETP in cholesteryl esters derived from HDL, in exchange for triglycerides. When such cholesteryl-ester-enriched particles are acted upon by HTGL, the removal of triglycerides decreases their size to an even greater extent creating the small-dense (sd) LDL, which are -not surprisingly- strongly atherogenic. It is thought that their presence might be responsible for an increased CVD risk in some patients who present with seemingly 'normal' plasma lipid concentrations – as happens in diabetes mellitus. Since only 1 molecule of apoB100 is present in each LDL particle, the presence of sd-LDL can manifest itself as **hyperapobetalipoproteinemia**, i.e an increased plasma apoB100 concentration with relatively normal cholesterol concentration. This condition is associated with increased CVD risk.

The overflow pathway of lipid metabolism

LDL particles are taken up by cells by the same route as remnant particles

The LDL are thus triacylglycerol-poor and relatively cholesterol-rich. They are generated from the VLDL remnants by the removal of almost all triglycerides by the HTGL. They are thus the 'overflow' products of the fuel transport pathway.

The LDL are small enough to penetrate the vascular wall. They contain only one apolipoprotein (the apoB100) and they are the main carrier of cholesterol in plasma. They remain in the circulation longer than the remnants, and are taken up through the apoB/E receptor, either in the liver (approximately 80% of particles) or in the peripheral cells (see Fig. 18.3). Their affinity for the receptor is lower than that of the apoE-containing remnant particles.

Intracellular cholesterol synthesis and its cellular uptake are interdependent

Most cells synthesize cholesterol for their own needs. However, when the concentration of intracellular cholesterol decreases, the cells can acquire it from the outside – and the lipoproteins constitute a pool of extracellular cholesterol upon which the cells draw. After internalization, the LDL–receptor complex is digested by the lysosomal enzymes. The released free cholesterol is esterified within the cell and the receptor protein recycles back to the membrane. The free cholesterol is a negative feedback regulator of its own synthesis. This is mediated by a family of transcription factors called sterol regulatory element-binding proteins (SREBPs). SREBPs regulate transcription of genes coding for enzymes in the pathway of cholesterol synthesis: the 3-hydroxy-3-methylglutaryl coenzyme A synthase and HMG-CoA reductase, but also the gene coding for the apoB/E receptor (Chapter 17). Depletion of hepatic sterols increases the SREBPs level and consequently both cholesterol synthesis and the expression of the apoB/E receptor. On the other hand, increased intracellular cholesterol concentration inhibits the SREBP pathway, decreasing cholesterol synthesis and receptor expression. This is dioscussed in more detail in Chapter 18 (Fig. 18.4; see also Fig. 17.7).

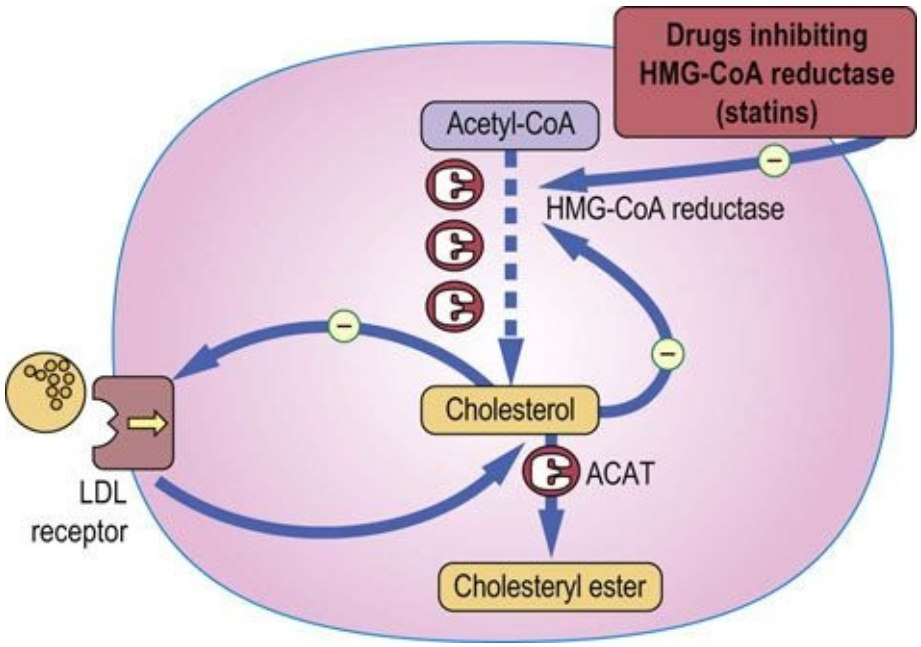


FIG. 18.4 Regulation of intracellular cholesterol concentration.

Intracellular cholesterol regulates the activity of the HMG-CoA reductase, the rate-limiting enzyme in cholesterol synthesis, and also the expression of LDL receptors on the cell membrane. The expression of relevant genes is controlled by the SREBP transcription factors. ACAT: acyl CoA:acylcholesterol transferase. See also Figure 17.7

The reverse cholesterol transport

HDL particles transport cholesterol from the peripheral cells to the liver – this makes them anti-atherogenic

The HDL particles transport cholesterol from the periphery to the liver – thus the 'reverse' transport, Fig. 18.5. This relieves the cholesterol burden of cells and makes the HDL anti-atherogenic: a high plasma concentration of HDL-cholesterol (HDL-C) is associated with longevity, and a decreased concentration of HDL-C (and apoAI) is associated with an increased CVD risk.

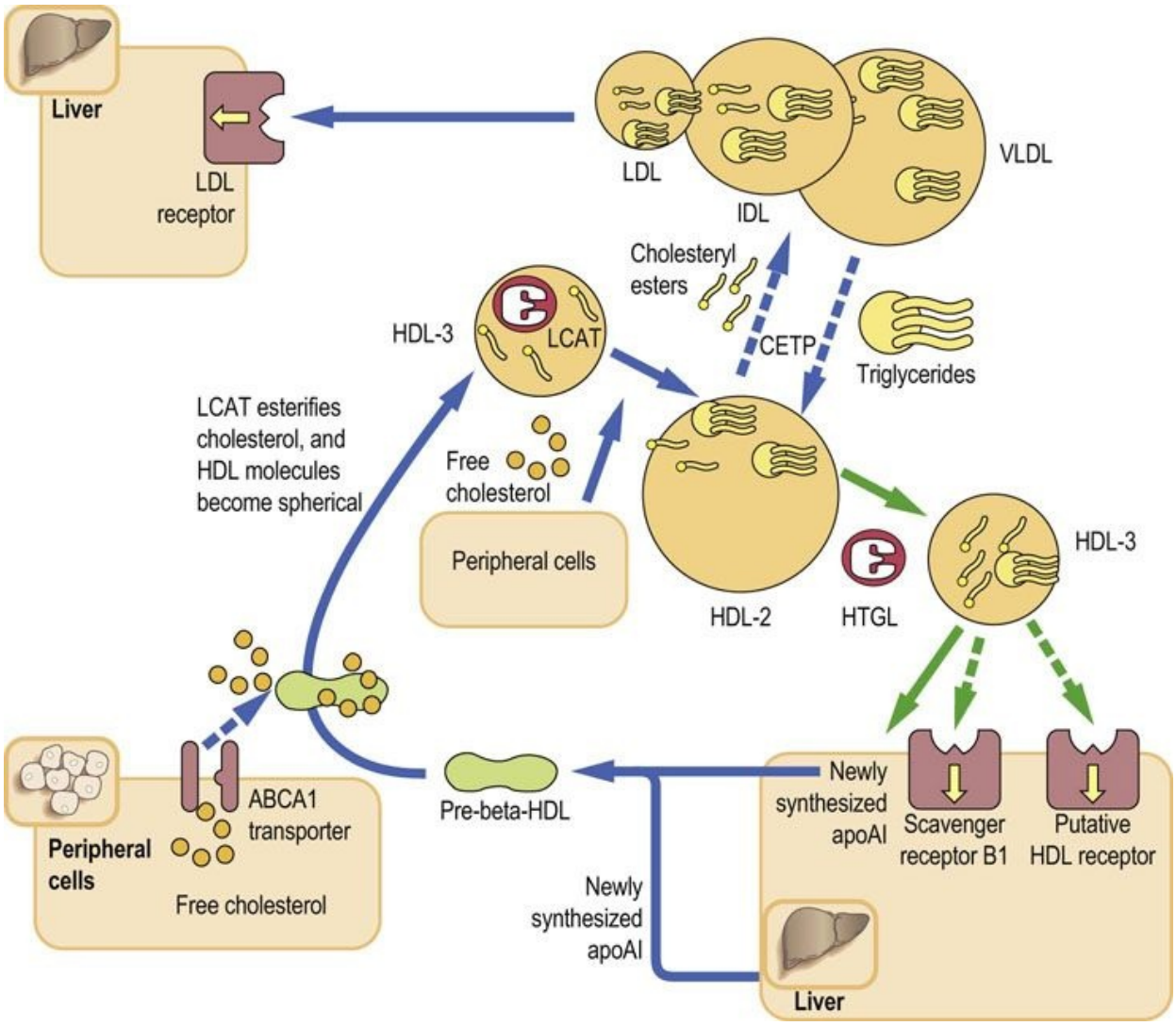


FIG. 18.5 Reverse cholesterol transport.
HDL are assembled in the liver and intestine as discoid particles. They acquire cholesterol from cell membranes through the ABCA1 transporter. The LCAT associated with HDL esterifies the acquired cholesterol. The cholesteryl esters move to the inside of the particle, making it spherical (HDL-2). The CETP facilitates exchange of apolipoproteins and cholesteryl esters between HDL and triglyceride-rich lipoproteins: this inserts the cholesteryl esters into the fuel transport pathway and is the main route of reverse cholesterol transport in humans. The HDL-2 particles which acquire triacylglycerols during the CETP-mediated exchange, increase in size, becoming HDL-3. They bind to the scavenger receptor BI on the hepatocyte membrane and transfer cholesteryl esters to the liver. When the transfer is completed, the size of the HDL particle decreases again. Some of the redundant surface material is released, forming apoAI-rich, lipid-poor pre-β-HDL, which re-enter the cholesterol removal cycle. LCAT, lecithin:cholesterol acyltransferase; CETP, cholesterol ester transfer protein; HTGL, hepatic triglyceride lipase.

The HDL particles are synthesized in the liver and the intestine. Their main

apolipoproteins are apoAI and apoAII. They also contain apoC and apoE. Importantly, the HDL exchange their components (apolipoproteins, phospholipids, triacylglycerols and cholesteryl esters) with the triglyceride-rich particles: the chylomicrons, VLDL, and the remnant particles.

Advanced concept box Peroxisome proliferator-activated receptors control carbohydrate and lipid metabolism

Peroxisome proliferator-activated receptors (PPARs) belong to the superfamily of nuclear receptors that function as transcription factors. They regulate genes controlling carbohydrate and lipid homeostasis – and thus play a role in controlling energy metabolism. PPARs form dimers with retinoid X receptor (RXR); and the dimer subsequently binds to response elements in the promoter regions of target genes. The majority of their activating ligands derive from the metabolism of fatty acids.

The action of PPARα is associated also with catabolism of fatty acids. It stimulates fatty acid catabolism, ketogenesis and gluconeogenesis. It is also involved in the assembly of lipoproteins and in cholesterol metabolism. PPARα regulates expression of the AI-CIII-AIV gene cluster; this increases the expression of apoAI and apoAII and reduces the expression of apoCIII gene. It also increases the expression of the LPL.

PPARβ/δ is involved in the control of cell proliferation and differentiation, and also in fatty acid catabolism. PPARγ influences energy homeostasis and adipose tissue differentiation. Its action improves insulin sensitivity. The actions of both PPARα and PPARγ are anti-inflammatory.

The PPARs are important targets of drug action. The commonly used lipid-lowering drugs, derivatives of fibric acid (fibrates) activate PPARα (see also Chapter 20). The anti-diabetic drugs thiazolidinediones activate PPAR γ.

Cholesterol is removed from cells to the HDL by specific transporter molecules

The HDL are formed as discoid, lipid-poor particles (pre-β-HDL) which contain mainly apoAI; they are partly constructed from the excess phospholipids shed from the VLDL during their hydrolysis by LPL. These nascent HDL particles accept cholesterol from cells through the action of a membrane protein known as the ATP-binding cassette transporter A1 (ABCA1; Chapter 8). ABCA1 uses ATP as a source of energy and is rate-limiting for the efflux of free cholesterol to apoAI. Another ATP-binding cassette transporter, ABCG1, transfers cholesterol from cells to mature HDL particles. Others still, known as ABCG5 and ABCG8, reside in the apical membranes of hepatocytes where they control the transfer of cholesterol to bile (Chapter 17).

Cholesteryl ester exchange between HDL and triglyceride-rich particles is important in humans

The free cholesterol acquired by the nascent HDL is esterified by LCAT. The cholesteryl esters move into the interior of the particle, which enlarges and becomes spherical – it is known as the HDL3. Then, aided by the CETP, it transfers some cholesteryl esters to triglyceride-rich lipoproteins in exchange for triglycerides. The acquisition of triglycerides enlarges the particle further – it now becomes HDL-2. Note that the CETP-mediated exchange re-inserts the cholesterol into the fuel transport pathway, and channels it back to the liver. This is important, because the exchange route seems to be the **main pathway of the reverse cholesterol transport in humans**.

The cholesterol remaining in the HDL-2 is also carried to the liver. There, the HDL-2 binds to the class B scavenger receptor, transferring cholesterol to the cell membrane. Note that the particle binds to the receptor but it is not taken up by cells. The transfer of cholesterol returns it to a smaller size. The now redundant parts of its ‘shell’ become the nascent HDL ready for the next cycle of transport. The HDL-2 can also be further digested by the HTGL, yielding a subclass of small HDL-3.

HDL also exert anti-atherogenic actions other than reverse cholesterol transport

Apart from the reverse cholesterol transport, the HDL has other atheroprotective properties. For instance, they increase the production of nitric oxide (NO) by

activating the endothelial nitric oxide synthase (eNOS). They also exert anti-inflammatory and free radical (reactive oxygen species, ROS)-scavenging actions, promote the integrity of the endothelial layer, and prevent endothelial adhesion of cells, platelet aggregation, and thrombosis.

Dyslipidemias

Defects in the lipoprotein metabolism lead to disorders known as dyslipidemias synonymously but less accurately called hyperlipidemias. Their original, now outdated, classification into types I to V had been based on the electrophoretic behavior of lipoproteins (Table 18.3). The electrophoretic classification was superseded by the genetic one (Table 18.4). Another clinically commonly used phenotypic classification divides the dyslipidemias into hypercholesterolemia, hypertriglyceridemia and mixed dyslipidemia. Figure 18.6 presents an overview of the abnormalities of lipoprotein metabolism.

Table 18.3

Phenotypic classification of dyslipidemias

Dyslipidemia type (Fredrickson)	Increased electrophoretic fraction (lipoprotein type)	Increased cholesterol	Increased triglycerides
I	Chylomicrons	Yes	Yes
IIa	Beta (LDL)	Yes	No
IIb	Pre-beta & beta (VLDL & LDL)	Yes	Yes
III	'Broad beta' band (IDL)	Yes	Yes
IV	Pre-beta (VLDL)	No	Yes
V	Pre-beta (VLDL) plus chylomicrons	Yes	Yes

On electrophoresis the α-lipoproteins (HDL) migrate furthest towards the anode ((+) electrode), followed by the pre-β-lipoproteins (VLDL) and the β-lipoproteins (LDL). The chylomicrons remain at the cathodic end, at the origin of the electrophoretic strip.

This is the classification developed by Fredrickson and adopted by the WHO; it is based on the electrophoretic separation of serum lipoproteins. It has been largely superseded by the genetic classification. Dyslipidemias are also simply classified as hypercholesterolemia, hypertriglyceridemia or mixed dyslipidemia.

Table 18.4

The most important genetic dyslipidemias

Dyslipidemia	Frequency/inheritance	Defect	Plasma lipid pattern	Increased cardiovascular risk
Familial hypercholesterolemia	1:500 Autosomal dominant	LDL receptor deficiency or functional impairment	Hypercholesterolemia or mixed hyperlipidemia (IIa or IIb)	Yes
Familial combined hyperlipidemia	1:50 Autosomal dominant	Overproduction of apoB100	Hypercholesterolemia or mixed hyperlipidemia (IIa or IIb) Characteristically variable patterns in different family members	Yes
Familial dysbetalipoproteinemia (type III hyperlipidemia)	1:5000 Autosomal recessive	Presence of APO E2/E2 genotype. Defective remnant binding to LDL receptor	Mixed hyperlipidemia	Yes

Mixed hyperlipidemia: increased both plasma cholesterol and plasma triglyceride concentration.

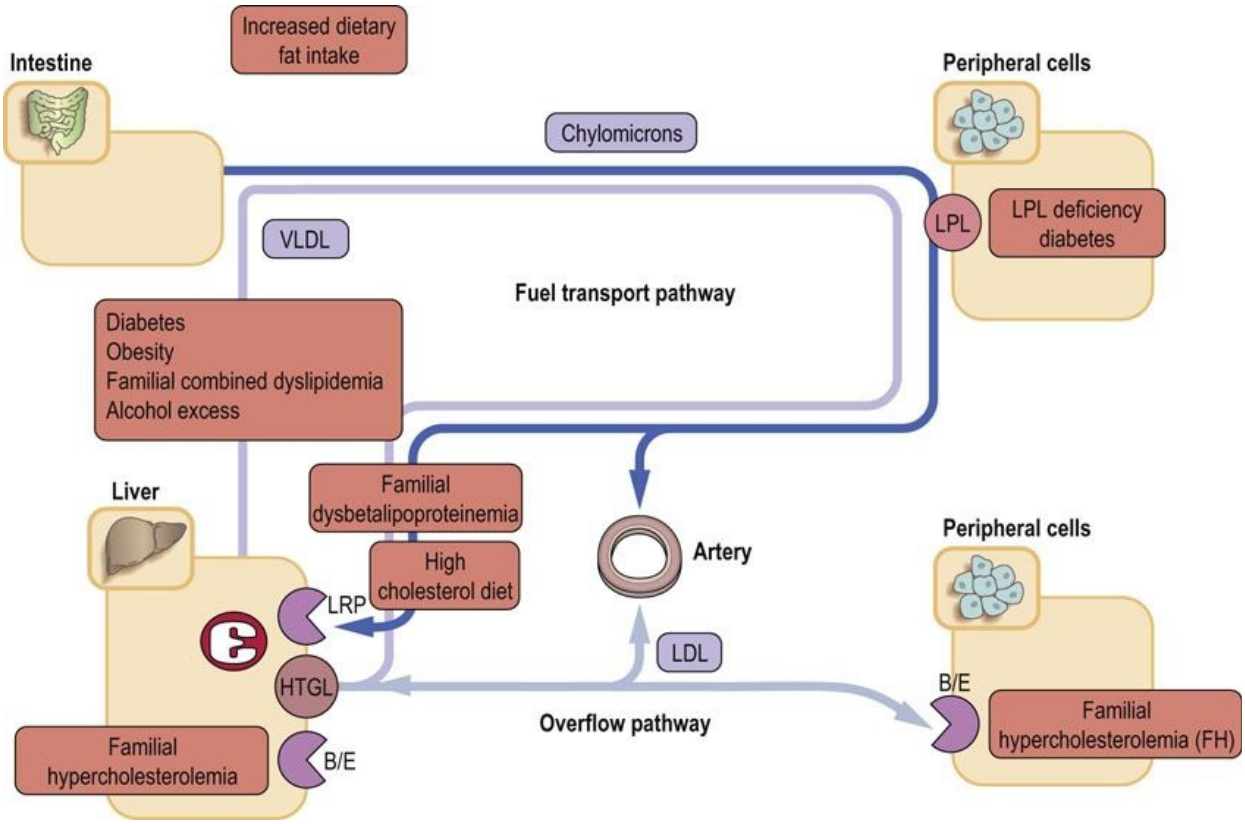


FIG. 18.6 An overview of the abnormalities of lipoprotein metabolism.
Compare this figure to Figure 18.3.
Conditions that primarily affect the fuel transport pathway: The fuel transport pathway is affected by excessive dietary intake of fats, obesity and diabetes. The LPL deficiency causes extreme elevation of chylomicrons and VLDL. Importantly, in this pathway the lipoprotein concentrations change with fast-feed cycle. The concentrations of atherogenic remnants increase postprandially and may be important in atherogenesis. Familial dysbetalipoproteinemia leads to an increased remnant concentration because of impaired uptake caused by the mutation in apoE.
Conditions that primarily affect the overflow pathway: The plasma LDL concentration can be affected by its increased generation (a HTGL-mediated digestion of VLDL remnants) from the fuel transport pathway, or by their impaired cellular uptake. The most important defect is the impaired uptake caused by mutations in the apoB/E receptor gene, which causes familial hypercholesterolemia (FH). In FH the overflow pathway is primarily affected, and plasma LDL concentration is elevated. Note, however, that the remnant uptake is also impaired.
Conditions that affect both pathways: Familial combined hyperlipidemia affects both pathways because of the increased synthesis of the VLDL and consequent increase in LDL generation. Similarly, a high-fat diet affects both pathways.

Clinical box Genetically determined dyslipidemias

In the industrialized countries, approximately 30% of people have undesirably high plasma cholesterol concentration. The most frequent dyslipidemia (known as the **common hypercholesterolemia**) is polygenic and is a result of combined genetic and environmental factors. Then there are rarer disorders with defined genetic background. The most important ones are mentioned below.

Familial hypercholesterolemia is a monogenic disorder caused by a mutation in the gene coding for the apoB/E receptor and it affects the overflow pathway. Cellular uptake of the remnant particles and the LDL is either impaired (heterozygous FH) or completely inhibited (a very rare homozygous FH). Other mutations disrupt LDL receptor recycling to plasma membrane. Patients with FH have very high plasma cholesterol and LDL-C concentrations. The mode of inheritance is autosomal dominant, so there usually is a prominent family history of early heart disease. Some patients develop lipid deposits on hand and knee tendons, and particularly on the Achilles tendons: these are known as xanthomas, and are diagnostic for the disorder. Familial hypercholesterolemia carries a high risk of early cardiovascular disease (note: premature infarction is defined as one occurring in a man aged below 55 years or a woman below 65 years).

A mutation in the gene coding for the LDL-receptor-related protein 6 has also been linked to premature coronary heart disease, showing autosomal dominant inheritance.

Familial combined hyperlipidemia is characterized by an overproduction of apoB100 rather than the impairment of the receptor-mediated clearance. There is an increased production of the VLDL and, consequently, increased generation of the LDL: again, both the fuel transport and the overflow pathways are affected. This dyslipidemia presents with variable plasma lipid patterns (either with hypercholesterolemia alone or hypercholesterolemia with hypertriglyceridemia). Familial combined hyperlipidemia is a relatively common cause of premature myocardial infarctions.

Familial dysbetalipoproteinemia affects primarily fuel transport pathway and is caused by a mutation in the apoE gene,

yielding apoE isoform with low affinity for the apoB/E receptor. In this condition, the remnants accumulate, and there is an increase in both cholesterol and triglyceride concentration in plasma. Characteristic palmar xanthomas are present. Familial dyslipidemia is associated with early coronary disease.

Lipoprotein lipase deficiency is a very rare dyslipidemia that affects the fuel transport pathway and is caused by the deficiency of LPL. Chylomicron and VLDL accumulation leads to very high plasma triacylglycerol concentrations. Clinical signs include characteristic rash-like skin xanthomas. The risk associated with the LPL deficiency is primarily that of pancreatic inflammation (pancreatitis, see Chapter 9) caused by the very high triacylglycerol concentrations.

Abetalipoproteinemia. Mutations in the gene coding for apoB can also lead to low VLDL, and consequently low LDL concentrations. A very rare dyslipidemia known as abetalipoproteinemia is associated with the mutation of the gene coding for the microsomal transfer protein (MTP), involved in the cellular assembly of the VLDL.

Clinical box The diagnosis of familial hypercholesterolemia

The UK Simon Broome criteria for the diagnosis of familial hypercholesterolemia used in Britain are:

- Plasma total cholesterol above 7.5 mmol/L (290mg/dL) or LDL-cholesterol above 4.9 mmol/L (189 mg/dL) in an adult.
- Cholesterol above 4 mmol/L (154 mg/dL) in a child under 16.

plus:

- Tendon xanthomas in a patient or first-degree relative (parent, sibling, child) or in a second-degree relative (grandparent, uncle, aunt).

or:

- DNA-based evidence of an LDL receptor mutation, familial-

defective apoB100 or PSK9 gene mutation.

Currently, the basic genetic screening to support the diagnosis of FH involves searching for several mutations of the LDL-receptor gene: one is the sequence1637G>A, which results in the substitution of glycine by aspartic acid (Gly546Asp). It results in a reduced activity of the LDL receptor. Another mutation of the APOB gene, Arg3527Gln, has been found in 5-7% of the FH patients, as well as a less frequent mutation of the PCSK9 gene, Asp374Tyr.

Clinical box Dyslipidemia is common in diabetes mellitus

Mr B is 67 years old, overweight (BMI 28 kg/m^2) and has type 2 diabetes and mild hypertension. When he visited the outpatient clinic, his cholesterol concentration was 6.9 mmol/L (265 mg/dL), triglycerides were 1.9 mmol/L (173 mg/dL), and HDL-C was 0.9 mmol/L (35 mg/dL). Fasting blood glucose was 8.5 mmol/L (153 mg/dL) and hemoglobin A_{1c} ($HbA1_{1c}$) 31 mmol/mol (desirable value below 48 mmol/mol). He was being treated with diet and metformin, which improves insulin sensitivity and lowers blood glucose concentration.

Comment.

Diabetes carries a 2–3 times increased risk of coronary heart disease. This patient's diabetes was well controlled (judged by the HbA_{1C} concentration) but his cholesterol level remained high, so that he required lipid-lowering drug treatment. Low HDL-C concentration is relatively common in type 2 diabetes. The patient was prescribed a statin in addition to metformin. Blood pressure responded to treatment with an angiotensin-converting enzyme inhibitor (see Chapter 20).

Clinical box Familial hypercholesterolemia is a cause of early myocardial infarctions

A 32-year-old man who smoked heavily developed a sudden crushing chest pain. He was admitted to the casualty department. Myocardial infarction was confirmed by EKG changes and by high cardiac troponin concentration. On examination the patient had tendon xanthomas on hands and thickened Achilles tendons. There was a strong family history of coronary heart disease (his father had had a coronary bypass graft at the age of 40 and his paternal grandfather died of myocardial infarction in his early fifties). His cholesterol was 10.0 mmol/L (390 mg/dL), triglycerides 2 mmol/L (182 mg/dL) and HDL-C 1.0 mmol/L (38 mg/dL).

Comment.

This patient has **familial hypercholesterolemia (FH)**, an autosomal dominant disorder characterized by a decreased number of LDL receptors. FH carries a very high **risk of premature coronary disease**, and heterozygotes may suffer heart attacks as early as the 3rd or 4th decade of life. The frequency of FH homozygotes in Western populations is approximately 1 : 500. This patient was immediately given an intravenous thrombolytic treatment. Subsequently he underwent coronary artery bypass graft and was treated with lipid-lowering drugs. His cholesterol concentration decreased to 4.8 mmol/L (185 mg/dL) and triglyceride to 1.7 mmol/L, with HDL-C increasing to 1.1 mmol/L (42 mg/dL).

Note: An early myocardial infarction is the one occurring in a man aged below 55 or a woman younger than 65 years.

Conditions affecting the fuel transport pathway

Increased flux through the fuel transport pathway occurs in obesity and diabetes mellitus

The increased flux through the fuel transport pathway is most commonly due to increased synthesis of the VLDL. This happens in two common conditions, obesity and diabetes mellitus. Diabetic dyslipidemia also affects the reverse cholesterol transport; a frequent pattern seen in diabetes is an increase in plasma triglyceride concentration combined with a decrease in HDL-C. In diabetes, the overflow pathway remains relatively unaffected: patients often have normal LDL-cholesterol (LDL-C) concentration. However, diabetic patients seem to generate the small dense LDL, and therefore diabetic LDL, although not increased, may be more atherogenic than the nondiabetic particles. Importantly, weight loss decreases the activity of this pathway. The combination of increased VLDL remnants (resulting in mild hypertriglyceridemia), increased sd-LDL and low HDL is sometimes referred to as ‘the atherogenic triad’.

Another common condition causing increase in VLDL production is alcohol abuse. However, in contrast to diabetes, alcohol, while raising VLDL, also raises the HDL concentration.

LPL deficiency leads to extreme hypertriglyceridemia

The fuel transport pathway also becomes overloaded if the hydrolysis of chylomicrons or VLDL is inefficient. In diabetes, the lack of insulin suppresses LPL, and contributes to the development of hypertriglyceridemia. A very rare condition, the LPL deficiency, results in extremely high plasma triglyceride concentrations, which may exceed 100 mmol/L (8850 mg/dL).

Impaired metabolism of remnant particles leads to familial dysbetalipoproteinemia

The rare but important condition affecting the very last step of the fuel transport pathway, the uptake of remnants, is an inherited condition known as familial dysbetalipoproteinemia. It is caused by the mutation of the remnants ‘driver’ apolipoprotein, apoE (the presence of E2/E2 phenotype), which decreases its binding to the apoB/E receptor. Familial dysbetalipoproteinemia is linked to premature CVD and was previously known as Type III hyperlipidemia.

Conditions affecting the overflow pathway

High LDL-C concentration can be a result of its overproduction or impaired cellular uptake

Abnormalities of the overflow pathway lead to hypercholesterolemias, the most important of which is the familial **hypercholesterolemia (FH)**. FH is characterized by defective cellular uptake of the LDL due to abnormalities (or the lack) of the apoB/E receptor caused by mutations in the relevant genes (see Box). Patients with FH have a high LDL concentration and some present with tendon xanthomas – deposits of cholesteryl esters in the Achilles tendons and tendons of the hands. FH is associated with premature cardiovascular disease and patients frequently present with a prominent family history of early coronary disease. **Familial defective apolipoprotein B** (FDB) is, on the other hand, caused by a mutation of the apoB100 molecule that impairs its binding to the receptor.

An increase in LDL can also be secondary to increased VLDL synthesis (according to the 'overflow' concept) and the consequent increased transformation of remnants into LDL. This happens in a condition known as the **familial combined dyslipidemia**. Finally, dietary intake of saturated fats affects the LDL concentration. A low-fat diet can decrease the LDL, and plasma cholesterol concentration, by 10–15%.

Note that the majority of mild-to-moderate hypercholesterolemias seen in clinical practice are polygenic (see Box on p. 225).

Conditions affecting the reverse cholesterol transport

Several rare mutations lead to a decreased HDL cholesterol concentration

Low HDL-cholesterol can result from mutations in genes coding for the apoA1, the ABCA1 transporter and the LCAT. Patients with apoAI deficiency present with low plasma HDL-C accompanied by xanthelasmas, corneal clouding, and arteriosclerosis. Heterozygotes occur in 1% of the population. They also develop amyloidosis.

Those with ABCA1 mutations, apart from low plasma HDL-C have large, orange-colored tonsils, hepatosplenomegaly, peripheral neuropathy and thrombocytopenia. It is known as Tangier disease.

The deficiency of the LCAT is known as fish-eye disease. It is characterized by HDL deficiency and also by corneal clouding, nephropathy and hemolytic anemia. In contrast, the deficiency of CETP leads to high HDL concentration.

Atherosclerosis, atherogenesis and atherothrombosis

Atherosclerosis is a process that leads to the narrowing, or a sudden complete occlusion, of the arterial lumen. The narrowing is due to slow-growing atherosclerotic plaques, whereas a sudden complete occlusion is caused by the thrombus that forms over a ruptured plaque. The complete occlusion may cause **myocardial infarction** (if the blockage is in a coronary artery), **stroke** (blockage in an artery supplying the brain) or **peripheral vascular disease** (blockage in leg arteries; this leads to a characteristic pain that occurs during walking, with fast relief when the person stops, known as intermittent claudication). The development of atherosclerotic plaque is termed **atherogenesis**. The term **atherothrombosis** is sometimes used to emphasize links between plaque growth and blood coagulation processes.

Atherogenesis involves lipid transport and deposition in the subendothelial layer of the arterial wall (intima). This occurs on the background of endothelial damage and is accompanied by inflammatory reaction that affects the intima, and involves elements of innate and adaptive immunity (Chapter 38). It leads to remodeling of the arterial wall, including new vessel formation (angiogenesis). Thrombosis (Chapter 7) is important in plaque maturation and destabilization (Fig.18.7).

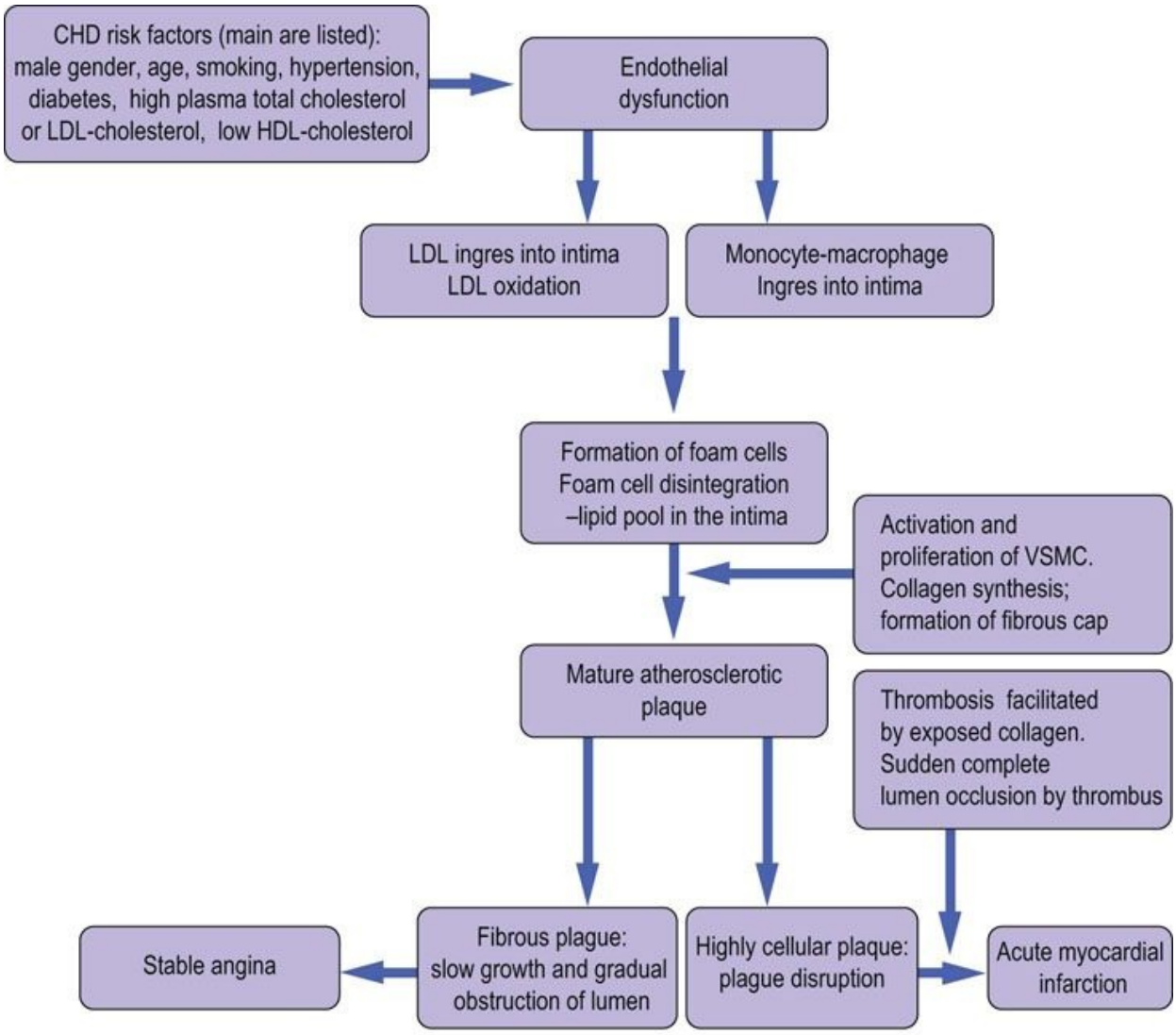


FIG. 18.7 The development of the atherosclerotic plaque.
The cardiovascular risk factors damage the endothelium. This leads to penetration of lipids and cells into the intima. Activation of macrophages leads to cytokine response and leads to activation and proliferation of smooth muscle cells. The lipid pool which results from disintegrating lipid-filled macrophages (the foam cells) becomes covered by a fibrous cap, forming the mature plaque. While stable fibrous plaques cause a slowly-progressing angina, the disruption of an unstable, highly cellular plaque leads to acute clinical events such as a myocardial infarction. VSMC, vascular smooth muscle cells.

Advanced concept box Plasma C-reactive protein concentration reflects chronic low-grade inflammation

associated with atherogenesis

The inflammatory reaction associated with infection can be detected by measuring the concentration of plasma C-reactive protein (CRP), a protein synthesized in the liver (but also in VSMC and in the endothelium) in response to stimulation by pro-inflammatory cytokines. Its name comes from its binding to the capsular (C) polysaccharide of bacteria such as *S. pneumoniae*, by which way it mediates their clearance.

Minute increases in CRP concentration, which require a highly sensitive (hs) analytical method able to detect CRP concentrations below 10 mg/L, may reflect chronic inflammatory processes in the vascular walls. Epidemiologic studies demonstrated an association between the hsCRP concentration and cardiovascular events. Importantly, this association is independent of the link between plasma cholesterol concentration and coronary disease. Increased plasma concentrations of other pro-inflammatory molecules, such as **interleukin-6** (IL-6) and **serum amyloid A**, have also been linked to coronary heart disease. hsCRP measurement enhances the predictive value of the LDL-C and its suggested use is to fine-tune cardiovascular risk assessment. It has been recently observed that adding either CRP or fibrinogen to the panel of conventional risk factors improves risk assessment.

It is suggested that CRP <1 mg/dL signifies low CVD risk, while CRP >3 mg/L is associated with high risk of coronary disease.

The role of vascular endothelium

Normal endothelium has anticoagulant and anti-adhesion properties

The lumen of a healthy artery is lined by a confluent layer of endothelial cells. Normal endothelial surface is strongly antithrombotic and anti-adhesive: it repels cells floating in plasma. The arterial wall itself consists of three layers: the

subendothelial layer (the intima), the middle one (the media, which contains vascular smooth muscle cells, VSMC), and the outer layer (the adventitia, composed of looser connective tissue and containing relevant nerves). Substances can penetrate the endothelium either through junctions between the endothelial cells or by transgressing the cells themselves. Particles with a diameter greater than approximately 60–80 nm can lodge in the vascular wall.

Endothelium controls vasodilatation by secreting nitric oxide known as the endothelium-derived relaxing factor (EDRF)

The endothelium controls the ability of blood vessels to dilate (vasodilatation) and to constrict (vasoconstriction) and thus to regulate tissue and organ blood flow. Nitric oxide is the most important vasodilatory substance. It is synthesized from L-arginine by the endothelial NO synthase (eNOS). Activity of eNOS is controlled by the intracellular calcium concentration. The eNOS is constitutively (constantly) expressed in the endothelium, while another isoenzyme, the inducible NOS (iNOS), is found in VSMC and in macrophages. NO signals via guanylate cyclase and cyclic GMP (see Chapters 37 and 40). A decrease in NO production contributes to arterial hypertension. The drug **glyceryl trinitrate**, commonly used to relieve chest pain caused by inadequate oxygen supply to the heart muscle (causing the characteristic chest pain known as **angina pectoris**), dilates coronary arteries by stimulating NO release.

Endothelial dysfunction, lipid deposition, and inflammatory reaction in the vascular wall are the key processes in atherogenesis

Crucially, there is cross-talk between endothelial cells, VSMC, plasma-derived inflammatory cells (monocytes, macrophages and lymphocytes) and blood platelets, all of which secrete an array of chemokines, cytokines, and growth factors (Chapter 42). This attracts cells to the atherosclerotic lesions, induces cell migration, proliferation, apoptosis, and production of collagenous extracellular matrix (Chapter 27, Fig. 27.9).

Endothelial dysfunction precedes formation of atherosclerotic lesions

Atherogenesis is initiated by endothelial damage (Figs 18.8 and 18.9). Initially, the damage is functional rather than structural. The endothelium switches from anti-atherogenic to pro-atherogenic, inflammatory phenotype by losing its cell-repellent quality, and admits inflammatory cells into the vascular wall. It also becomes more permeable to lipoproteins, which subsequently deposit in the intima. Later, there is structural damage or a complete destruction of endothelial cells.

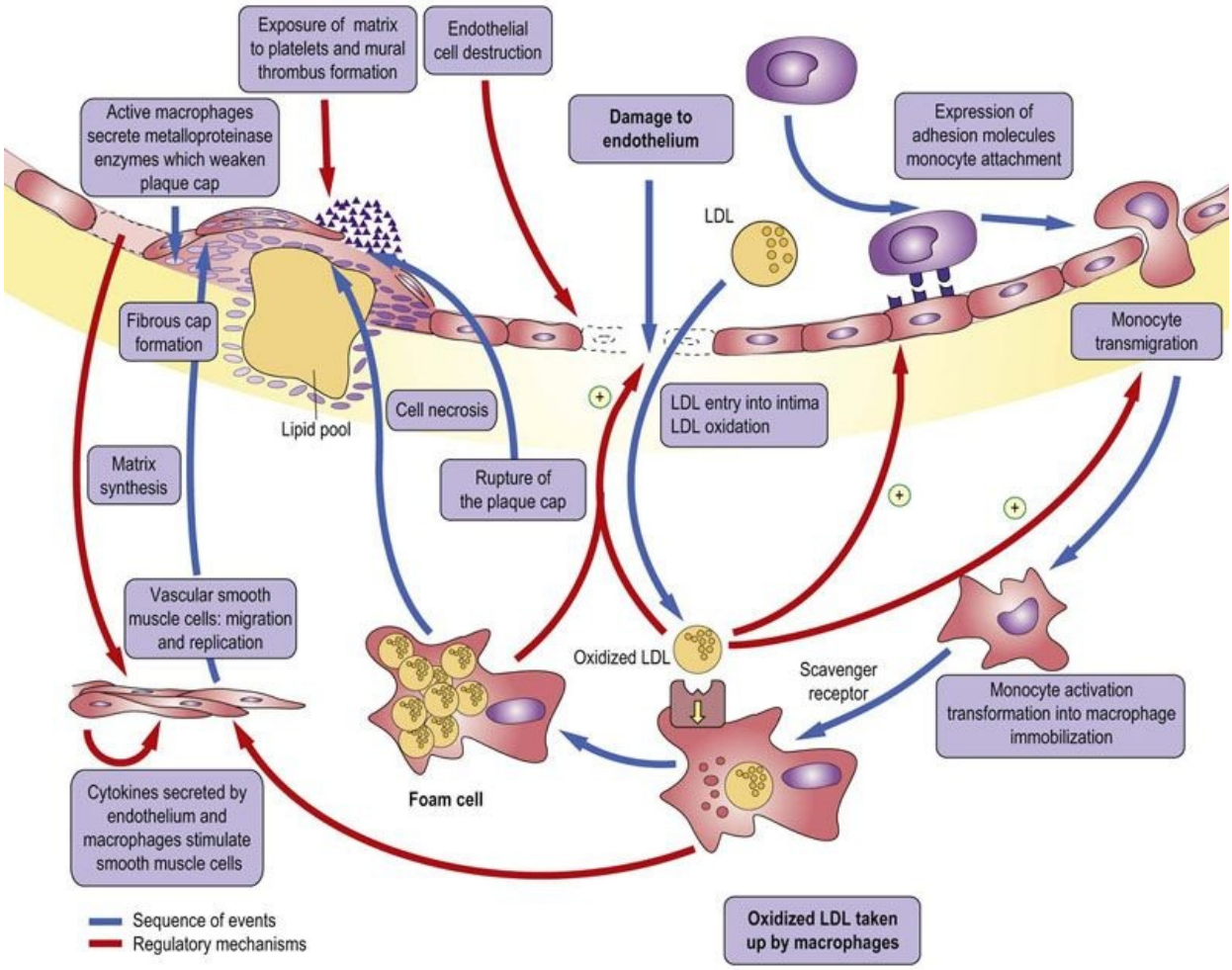


FIG.18.8 Atherogenesis: the process.
Atherogenesis involves endothelial dysfunction, deposition of lipids in the arterial intima, low-grade inflammatory reaction, migration and proliferation of the vascular smooth muscle cells, and thrombosis. Note the role of oxidized lipids in the formation of lipid-laden cells and subsequently the lipid pool that becomes the center of the atherosclerotic plaque. The sequence of events is described in the text.

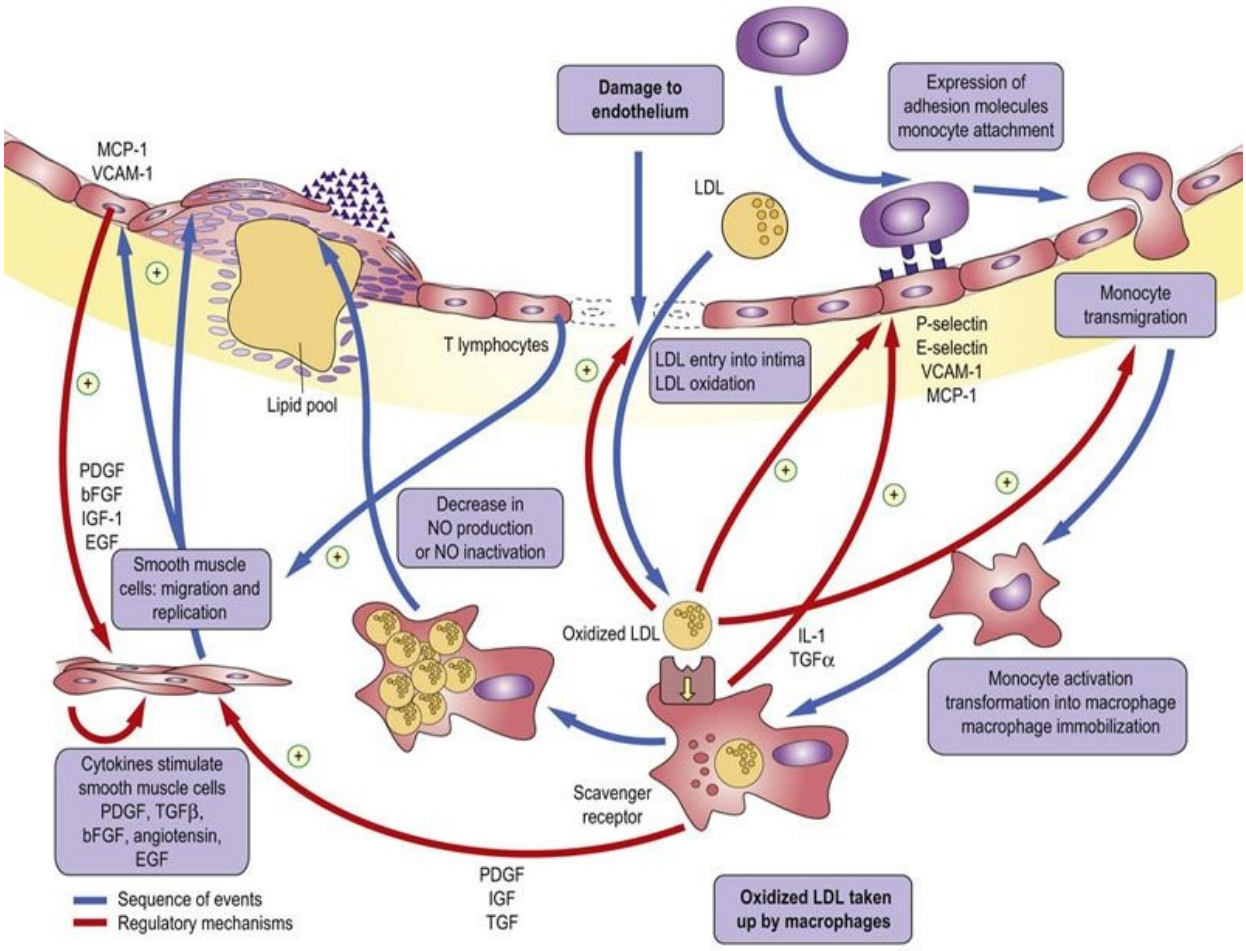


FIG. 18.9 Atherogenesis: the role of growth factors and cytokines. Atherogenesis is driven by signals mediated by cytokines and growth factors generated by endothelial cells, macrophages, T lymphocytes and vascular smooth muscle cells (VSMC). There are multiple activation paths: for instance, the expression of MCP-1 and VCAM-1 may be stimulated by signals generated by the macrophages as well as by the oxidized LDL. VSMC may be stimulated by the dysfunctional endothelial cells, by macrophages, and by T lymphocytes (note also the autocrine activation). A hormone, angiotensin II, also participates in these processes. MCP-1, monocyte chemoattractant protein 1; VCAM-1, vascular cell adhesion molecule 1; ICAM-1, intracellular cell adhesion molecule 1; TNFβ, tumor necrosis factor-β; TNFα, tumor necrosis factor-α; IFNγ, interferon-γ; NO, nitric oxide; PDGF, platelet-derived growth factor; bFGF, basic fibroblast growth factor; IGF-1, insulin-like growth factor 1; EGF, epidermal growth factor; TGFβ, transforming growth factor-β; IL-1, interleukin 1.

Cell adhesion to endothelium is mediated by the adhesion molecules present on its surface. Selectins (P-selectin and E-selectin) mediate the initial interactions with circulating cells. The vascular cell adhesion molecule 1 (VCAM-1) promotes adhesion of monocytes and T lymphocytes. In experimental animals, VCAM-1 deficiency decreases formation of

atherosclerotic plaques. Adhering cells are subsequently stimulated by the monocyte chemoattractant protein 1 (MCP-1) to cross the endothelium and lodge in the intima. This is further facilitated by a protease (matrix metalloproteinase 9, MMP-9), secreted by the monocytes. Production of NO in the damaged endothelium decreases and this promotes vasoconstriction. Since NO normally reduces monocyte adhesion and VSMC migration and proliferation, in its absence these processes intensify.

Importantly, endothelial expression of the adhesion molecules is stimulated by most cardiovascular risk factors, including hypercholesterolemia, hypertension, components of the cigarette smoke, high saturated fat diet and also diabetes mellitus and obesity. Activation of the renin–angiotensin–aldosterone system, apart from its effect on blood pressure, also has pro-atherogenic effects. Angiotensin II (Chapter 23) increases expression of VCAM-1 and MCP-1. Clinical studies show that drugs inhibiting that axis (the ACE inhibitors) are beneficial in cardiovascular prevention.

Advanced concept box Genetics of atherosclerosis

Genes coding for the LDL receptors, apolipoproteins and LRP6 are currently the only ones that have been directly linked with atherosclerotic disorders. Deep sequencing (sequencing the genome a large number of times to minimize error rate) identified two variants of the **PCSK9 gene** (in patients of African descent) that are responsible for low lipid levels and a decreased risk of myocardial infarction. Altogether, approximately 95 loci have been associated with lipids LDL, HDL and triglycerides – their combined effect was responsible for approx 25% of variation in the LDL and HDL concentrations.

However, the majority of cardiovascular diseases are **polygenic**. The current working hypothesis (the common disease variant hypothesis) is that the common variants, occurring with a frequency below 5%, play a role in pathophysiology of polygenic diseases. **The genome-wide association studies** (GWAS) test associations between a disease and common variants across the entire genome. There are presently about 30 loci associated with myocardial infarction and coronary artery disease.

The technical ability to perform genome-wide scans changed the way this type of research is conducted. Formerly, accumulation of results from many single studies would lead to elucidation of mechanisms and then perhaps a search for their role in disease. In contrast, genome-wide studies often provide evidence of association between a pathologic condition and a gene before anything is known about the underlying mechanisms. Only later investigators would search for processes that underpin such association.

Monocytes migrate into the intima and transform into resident macrophages

Monocytes are attracted to the developing plaques by the chemoattractant cytokine CCL2, which binds to their receptors. At the next stage, monocytes transform into macrophages, under the influence of interferon-γ, tumor necrosis factor-α (TNFα), granulocyte-macrophage colony-stimulating factor, and the monocyte colony-stimulating factor 1 (MCSF-1), secreted by the endothelial cells and the VSMC. Macrophages also generate reactive oxygen species, which oxidize LDL in the intima. Some macrophages produce cytokines such as interleukin 1β (IL-1β), IL-6 and TNFα. Other express scavenger receptors (class A receptors, and CD36), which become active in endocytosis of the oxidized LDL.

Lipids enter the intima

The smaller lipoproteins, *i.e.* the remnants and the LDL, enter the vascular wall more easily than other particles, and thus are most atherogenic. While in the plasma, the LDL particles are protected against oxidation by antioxidants such as vitamin C and β-carotene. In the intima, however, they are deposited in association with proteoglycans. This removes their access to antioxidants, and the fatty acids and phospholipids in the LDL become prone to oxidation mediated by lipoxygenases, myeloperoxidase, and NADPH oxidases expressed by the macrophages. Oxidized LDL further stimulates expression of VCAM-1 and MCP-1, maintaining the ingress of cells into intima; they are also mitogenic for macrophages. The apoB100, once oxidized, binds to the scavenger receptors

rather than to the apoB/E receptor on cells. Because the scavenger receptors are not feedback-regulated by intracellular cholesterol, the macrophages overload themselves with oxidized lipids and take the appearance of **foam cells**; conglomerates of such cells are visible in the arterial walls as the so-called **fatty streaks**. Dying foam cells release lipids, which form pools within the intima. These become centers of mature atherosclerotic plaques.

Migrating vascular smooth muscle cells change the structure of the vascular wall

Growth factors secreted by the endothelial cells and macrophages (the platelet-derived growth factor PDGF, the epidermal growth factor EGF, and the insulin-like growth factor 1, IGF-1) activate the VSMC present in the arterial media. Under the influence of PDGF and TGFβ, the VSMC proliferate and migrate into the intima. Migration is further facilitated by the MMP-9. The cells secrete adhesion molecules, as well as inflammatory cytokines IL-1 and TNFα. Activated VSMC also synthesize extracellular matrix, in particular collagen, which deposits in the growing plaque. All this disrupts the structure of the arterial wall: a newly formed plaque may protrude into the lumen of the artery, obstructing the blood flow.

Inflammation is fundamental to atherogenesis

The exit of the monocytes and T leukocytes from plasma and their activation in the intima are parts of the inflammatory response. Normally, such response is initiated by an antigen or trauma. Intriguingly, no specific antigen capable of initiating atherogenesis has been identified. There could be a molecular mimicry between such putative antigen(s) and the exogenous pathogens (Chapter 38). The antigen(s) might be infectious agents or modified molecules generated by reactive oxygen species. For instance, the phosphorylcholine group found in the oxidized LDL is also a component of the capsular polysaccharide of bacteria. The oxidized LDL remains a candidate antigen that could be responsible for the stimulation of inflammatory reaction in atherogenesis.

Atherogenesis involves both innate and adaptive immunity. Innate immunity includes recognition of molecules by scavenger receptors A and the membrane glycoprotein CD36. When molecules that possess patterns encoded in immune memory bind to these receptors, they activate cells through, for instance, the pathway involving the transcription factor NFκB. With regard to adaptive

immunity, the T cells predominantly of CD4 subtype are present in atherosclerotic lesions, and circulating IgG and IgM-type antibodies against modified LDL have been identified in plasma.

Macrophages (and foam cells) present in the plaque continue to secrete cytokines, growth factors and adhesion molecules (IL-1β and TNFα, VCAM-1, IL8, IL-6) and the MMPs. IL-18 stimulates the secretion of the interferon-γ, which in turn stimulates chemokine-inducible protein 10, and the T-cell-α chemoattractant, which further facilitate T-cell ingress. Interferon also facilitates activation of T-helper 1 cells to effector cells that then secrete CD40 ligand (CD40L), a cytokine that is a member of the TNF family.

Lastly, there is new vessel formation: the newly formed vessels facilitate intraplaque hemorrhages. The thrombin generated there activates monocytes, macrophages, endothelial cells, VSMC and platelets to secrete inflammatory mediators such as CD40L (CD40L binding to CD40 receptor further increases secretion of MMPs, cytokines and adhesion molecules).

The adipokines secreted by the adipose tissue may also contribute to the atherogenic milieu. Adiponectin (Chapter 22) is an insulin-sensitizing adipokine. It has anti-inflammatory actions. It stimulates maturation of preadipocytes and decreases adipocyte mass. Its effects on plasma lipids include a decrease in plasma triglyceride concentration, and an increase in HDL-C. It also increases the activity of HTGL and LPL. Besides, it decreases the expression of VCAM-1, the scavenger receptor and TNFα. Obesity decreases adiponectin concentration, de-inhibiting pro-atherogenic processes.

The role of platelets

Platelets mediate leukocyte adhesion and secrete CD40L. Initial adhesion of platelets occurs through platelet glycoprotein receptors to von Willebrand factor and fibrinogen. Adhesion is further mediated by β_3-integrins. The adherent platelets contribute to inflammation by secreting cytokines, chemokines, growth factors and adhesion molecules. There is binding between platelets and circulating cells, and such aggregates increase leukocyte activation, adhesion and cell migration.

The role of prostaglandins and leukotrienes

Prostaglandin synthesis from arachidonic acid is catalyzed by the enzyme cyclooxygenase (COX). COX-1 in platelets is inhibited by aspirin; this decreases

production of pro-thrombotic thromboxane A_2 and is responsible for cardioprotective effect of the drug. The pro-inflammatory prostaglandins such as prostaglandin E_2 (PGE_2) contribute to inflammatory process. Leukotrienes (leukotriene A_4 and B_4: lipids derived from the arachidonic acid) are generated by the enzyme 5-lipoxygenase and contribute to the recruitment of lymphocytes to atherosclerotic lesions.

Inflammatory activity destabilizes the plaque, making it prone to rupture

In the mature plaque (Fig. 18.10) the lipid pool is surrounded by foam cells, lymphocytes and VSMC that have migrated into the intima. The plaque 'cap' contains collagenous matrix synthesized by VSMC. The cap also contains active macrophages and T lymphocytes. The advanced lesions may calcify.

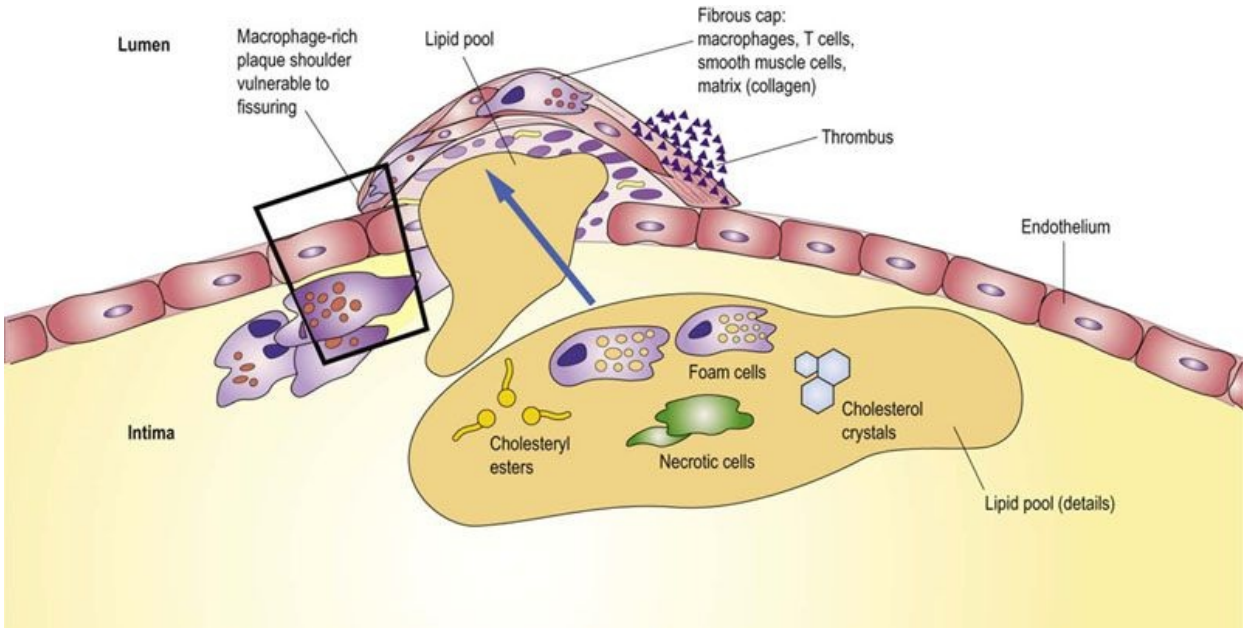


FIG. 18.10 The atherosclerotic plaque.
The lipid center and fibrous cap are the main parts of the mature atherosclerotic plaque which emerges from the structurally remodeled vascular wall. The plaque which is cell-poor and collagen-rich is relatively stable, and grows slowly over the years. On the other hand, the plaque which is cell-rich and collagen-poor becomes unstable and may easily rupture. The key process leading to plaque rupture is digestion of the collagenous matrix of the plaque cap by the metalloproteinase enzymes (MMPs). The figure illustrates areas vulnerable to breakage and shows the obstructing thrombus formed at the rupture site.

Growth of the plaque is accelerated by cycles of plaque mini-ruptures and thrombosis. Active macrophages and T lymphocytes reside preferentially at plaque edges. The unstable plaque has a decreased VSMC content and contains an increased number of macrophages. The macrophages continue to secrete MMPs (collagenases, gelatinases and stromyelysin) that degrade matrix. In addition, lysosomal proteases (cathepsins) degrade collagen and elastin. The activated T cells secrete IFNγ and pro-inflammatory cytokines that induce macrophages to release MMPs and inhibit VSMC collagen synthesis, further weakening the plaque cap. VSMC present in the most vulnerable edge regions of the plaque may undergo apoptosis.

The mature plaque also contains elements of the coagulation system. Tissue factor, a transmembrane cytokine receptor and the primary physiologic trigger of coagulation cascade (Chapter 7), is normally expressed in the VSMC and fibroblasts in the adventitia. In the plaque, it may be expressed in the VSMC and in macrophages. The MMP-9 activity in the plaque stimulates its expression. Apart from this, the tissue factor complexes with the coagulation factor VII (FVII) and this induces cell signaling through the protease activated receptor 2 (PAR2), stimulating a range of events from monocyte, chemotaxis, through VSMC migration and proliferation to angiogenesis and apoptosis.

Finally, thrombin continues to be generated in both the early and advanced plaques, and formation of small thrombi within plaques contributes to their instability. These minute thrombi accelerate plaque growth. On the other hand, after a major rupture, a thrombus forming on the plaque surface may completely occlude the lumen of the affected artery. Such an occlusion cuts off oxygen supply and causes tissue necrosis, precipitating sudden – and sometimes catastrophic – clinical events.

THE Assessment of cardiovascular risk

Cardiovascular risk means the probability of a clinical event occurring within a defined time frame

Cardiovascular risk is the probability that a person will suffer (or die of) heart attack or stroke in a defined period in the future. The main cardiovascular risk factors are listed in Table 18.5. Epidemiologic studies show that the risk of cardiovascular disease is strongly related to plasma concentrations of total cholesterol and LDL-C. It is also inversely related to the concentration of HDL-C. Recent research makes it quite clear that the plasma concentration of triglycerides also contributes to risk, albeit to a lesser extent.

Table 18.5

The main and emerging cardiovascular risk factors

Risk factor	Comment
Male sex	Cardiovascular risk between sexes equalizes in postmenopausal women
Age	In elderly people, age and gender alone may determine the high risk
Smoking	
Hypertension	
High plasma total cholesterol (high LDL-cholesterol)	
Low plasma HDL-cholesterol	
Diabetes mellitus	CVD is the main cause of death in diabetes
Impaired renal function	
Family history of premature CVD	Family history of premature CVD increases the calculated risk by the factor of 1.7–2.0
High plasma apoB, low apoA	Newer studies show that assessment of risk based on apolipoproteins is better than that based on cholesterol concentration
High lp(a)	Refines risk assessment
High hsCRP/fibrinogen	Refine risk assessment
Low adiponectin	Important in obesity and diabetes
Central obesity	
Sedentary lifestyle	
Increased carotid intima-media thickness	
Social deprivation	
Autoimmune inflammatory conditions (rheumatoid arthritis, SLE, psoriasis)	

According to the US National Cholesterol Education Program Adult Treatment Panel III (ATP III), the desirable level of total cholesterol is below 5.2 mmol/L (200 mg/dL) and the optimal level of LDL-C is below 2.6 mmol/L (100 mg/dL). The risk steeply increases when total plasma cholesterol concentration increases above 5.2 mmol/L (200 mg/dL). It seems that there is no lower threshold of cholesterol concentration at which the risk would plateau (in other words, the lower the better).

Plasma lipid concentrations are an essential component of cardiovascular risk assessment

It is now accepted that cholesterol concentration that should be achieved with treatment needs to be lowest in people who are at the greatest risk of cardiovascular events, such as individuals with several risk factors, or those who already have atherosclerosis-related disease, diabetes or renal disease. In such persons the ATP III recommends lowering the LDL-C to below 2.6 mmol/L (100 mg/dL) and even 1.8 mmol/L (70 mg/dL).

The Joint British Societies' Guidelines (JBS2) published in 2005 recommend that the optimal total cholesterol concentration in persons who have an increased cardiovascular risk is 4 mmol/L (155 mg/dL) and optimal LDL-C concentration is 2.0 mmol/L (77 mg/dL). Triglyceride concentration should be no higher than 1.7 mmol/L.

A low concentration of HDL-C also signifies increased risk. HDL-C concentration below 1 mmol/L (40 g/dL) in men or 1.2 mmol/L (47 mg/dL) in women is regarded as low. Conversely, it seems that a concentration above 1.6 mmol/L (60 mg/dL) provides some protection against coronary disease. The principles of lipid testing in clinical practice are summarized in Figure 18.11.

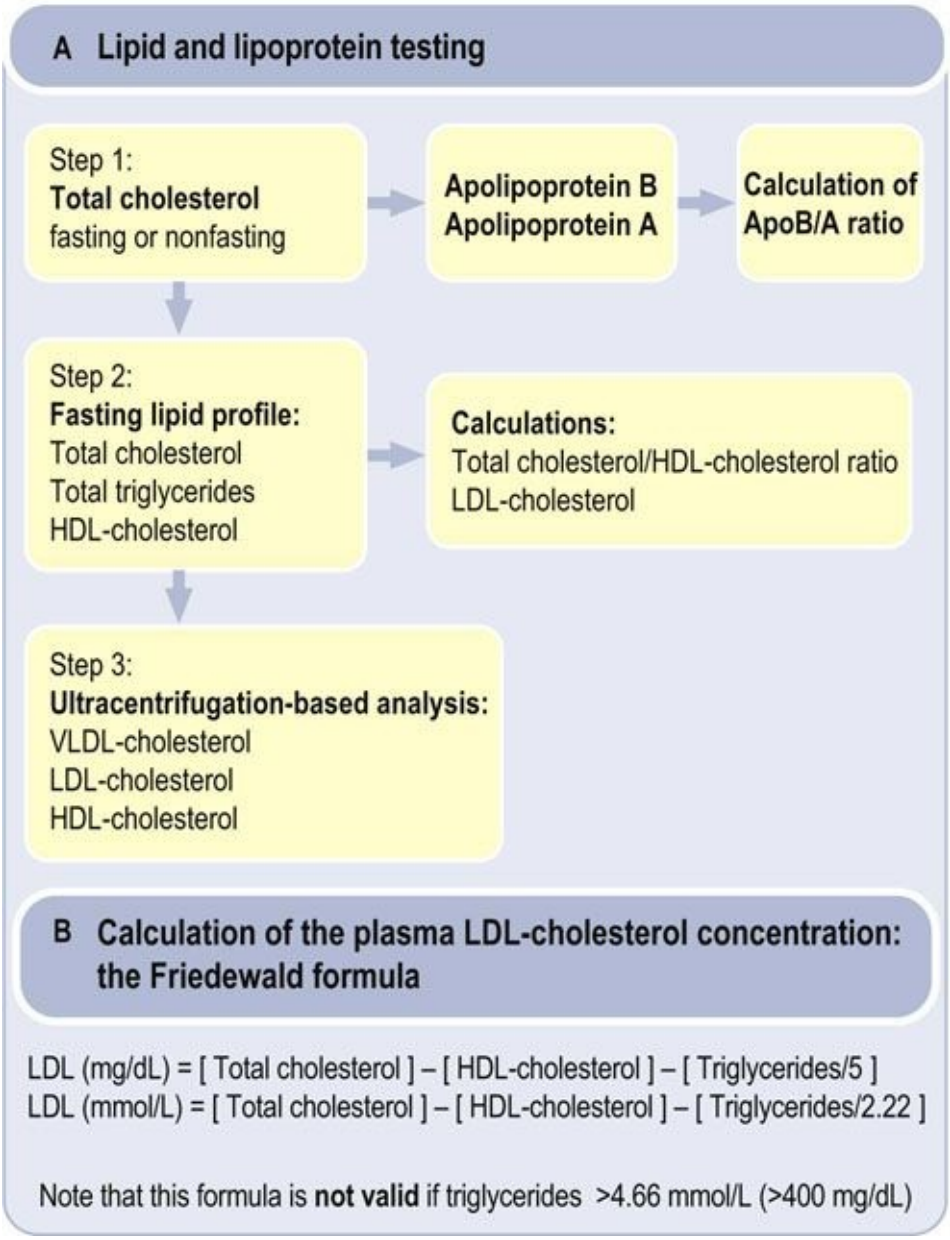


FIG. 18.11 Laboratory diagnosis of dyslipidemias.
(A) The measurement of plasma lipids and apolipoproteins. Several studies show that apolipoprotein B and apolipoprotein A measurements provide better assessment of the lipid-associated cardiovascular risk than total cholesterol and HDL-cholesterol, respectively. **(B)** Calculation of plasma LDL-cholesterol concentration. Note that the commonly used laboratory methods do not measure the concentration of lipoprotein particles such as LDL or HDL. These methods only measure the fraction of total cholesterol which is present in these particles: therefore we talk about plasma LDL-cholesterol (LDL-C), HDL-cholesterol (HDL-C) or VLDL-cholesterol (VLDL-C). Apo, apolipoprotein.

One should remember that the risk associated with plasma concentrations of

the ‘main’ lipids can be modified by other factors, such as increased plasma concentration of Lp(a), fibrinogen or C-reactive protein (hsCRP, see Box). A sobering thought is that about 50% of persons who suffer from myocardial infarction have ‘average’ cholesterol and LDL-C concentrations. The search for new biomarkers continues.

Clinical box Lifestyle change improves plasma lipid profile

A 57-year-old man was referred to the lipid clinic because of hypertriglyceridemia. His triglycerides were 6 mmol/L (545 mg/dL), cholesterol was 5 mmol/L (192 mg/dL), and HDL-C was 1 mmol/L (39 mg/dL). He was obese, took 30 units of alcohol per week, and led a sedentary lifestyle. After initial difficulties, he eventually managed to lose 7 kg of weight over 6 months, cut drinking to below 20 units per week, and started to exercise regularly. Twelve months later, his triglycerides were 2.5 mmol/L (227 mg/dL), cholesterol 4.8 mmol/L (186 mg/dL), and HDL-C 1.2 mmol/L (46 mg/dL).

Comment.

Lifestyle change can result in appreciable improvements in the lipid profile. To achieve this, individuals need to become committed to changing lifestyle and to maintaining the change over a prolonged period of time.

Note: 1 unit of alcohol is one measure (60 mL) of liquor, one glass (170 mL) of wine, or a half-pint (300 mL) of beer.

The overall CVD risk is calculated using population-based risk calculators

There is a defined risk of a cardiac event associated with every risk factor. However, what interests the clinician most is the overall risk. Calculations of such risk are the basis for preventive measures and drug treatment. The risk algorithms have been developed on the basis of data from long-term

epidemiologic studies. The most widely used algorithm has been derived from the Framingham Study in the US. It is based on age, the presence of diabetes, smoking habits, systolic blood pressure, and total cholesterol and HDL-C concentrations. It relates to a population aged 34–74 years without CVD at baseline, and involves up to 12 years follow-up. Importantly, this algorithm does not take into account the family history of early cardiovascular disease.

In Europe, there is the Assign score, the Systematic Coronary Risk Evaluation (SCORE), as well as the Prospective Cardiovascular Munster Study (PROCAM) algorithm. The SCORE algorithm assesses a10-year risk of a first fatal event. The definition of a high risk there is the 10-year risk of CVD death equal or above 5%. This translates roughly into a risk of any CVD event (nonfatal or fatal: the basis of the Framingham algorithm) of about 15%. SCORE also differentiates between high-and low-risk populations in Europe. Yet another algorithm has been developed by the World Health Organization (WHO).

Drugs used in the treatment of dyslipidemias

Clinical box The presence of xanthelasma does not necessarily indicate dyslipidemia

A 28-year-old woman developed unsightly yellow marks around both eyes (xanthelasma). She was asymptomatic and had a good exercise tolerance. Her cholesterol was 5.0 mmol/L (192 mg/dL), triglycerides 0.7 mmol/L (64 mg/dL), and her HDL-C was 1.4 mmol/L (53 mg/dL). There was no family history of early coronary disease.

Comment.

Xanthelasma may occur in individuals with completely normal lipid levels. On the other hand, lipid deposits in tendons (tendon xanthomas) are always diagnostic of a familial lipid disorder. The patient was reassured and referred for cosmetic surgery.

Clinical box Defining cardiovascular risk

The concept of cardiovascular risk is based on the probability of a cardiovascular event occurring within 10 years from the point of assessment. Different risk calculations use different endpoints: the Framingham Study uses the probability of any cardiovascular disease (CVD), whereas the European SCORE system focuses on the probability of a fatal event decision. Here are some examples of the criteria used to assess the risk as low, moderate or high:

The SCORE risk classification developed by the European Society of Cardiology/European Atherosclerosis Society (ESC/EAS) assesses the 10-year risk of fatal CVD:
very high risk is ≥10% 10-year probability of fatal CVD;
high risk is ≥5% but <10%;
and moderate risk is ≥1% but <5%.

To convert the risk of fatal CVD to risk of total (fatal + nonfatal) CVD, one needs to multiply the SCORE-calculated risk by 3 in men and 4 in women.

The risk stratification used in the US National Cholesterol Education Program Adult Treatment Panel III guidelines (ATP III) is based on data from the Framingham Study, and uses three levels of 10-year risk of fatal or nonfatal CHD: >20%, 10–20%, and <10%.

Note that diagnostic cutoff points and recommended treatment targets change, and are periodically updated by the relevant bodies. For instance, ATP IV guidelines are now under development. For current recommendations consult the websites of organizations listed at the end of the chapter.

Clinical box Cholesterol, triglycerides and glucose

concentrations: conventional and si units

These are approximate equivalent values for cholesterol:

mmol/L	mg/dL
4	150
5	190
6	230
7	270
8	310

The conversion factors to use for obtaining exact values are:
Cholesterol: to convert mmol/L to mg/dL, multiply by 38.6;
Triglycerides: to convert mmol/L to mg/dL, multiply by 88.5;
Glucose: to convert mmol/L to mg/dL, multiply by 18.

Management of dyslipidemias combines lifestyle measures and drug treatment

Effective cardiovascular prevention needs an approach that combines lifestyle modification (smoking cessation, diet and regular exercise) with drug treatment of dyslipidemia, hypertension and diabetes (Table 18.5). The concentration of plasma LDL (and total plasma cholesterol) can decrease by approximately 15% when a person consistently follows a low-cholesterol diet. When lifestyle measures fail to correct the abnormalities, one resorts to drug treatment. There are several classes of drugs that lower plasma cholesterol concentration.

Statins inhibit HMG-CoA reductase

Statins such as simvastatin, pravastatin, atorvastatin and rosuvastatin are competitive inhibitors of HMG-CoA reductase, the rate-limiting enzyme in cholesterol synthesis. They primarily lower plasma LDL. The inhibition of this enzyme results in a decrease in intracellular cholesterol concentration. This decrease, through SREBP transcription factors (Chapter 17, Fig. 17.7), increases the expression of LDL receptors on the cell membrane. This leads to increased cellular uptake of LDL and, consequently, to a lower plasma cholesterol concentration. The treatment with statins decreases plasma cholesterol concentration by 30–60%, and decreases the risk of future cardiovascular events

by 20–30%. Statins also seem to decrease the inflammatory phenomena in the arterial wall.

Fibrates act through PPARα transcription factor

Derivatives of fibric acid (fibrates) are agonists of the transcription factor PPARα. They stimulate the activity of LPL, decrease plasma triglyceride concentrations, and increase the concentration of HDL-C. Their effect on the LDL and total cholesterol is less pronounced than that of the statins.

Inhibitors of intestinal absorption bind bile acids and inhibit cholesterol transporter

Inhibitors of the intestinal absorption of cholesterol include older drugs, the bile acid-binding resins, that are now rarely used. They decrease plasma cholesterol concentration by interrupting the recirculation of cholesterol from the intestine and increasing its excretion. A newer drug, ezetimibe, inhibits the intestinal cholesterol transporter, the Niemann–Pick C1-like 1 (NPC1L1) protein in the intestinal brush border, and lowers total cholesterol by approximately 20%. Longer-term studies of its clinical benefit are in progress.

Omega-3 fatty acids lower plasma triglyceride concentration

A substantial decrease in plasma triglyceride concentration can be achieved by treatment with omega-3 fatty acids (fish oil). Interestingly, fish oil preparations are also anti-arrhythmic, particularly in patients who have already suffered myocardial infarction.

Role of antioxidants continues to be studied

In animals, antioxidants such as probucol or vitamin E inhibit development of atherosclerosis. Epidemiologic studies have shown that those who take antioxidants such as vitamin E and C or β-carotene have a decreased risk of cardiovascular disease. However, prospective clinical trials of antioxidant treatment failed to confirm such preventive benefit. One tentative explanation is that it is the natural antioxidants (such as those contained in fruits) or their combinations that are protective, rather than pure substances.

Summary

■ Lipoproteins transport hydrophobic lipids between organs and tissues.
■ Chylomicrons mediate the transport of dietary fat.
■ VLDL mediate the transport of endogenously synthesized fat.
■ Chylomicrons, VLDL and remnant lipoproteins are part of the organism's fuel distribution network: the fuel transport pathway.
■ LDL are cholesterol-rich lipoproteins which emerge from the fuel transport pathway. When present in excess they may enter the arterial wall.
■ HDL mediate reverse cholesterol transport, *e.g.* removal of cholesterol from the peripheral cells to the liver.
■ Atherogenesis involves endothelial dysfunction, lipid deposition, inflammatory reaction in the arterial wall, activation and proliferation of the arterial smooth muscle cells, and thrombosis. The major atherosclerosis-related diseases are coronary heart disease, stroke and peripheral vascular disease.
■ Interactions between cells participating in atherogenesis are mediated by an array of cytokines, growth factors and adhesion molecules.
■ Atherogenesis disrupts the structure of the arterial wall and results in the formation of atherosclerotic plaque, which narrows the lumen of the affected artery. However, the immediate cause of a heart attacks is not the slow growth of the plaque, but its sudden rupture.
■ Assessment of cardiovascular risk involves measurements of several lipid parameters and identification other risk factors such as hypertension, smoking and the presence of diabetes. The intensity of treatment is geared to the degree of calculated risk.

Active learning

1. Compare the composition of the VLDL and LDL.
2. What are the differences between the transport to the peripheral tissues of dietary triacylglycerols and triacylglycerols synthesized in the liver?
3. Describe the transport pathway for dietary fatty acids.
4. Give examples of interactions between different cell types in atherogenesis.
5. How does atherosclerotic plaque rupture?
6. In what way does endothelial dysfunction contribute to

atherosclerosis?

Further reading

Borissoff, JI, Spronk, HMH, ten Cate, H. The hemostatic system as a modulator of atherosclerosis. *N Engl J Med*. 2011; 364(18):1746–1760.

Dominiczak, MH. Risk factors for coronary disease: the time for a paradigm shift? *Clin Chem Lab Med*. 2001; 39:907–919.

Dominiczak, MH, Caslake, MJ. Apolipoproteins: metabolic role and clinical biochemistry applications. *Ann Clin Biochem*. 2011; 48:498–515.

Durrington, P. Dyslipidaemia. *Lancet*. 2003; 362:717–731.

Duval, C, Muller, M, Kersten, S. PPAR alpha and dyslipidemia (review). *Biochim Biophys Acta Mol Biol Cell Biol Lipids*. 2007; 1771:961–971.

Libby, P, Aikawa, M. Stabilization of atherosclerotic plaques: New mechanisms and clinical targets. *Nature Medicine*. 2002; 8:1257–1262.

Michalik, L, Auwerx, J, Berger, JP, *et al.* International Union of Pharmacology. LXI. Peroxisome proliferator-activated receptors. *Pharmacol Rev*. 2006; 58:726–741.

O'Donnell, CJ, Nabel, EG. Genomics of cardiovascular disease. In: Genomic medicine. Feero WG, Guttmacher, AE, editors. *N Engl J Med*. 2011; 365:2098–2109.

Packard, RRS, Libby, P. Inflammation in atherosclerosis: from vascular biology to biomarker discovery and risk prediction. *Clin Chem*. 2008; 54:24–38.

Rader, DJ, Daugherty, A. Translating molecular discoveries into new therapies for atherosclerosis. *Nature*. 2008; 451:904–913.

Websites and downloads

Familial hypercholesterolemia. Simon Broome Diagnostic criteria for index individuals and relatives. http://www.nice.org.uk/nicemedia/live/12048/41707/41707.pdf. [(Accessed October 2012)].

General Cardiovascular disease. Framingham Heart Study. www.framinghamheartstudy.org/risk/gencardio.html. [(Accessed May 2013)].

The Task Force for the Management of Dyslipidaemias of the European Society of Cardiology (ESC) and the European Atherosclerosis Society (EAS). ESC/EAS Guidelines for the management of dyslipidaemias. www.escardio.org/guidelines-surveys/esc-guidelines/guidelinesdocuments/guidelines-dyslipidemias-ft.pdf. [(Accessed May 2013)].

Marks, D, Thorogood, M, Neil, HA, *et al.* A review on the diagnosis, natural history, and treatment of familial hypercholesterolaemia. *Atherosclerosis*. 2003; 168:1–14.

Morris, PB. ATP IV, CVD risk assessment and dyslipidemia. www.acli.com/Events/Documents/Tue22812%20-%20Lipidology%20-%20Pamela%20Morris.pdf. [(Accessed May 2013)].

National Cholesterol Education Program Adult Treatment Panel III (ATP III.

www.nhlbi.nih.gov/guidelines/cholesterol/index.htm [(Accessed May 2013)].

National High Blood Pressure Education Program. Joint National Committee on the Prevention, Detection, Evaluation, and Treatment of High Blood Pressure (7th Report: known as JNC7) JNC8 is in development www.nhlbi.nih.gov/guidelines/hypertension/. [(Accessed May 2013)].

National Institute for Health and Clinical Excellence (NICE), United Kingdom. http://publications.nice.org.uk/lipid-modification-cg67 [(Accessed May 2013)].

Von Eckardstein, A. HDL in motion: molecular dynamics and disease pathophysiology. www.medscape.org/noscan/slideshow. [(Accessed May 2013)].

CHAPTER 19

Biosynthesis and Degradation of Amino Acids

Allen B. Rawitch

Learning objectives

After reading this chapter you should be able to:

- Describe the three mechanisms used by humans for removal of the nitrogen from amino acids prior to the metabolism of their carbon skeletons.
- Outline the sequence of reactions in the urea cycle and trace the flow of nitrogen from amino acids into and out of the cycle.
- Describe the role of vitamin B_6 in aminotransferase reactions.
- Define the terms and give examples of glucogenic and ketogenic amino acids.
- Summarize the factors that contribute to the input and the depletion of the pool of free amino acids in animals.
- Summarize the sources and use of ammonia in animals and explain the concept of nitrogen balance.
- Identify the essential amino acids and the metabolic sources of the nonessential amino acids.
- Explain the biochemical basis and the therapeutic rationale for treatment of phenylketonuria and maple syrup urine disease.

Introduction

Amino acids are a source of energy from the diet and during fasting

In addition to their roles as building blocks for peptides and proteins, and as precursors of neurotransmitters and hormones, the carbon skeletons of some amino acids can be used to produce glucose through gluconeogenesis, thereby providing a metabolic fuel for tissues that require or prefer glucose; such amino acids are designated as glucogenic or glycogenic amino acids. The carbon skeletons of some amino acids can also produce the equivalent of acetyl-CoA or acetoacetate and are termed ketogenic, indicating that they can be metabolized to give immediate precursors of lipids or ketone bodies. In an individual consuming adequate amounts of protein, a significant quantity of amino acids may also be converted to carbohydrate (glycogen) or fat (triacylglycerol) for storage. Unlike carbohydrates and lipids, amino acids do not have a dedicated storage form equivalent to glycogen or fat.

When amino acids are metabolized, the resulting excess nitrogen must be excreted. Since the primary form in which the nitrogen is removed from amino acids is ammonia, and because free ammonia is quite toxic, humans and most higher animals rapidly convert the ammonia derived from amino acid catabolism to urea, which is neutral, less toxic, very soluble, and excreted in the urine. Thus **the primary nitrogenous excretion product in humans is urea, produced by the urea cycle in liver.** Animals that excrete urea are termed ureotelic. In an average individual, more than 80% of the excreted nitrogen is in the form of urea (25–30 g/24 h). Smaller amounts of nitrogen are also excreted in the form of uric acid, creatinine, and ammonium ion.

The carbon skeletons of many amino acids may be derived from metabolites in central pathways, allowing the biosynthesis of some, but not all, the amino acids in humans. Amino acids that can be synthesized in this way are therefore not required in the diet (nonessential amino acids), whereas amino acids having carbon skeletons that cannot be derived from normal human metabolism must be supplied in the diet (essential amino acids). For the biosynthesis of nonessential amino acids, amino groups must be added to the appropriate carbon skeletons. This generally occurs through the transamination of an α-keto acid corresponding to that specific amino acid.

Metabolism of dietary and endogenous proteins

Relationship to central metabolism

Muscle protein and adipose lipids are consumed to support gluconeogenesis during fasting and starvation

Although body proteins represent a significant proportion of potential energy reserves (Table 19.1), under normal circumstances they are not used for energy production. In an extended fast, however, muscle protein is degraded to amino acids for the synthesis of essential proteins, and to keto acids for gluconeogenesis to maintain blood glucose concentration as well to provide carbons for energy production. This accounts for the loss of muscle mass during fasting.

Table 19.1

Storage forms of energy in the body

Stored fuel	Tissue	Amount (g)*	Energy (kJ)	(kcal)
Glycogen	Liver	70	1176	280
Glycogen	Muscle	120	2016	480
Free glucose	Body fluids	20	336	80
Triacylglycerol	Adipose	15,000	567,000	135,000
Protein	Muscle	6000	100,800	24,000

Proteins represent a substantial energy reserve in the body.

*In a 70-kg individual.

(Adapted with permission from Cahill GF Jr, *Clin Endocrinol Metab* 1976;**5**:398.)

In addition to its role as an important source of carbon skeletons for oxidative

metabolism and energy production, dietary protein must provide adequate amounts of those amino acids that we cannot make, to support normal protein synthesis. The relationships of body protein and dietary protein to central amino acid pools and to central metabolism are illustrated in Figure 19.1.

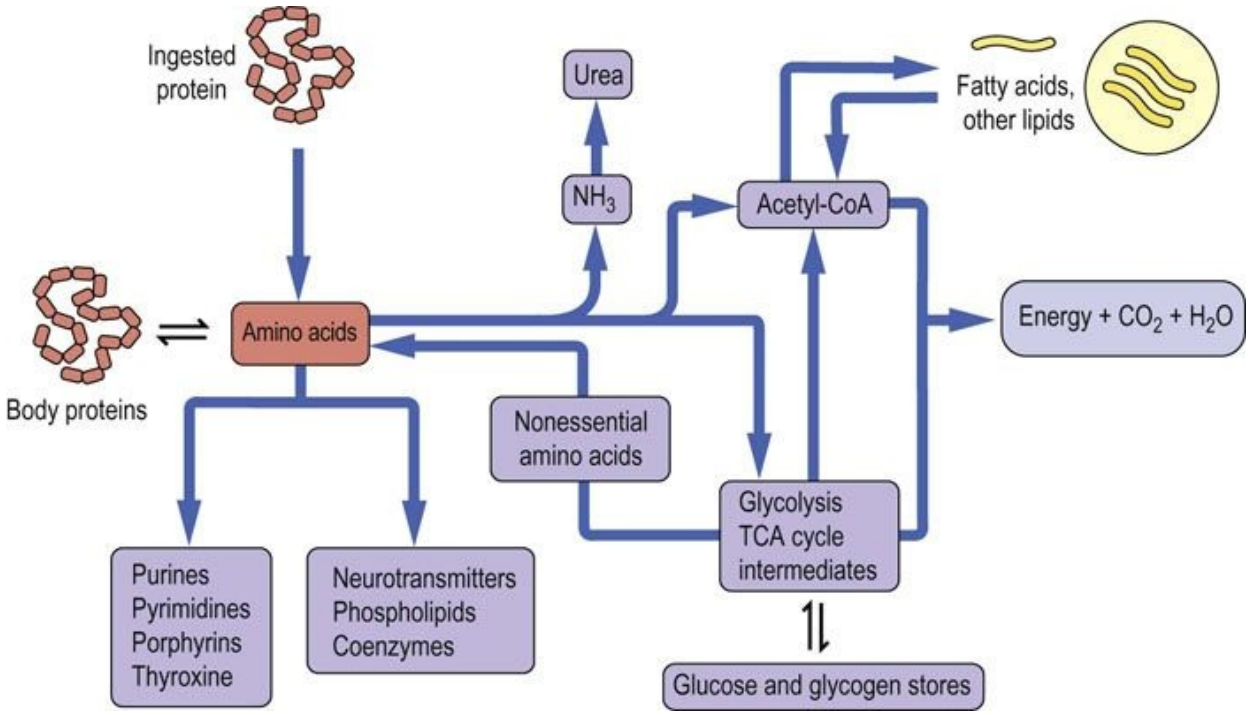


FIG. 19.1 Metabolic relationships among amino acids.
The pool of free amino acids is derived from the degradation and turnover of body proteins and from the diet. The amino acids are precursors of important biomolecules, including hormones, neurotransmitters and proteins, and also serve as a carbon source for central metabolism, including gluconeogenesis, lipogenesis and energy production.

Advanced concept box Alanine and interorgan carbon and nitrogen flow

Much of the carbon flow that occurs between peripheral tissues, such as skeletal muscle, and the liver is facilitated by the release of alanine into the blood by the peripheral tissues. The alanine is converted to pyruvate in the liver and the nitrogen component is incorporated into urea. The pyruvate can be used for

gluconeogenesis to produce glucose, which is released into the blood for transport back to peripheral tissues. This **glucose–alanine cycle** allows the net conversion of amino acid carbons to glucose, the elimination of amino acid nitrogen as urea, and the return of carbons to the peripheral tissues in the form of glucose.

The glucose-alanine cycle works in a fashion similar to the Cori cycle (Chapter 21) in which lactate, released from skeletal muscle, is used for hepatic gluconeogenesis, the key difference being that alanine also carries a nitrogen atom to the liver. Alanine and glutamine are released in approximately equal quantities from skeletal muscle and represent almost 50% of the amino acids released by skeletal muscle into the blood – an amount that far exceeds the proportion of these amino acids in muscle proteins. Thus, there is substantial remodeling of protein-derived amino acids by transamination reactions, prior to their release from muscle.

Digestion and absorption of dietary protein

In order for dietary protein to contribute to either energy metabolism or pools of essential amino acids, the protein must be digested to the level of free amino acids or small peptides and absorbed across the gut. Digestion of protein begins in the stomach with the action of pepsin, a carboxyl protease that is active in the very low pH found in the gastric environment. Digestion continues as the stomach contents are emptied into the small intestine and mixed with pancreatic secretions. These pancreatic secretions are alkaline and contain the inactive precursors of several serine proteases including trypsin, chymotrypsin and elastase along with carboxypeptidases. The process is completed by enzymes in the small intestine (see Chapter 10). After any remaining di-and tripeptides are broken down in enterocytes, the free amino acids are transported to the portal vein and carried to the liver for energy metabolism or biosynthesis, or distributed to other tissues to meet similar needs.

Turnover of endogenous proteins

In addition to the ingestion, digestion and absorption of amino acids from dietary protein, all the proteins in the body have a half-life or life span and are routinely degraded to amino acids and replaced with newly synthesized protein. This process of protein turnover is carried out in the lysosome or by proteasomes. In the case of lysosomal digestion, protein turnover begins with engulfment of the protein or organelle in vesicles known as autophagosomes, by a process known as autophagy. The vesicles then fuse with lysosomes and the protein, lipid and glycans are degraded by lysosomal acid hydrolases. Cytosolic proteins are degraded primarily by proteasomes which are high-molecular-weight complexes containing multiple proteolytic activities. There are both ubiquitin-dependent (Chapter 34) and ubiquitin-independent pathways for degradation of cytosolic proteins.

Amino acid degradation

Amino acids destined for energy metabolism must be deaminated to yield the carbon skeleton

There are three mechanisms for removal of the amino group from amino acids.
■ **Transamination** – the transfer of the amino group to a suitable keto acid acceptor (Fig. 19.2).

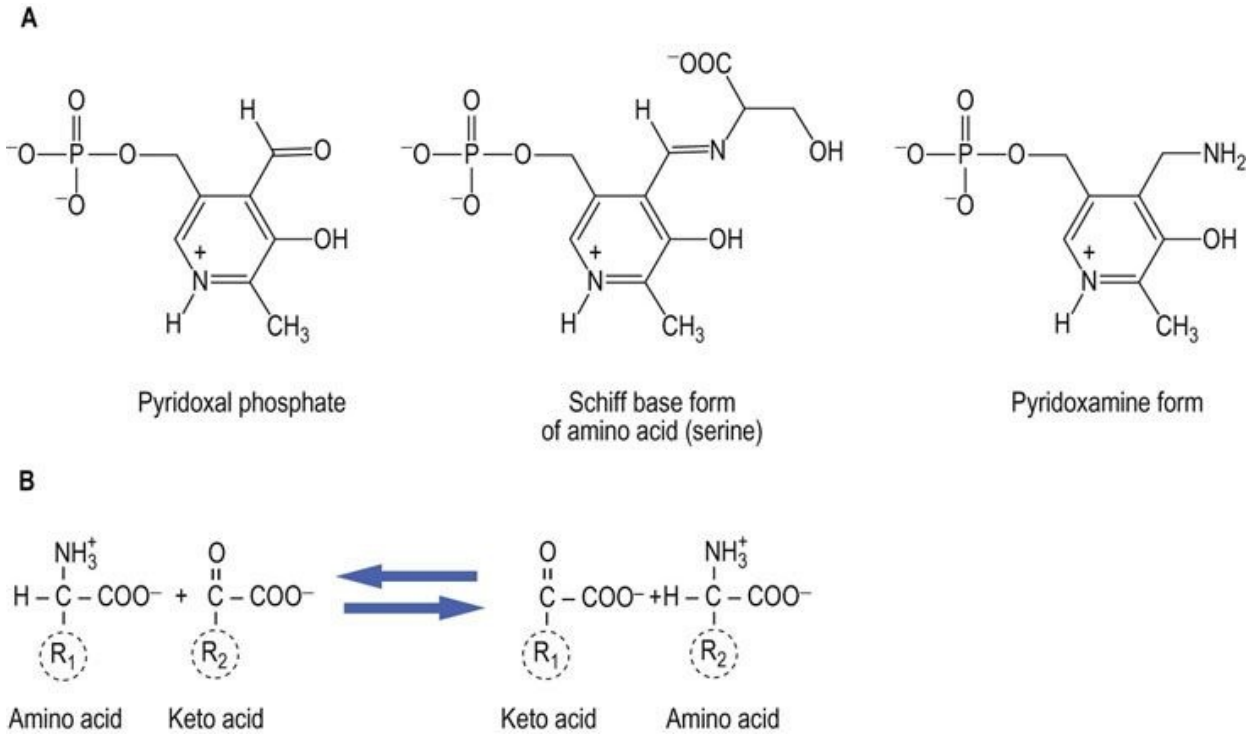


FIG. 19.2 The catalytic role of pyridoxal phosphate.
Aminotransferases or transaminases use pyridoxal phosphate as a cofactor. A pyridoxamine adduct acts as an intermediate in transfer of an amino group between an α-amino acid and an α-keto acid. **(A)** Structures of the components involved. The cofactor, pyridoxal phosphate, is used in a variety of enzyme-catalyzed reactions involving both amino and keto compounds, including transamination and decarboxylation reactions. **(B)** Transamination involves both a donor α-amino acid (R_1), and an acceptor α-keto acid (R_2). The products are an α-keto acid derived from the carbon skeleton of R_1 and an α-amino acid from the carbon skeleton of R_2.

■ **Oxidative deamination** – the oxidative removal of the amino group, resulting in keto acids and ammonia (see Fig. 19.3).

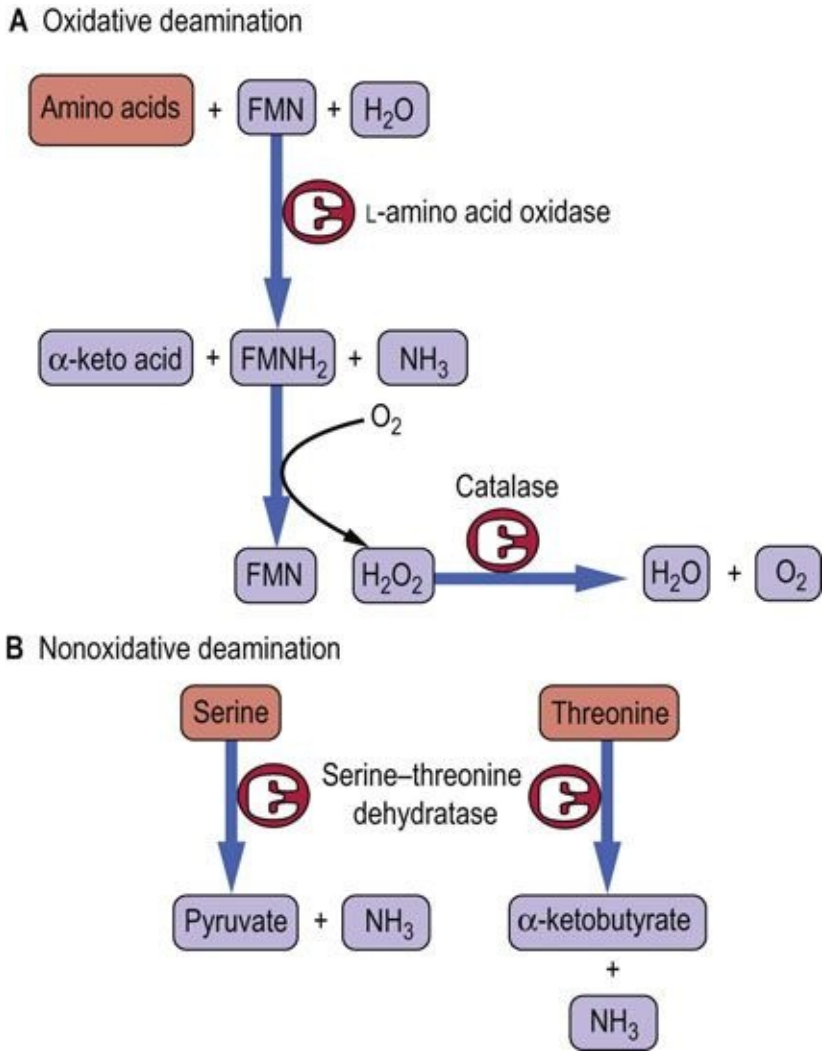


FIG. 19.3 Deamination of amino acids.
The primary route for amino group removal is via transamination, but there are additional enzymes capable of removing the α-amino group. **(A)** L-amino acid oxidase produces ammonia and an α-keto acid directly, using flavin mononucleotide (FMN) as a cofactor. The reduced form of the flavin must be regenerated using molecular oxygen; this reaction is one of several that produce H_2O_2. The peroxide is decomposed by catalase. **(B)** A second means of deamination is possible only for hydroxyamino acids (serine and threonine), through a dehydratase mechanism; the Schiff base, imine intermediate hydrolyzes to form the keto acid and ammonia.

■ **Removal of a molecule of water by a dehydratase** – *e.g.* serine or threonine

dehydratase; this reaction produces an unstable, imine intermediate that hydrolyzes spontaneously to yield an α-keto acid and ammonia (see Fig. 19.3).

Clinical test box Measurement of blood urea nitrogen

Serum urea measurements (also reported by laboratories as BUN or blood urea nitrogen) are critical in monitoring patients with a variety of metabolic diseases in which the metabolism of amino acids may be affected and in tracking the condition of individuals with renal problems. The traditional methodology used for measuring blood urea has relied on the action of the enzyme urease, which converts urea to CO_2 and ammonia. The resulting ammonia can be detected spectrophotometrically by formation of a colored compound on reaction with phenol or a related compound (the Berthelot reaction).

Clinical box Reaction to monosodium glutamate

A healthy 30-year-old woman experienced the sudden onset of headache, sweating, and nausea after eating at an Oriental restaurant. She felt weak and experienced some tingling and a sensation of warmth in her face and upper torso. The symptoms passed after about 30 minutes and she experienced no further problems. Upon visiting her doctor the next day, she learned that some individuals react to foods containing high levels of the food additive monosodium glutamate, the sodium salt of glutamic acid. Monosodium glutamate is a common food additive used to enhance savory flavor in many foods. It is one of the principal substances responsible for the umami or savory taste sensation which enhances the flavor effects of the other basic taste and taste combination sensations.

Comment.

The flu-like symptoms that develop, previously described as 'Chinese restaurant syndrome', have been attributed to central nervous system (CNS) effects of glutamate or its derivative, the inhibitory neurotransmitter γ-amino butyric acid (GABA). Interestingly, studies have shown that this phenomenon causes no permanent CNS damage and that, although bronchospasm may be triggered in individuals with severe asthma, the symptoms are generally brief and completely reversible. Monosodium glutamate continues to be a widely used additive in many processed foods and is approved by the FDA.

The principal mechanism for removal of amino groups from the common amino acids is via transamination, or the transfer of amino groups from the amino acid to a suitable α-keto acid acceptor, most commonly to α-ketoglutarate or oxaloacetate, forming glutamate and aspartate, respectively. Several enzymes, called **aminotransferases (or transaminases)**, are capable of removing the amino group from most amino acids and producing the corresponding α-keto acid. Aminotransferase enzymes use pyridoxal phosphate, a cofactor derived from the **vitamin B_6 (pyridoxine)**, as a key component in their catalytic mechanism; pyridoxamine is an intermediate in the reaction. The structures of the various forms of vitamin B_6 and the net reaction catalyzed by aminotransferases are shown in Figure 19.2.

Nitrogen atoms are incorporated into urea from two sources, glutamate and aspartate

The transfer of an amino group from one keto acid carbon skeleton to another may seem to be unproductive and not useful in itself; however, when one considers the nature of the primary keto acid acceptors that participate in these reactions (α-ketoglutarate and oxaloacetate) and their products (glutamate and aspartate), the logic of this metabolism becomes clear. The two nitrogen atoms in urea are derived exclusively from these two amino acids (Fig. 19.4), thereby linking amino acid catabolism to energy metabolism. Ammonia, which is

produced primarily from glutamate via the glutamate dehydrogenase (GDH) reaction (Fig. 19.5B), enters the urea cycle as **carbamoyl phosphate**. The second nitrogen is contributed to urea by aspartic acid. Fumarate is formed in this process and may be recycled through the tricarboxylic acid (TCA) cycle to oxaloacetate, which can accept another amino group and re-enter the urea cycle, or the fumarate may be used for energy metabolism or gluconeogenesis. Thus the funneling of amino groups from other amino acids into glutamate and aspartate provides the nitrogen for urea synthesis in a form appropriate for the urea cycle (Fig. 19.4). The other pathways that lead to the release of amino groups from some amino acids through the action of amino acid oxidase or dehydratases (Fig. 19.3) make relatively minor contributions to the flow of amino groups from amino acids to urea.

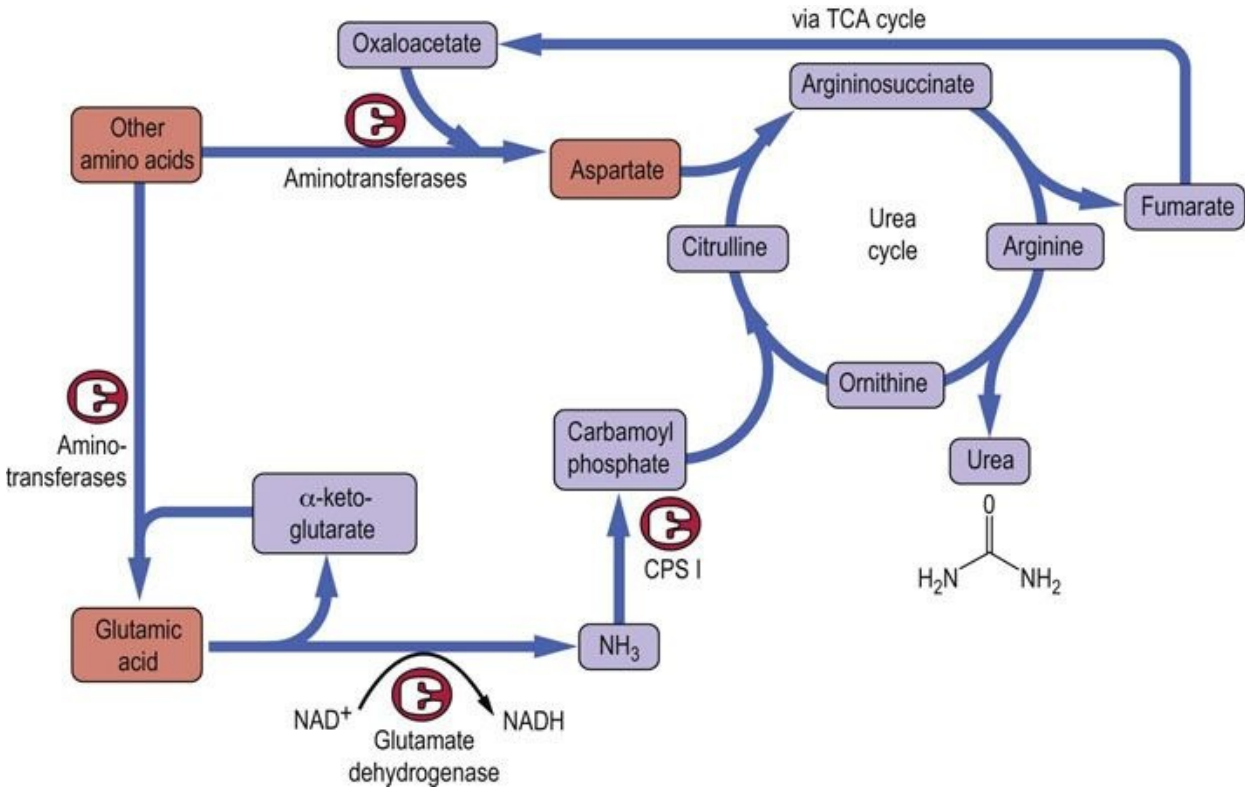


FIG. 19.4 Sources of nitrogen atoms in the urea cycle.
Nitrogen enters the urea cycle from most amino acids via transfer of the α-amino group to either oxaloacetate or α-ketoglutarate, to form aspartate or glutamate, respectively. Glutamate releases ammonia in the liver through the action of GDH (Fig. 19.5). The ammonia is incorporated into carbamoyl phosphate, and the aspartate combines with citrulline to provide the second nitrogen for urea synthesis. Oxaloacetate and α-ketoglutarate can be repeatedly recycled to channel nitrogen into this pathway. CPS I,

carbamoyl phosphate synthetase-I.

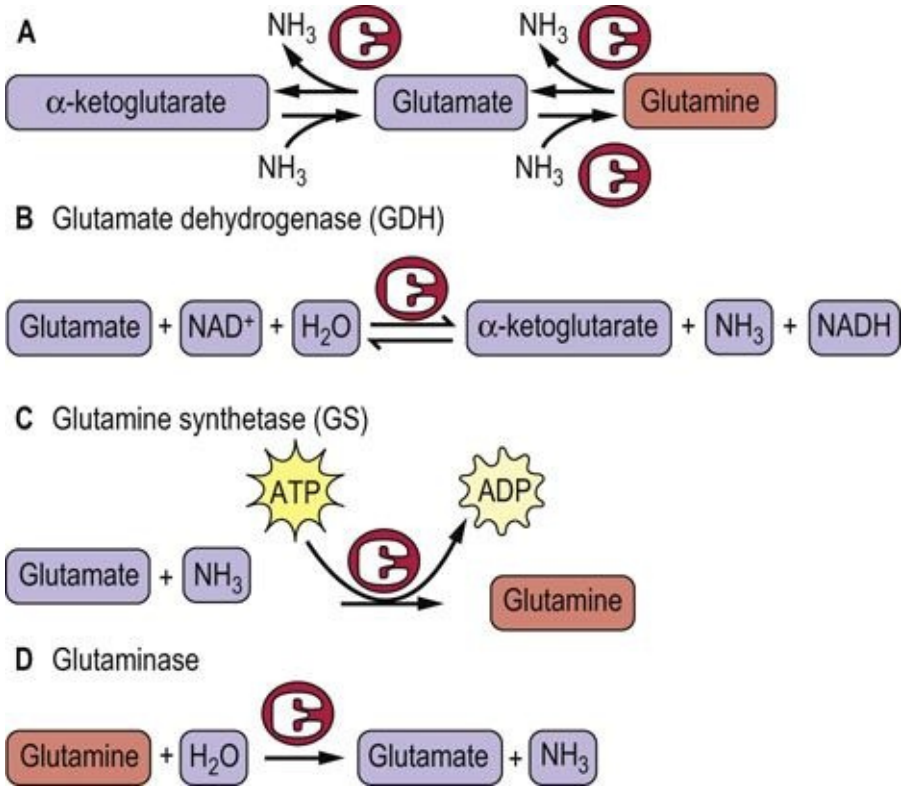


FIG. 19.5 Relationships between glutamate, glutamine and α-ketoglutarate. The several forms of the carbon skeleton of glutamic acid have key roles in the metabolism of amino groups. **(A)** Three forms of the same carbon skeleton. **(B)** The glutamine dehydrogenase reaction is a reversible reaction that can produce glutamate from α-ketoglutarate or convert glutamate to α-ketoglutarate and ammonia. The latter reaction is important in the synthesis of urea because amino groups are fed to α-ketoglutarate via transamination from other amino acids. **(C)** Glutamine synthetase catalyzes an energy-requiring reaction with a key role in transport of amino groups from one tissue to another; it also provides a buffer against high concentrations of free ammonia in tissues. **(D)** The second half of the glutamine transport system for nitrogen is the enzyme glutaminase, which hydrolyzes glutamine to glutamate and ammonia. This reaction is important in the kidney for management of proton transport and pH control.

The central role of glutamine

Ammonia is detoxified by incorporation into glutamine, then eventually into urea

In addition to the role of glutamate as a carrier of amino groups to GDH, glutamate serves as a precursor of glutamine, a process that consumes a molecule of ammonia. This is important because glutamine, along with alanine, is a key transporter of amino groups between various tissues and the liver, and is present in greater concentrations than most other amino acids in blood. The three forms of the same carbon skeleton, α-ketoglutarate, glutamate, and glutamine, are interconverted via aminotransferases, glutamine synthetase, glutaminase, and GDH (see Fig. 19.5). Thus glutamine can serve as a buffer for ammonia utilization, as a source of ammonia, and as a carrier of amino groups. Because ammonia is quite toxic, a balance must be maintained between its production and utilization. A summary of the sources and pathways that use or produce ammonia is shown in Figure 19.6. It should be noted that the GDH reaction is reversible under physiologic conditions if amino groups are required for amino acid and other biosynthetic processes.

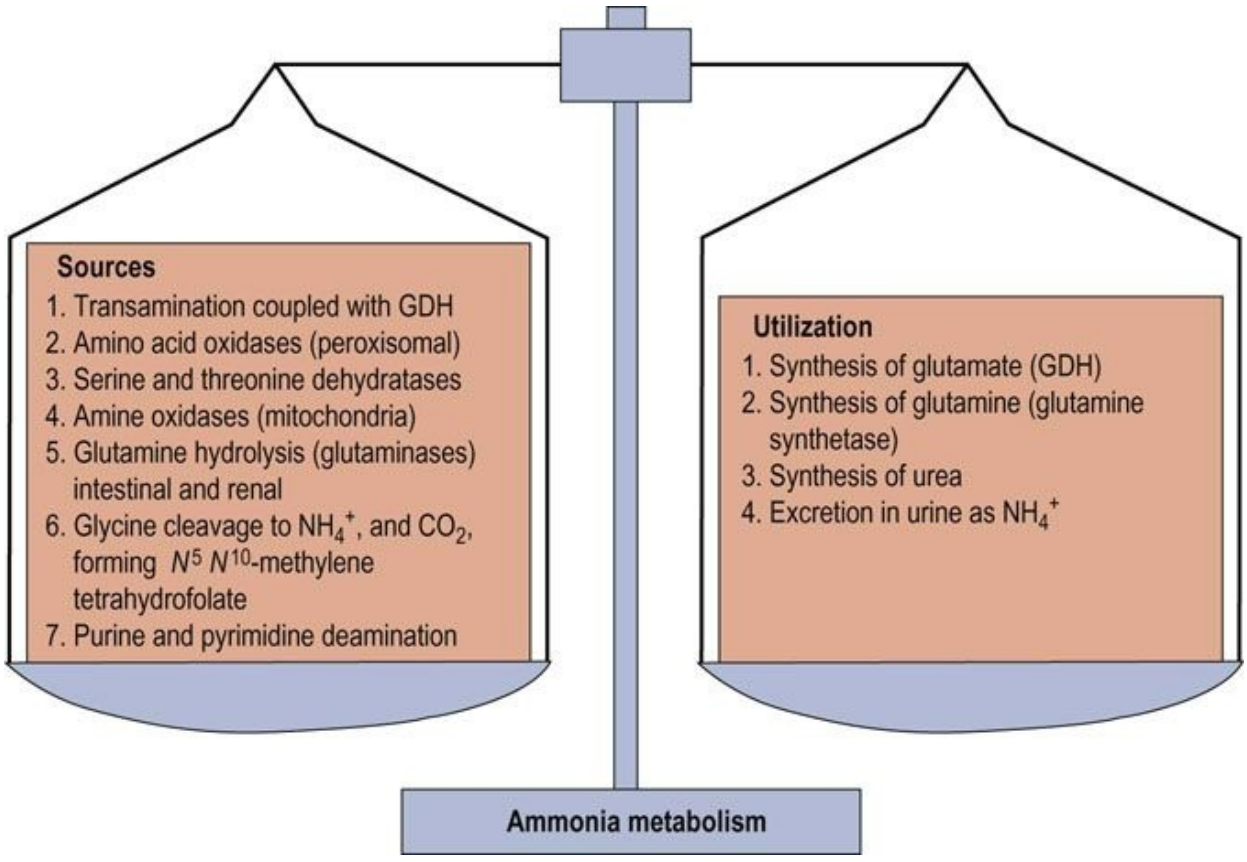


FIG. 19.6 Balance in ammonia metabolism.
The balance between production and utilization of free ammonia is critical for

maintenance of health. This figure summarizes the sources and pathways that use ammonia. Although most of these reactions occur in many tissues, synthesis of urea and the urea cycle are restricted to the liver. Glutamine and alanine function as the primary transporters of nitrogen from peripheral tissues to the liver.

The urea cycle and its relationship to central metabolism

The urea cycle is an hepatic pathway for disposal of excess nitrogen

Urea is the principal nitrogenous excretion product in humans (Table 19.2). The urea cycle (see Fig. 19.4) was the first metabolic cycle to be well defined; its description preceded that of the TCA cycle. The start of the urea cycle may be considered the synthesis of carbamoyl phosphate from an ammonium ion, derived primarily from glutamate via GDH (Fig. 19.5), and bicarbonate from liver mitochondria. This reaction requires two molecules of ATP and is catalyzed by the enzyme **carbamoyl phosphate synthetase I (CPS I)** (Fig. 19.7), which is found at high concentration in the mitochondrial matrix.

Table 19.2

Urinary nitrogen excretion

Urinary metabolite	g excreted/24 h*	% of total
Urea	30	86
Ammonium ion	0.7	2.8
Creatinine	1.0–1.8	4–5
Uric acid	0.5–1.0	2–3

*Approximate values in an average adult male.

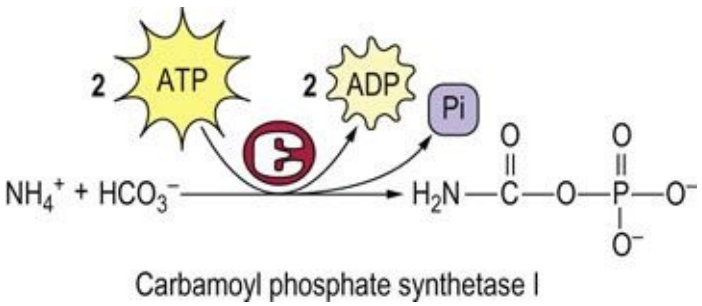


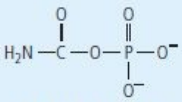
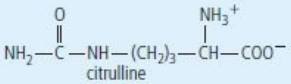
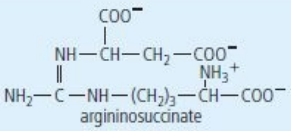
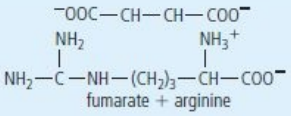
FIG. 19.7 Synthesis of carbamoyl phosphate.
The first nitrogen, derived from ammonia, enters the urea cycle as carbamoyl phosphate, synthesized by carbamoyl phosphate synthetase I in the liver.

The mitochondrial isozyme, CPS I, is unusual in that it requires *N*-acetylglutamate as a cofactor. It is one of two carbamoyl phosphate synthetase enzymes that have key roles in metabolism. The second, CPS II, is found in the cytosol, does not require *N*-acetylglutamate, and is involved in pyrimidine biosynthesis (Chapter 31).

Ornithine transcarbamoylase catalyzes the condensation of carbamoyl phosphate with the amino acid **ornithine**, to form **citrulline**: see Figure 19.4 for pathway and Table 19.3 for structures. In turn, the citrulline is condensed with aspartate to form argininosuccinate. This step is catalyzed by argininosuccinate synthetase and requires ATP; the reaction cleaves the ATP to adenosine monophosphate (AMP) and inorganic pyrophosphate (PPi) (2 ATP equivalents). The formation of argininosuccinate incorporates the second nitrogen atom destined for urea. Argininosuccinate is cleaved by argininosuccinase to arginine and fumarate, and the arginine is then cleaved by **arginase** to yield urea and ornithine. The ornithine and fumarate can re-enter the urea cycle, while the urea diffuses into the blood, is transported to the kidney and excreted in urine. The net process of ureogenesis is summarized in Table 19.4.

Table 19.3

Enzymes of the urea cycle

Enzyme	Reaction catalyzed	Remarks	Reaction product
Carbamoyl phosphate synthetase	Formation of carbamoyl phosphate from ammonia and CO_2	Fixes ammonia released from amino acids, uses 2 ATP, located in the **mitochondrion,** deficiency leads to high blood concentrations of ammonia and related toxicity	 carbamoyl phosphate
Ornithine transcarbamoylase	Formation of citrulline from ornithine and carbamoyl phosphate	Releases Pi, an example of a transferase, located in the **mitochondrion,** deficiency leads to high blood concentrations of ammonia and orotic acid, as carbamoyl phosphate is shunted to pyrimidine biosynthesis	 citrulline
Argininosuccinate synthetase	Formation of arginino-succinate from citrulline and aspartate	Requires ATP, which is cleaved to AMP + PPi – an example of a ligase, located in the **cytosol,** deficiency leads to high blood concentrations of ammonia and citrulline	 argininosuccinate
Argininosuccinase	Cleavage of argininosuccinate to arginine and fumarate	An example of a lyase, located in **cytosol,** deficiency leads to high blood concentrations of ammonia and citrulline	 fumarate + arginine
Arginase	Cleavage of arginine to ornithine and urea	An example of a hydrolase, located in the **cytosol** and primarily in the liver, deficiency leads to moderately increased blood ammonia and high blood concentrations of arginine; urea is excreted and ornithine re-enters the urea cycle	 urea + ornithine

Five enzymes catalyze the urea cycle in liver. The first enzyme, CPS I, which fixes NH_4^+ as carbamoyl phosphate, is the regulatory enzyme and is sensitive to the allosteric effector, *N*-acetylglutamate.

Table 19.4

Urea synthesis

Component reactions in urea synthesis

$CO_2 + NH_3 + 2$ ATP	→	Carbamoyl phosphate + 2 ADP + Pi
Carbamoyl phosphate + ornithine	→	Citrulline + Pi
Citrulline + aspartate + ATP	→	Argininosuccinate + AMP + PPi
Argininosuccinate	→	Arginine + fumarate
Arginine	→	Urea + ornithine
$CO_2 + NH_3 + 3$ ATP + aspartate	→	Urea + 2 ADP + AMP + 2 Pi + PPi + fumarate

The urea cycle is split between the mitochondrial matrix and the cytosol

The first two steps in the urea cycle occur in the mitochondrion. The citrulline which is formed in the mitochondrion then moves into the cytosol by a specific passive transport system. The cycle is completed in the cytosol with the release of urea from arginine and the regeneration of ornithine. Ornithine is transported back across the mitochondrial membrane to continue the cycle. Carbons from fumarate, released in the argininosuccinase step, may also re-enter the mitochondrion and be recycled by enzymes in the TCA cycle to oxaloacetate and ultimately to aspartate (Fig. 19.8), thus completing the second part of the urea cycle. Urea synthesis occurs virtually exclusively in the liver and the role of the enzyme, arginase, in other tissues is probably related more closely to ornithine requirements for polyamine synthesis than to the production of urea.

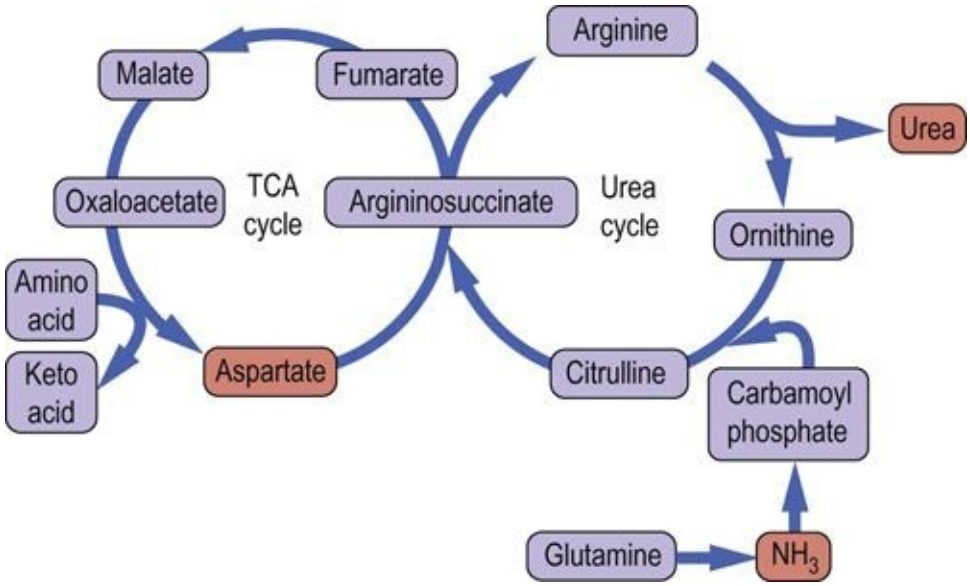


FIG. 19.8 The tricarboxylic acid and urea cycles.
Analysis of the urea cycle reveals that it is really two cycles, with the carbon flow split between the primary urea synthetic process and the recycling of fumarate to aspartate; the latter cycle occurs in the mitochondrion and involves parts of the TCA cycle.

Advanced concept box Carbamoyl phosphate synthesis

The enzyme carbamoyl phosphate synthetase I (CPS I) is found in the mitochondrion and primarily in the liver; a second enzyme, CPS II, is found in the cytosol and in virtually all tissues. Although the product of both these enzymes is the same, namely carbamoyl phosphate, the enzymes are derived from different genes and function in ureogenesis (CPS I) or pyrimidine biosynthesis (CPS II), respectively. Additional differences between the two enzymes include their source of nitrogen (NH_3 for CPS I, and glutamine for CPS II) and their requirement for *N*-acetylglutamate (required by CPS I but not by CPS II). Under normal circumstances, CPS I and II function independently and in different cellular compartments; however, when the urea cycle is blocked, *e.g.* as a result of a deficiency in ornithine transcarbamoylase, the accumulated mitochondrial carbamoyl phosphate spills over into the cytosolic compartment and may stimulate excess pyrimidine synthesis, resulting in a build-up of orotic acid in the blood and urine.

Advanced concept box Ammonia encephalopathy

The mechanisms involved in ammonia toxicity – the encephalopathy in particular – are not well defined. It is clear, however, that when its concentration builds up in the blood and other biological fluids, ammonia diffuses into cells and across the blood–brain barrier. The increase in ammonia causes an increased synthesis of glutamate from α-ketoglutarate and increased synthesis of glutamine. Although this is a normal detoxifying reaction in cells, when concentrations of ammonia are significantly increased, supplies of α-ketoglutarate in cells of the CNS may be depleted, resulting in inhibition of the TCA cycle and a decrease in ATP production. There may be additional mechanisms accounting for the bizarre behavior observed in individuals with high blood concentrations of ammonia. Either glutamate, a major inhibitory neurotransmitter, or its derivative, γ-amino butyric acid (GABA), may also contribute to the CNS effects.

Clinical box Hereditary hyperammonemia

An apparently healthy 5-month-old female infant was brought to a pediatrician's office by her mother, with a complaint of periodic bouts of vomiting and a failure to gain weight. The mother also reported that the child would oscillate between periods of irritability and lethargy. Subsequent examination and laboratory results revealed an abnormal electroencephalogram, a markedly increased concentration of plasma ammonia (323 mmol/L, 550 mg/dL; the normal range is 15–88 mmol/L, 25–150 mg/dL), and greater than normal concentrations of glutamine, but low concentrations of citrulline. Orotate, a pyrimidine nucleotide precursor, was found in her urine.

Comment.

The infant was admitted to hospital and treated with intravenous phenylacetate and benzoate along with arginine. The benzoate and phenyllactate are metabolized to glycine and glutamate conjugates, which are excreted, with their nitrogen content, into urine; arginine stimulates residual urea cycle activity. The infant improved rapidly and was discharged from hospital on a low-protein diet with arginine supplementation. Subsequent biopsy of the patient's liver indicated that her hepatic ornithine transcarbamoylase activity was about 10% of normal.

Clinical box Parkinson's disease

An otherwise healthy, 60-year-old man noticed an occasional tremor in his left arm when relaxing and watching television. He also noticed occasional muscle cramping in his left leg, and his spouse noticed that he would occasionally develop a trance-like stare. A complete physical examination and consultation with a neurologist confirmed a diagnosis of Parkinson's disease. He was prescribed a medication that contained L-dihydroxyphenylalanine (L-DOPA) and a monoamine oxidase inhibitor (MAOI). L-DOPA is a precursor of the neurotransmitter dopamine, while monoamine oxidase is the enzyme responsible for the oxidative deamination and degradation of dopamine. His symptoms improved immediately, but he gradually experienced significant side effects from the medication, especially the occurrence of involuntary movements.

Comment.

Parkinson's disease is caused by the death of dopamine-producing cells in the substantia nigra and the locus ceruleus. Although medication can markedly reduce the symptoms, the disease is progressive and may result in severe disability. Dopaminergic agonists often have side effects and also have limited effect on

tremor, so that other treatments such as deep brain stimulation or ablation are used in selected cases. Monoamine oxidase is also involved in deamination of other amines in the brain, so that MAOIs have many undesirable side effects. Transplantation of dopaminergic fetal tissue into the brain has been attempted but is a controversial experimental treatment at present.

Regulation of the urea cycle

N-acetylglutamate, and indirectly arginine, is an essential allosteric regulator of the urea cycle

The urea cycle is regulated in part by control of the concentration of ***N*-acetylglutamate**, the essential allosteric activator of CPS I. Arginine is an allosteric activator of *N*-acetylglutamate synthase and also a source of ornithine (via arginase) for the urea cycle. Concentrations of urea cycle enzymes also increase or decrease in response to a high-or low-protein diet, and urea synthesis and excretion are decreased and NH_4^+ excretion is increased during acidosis as a mechanism to excrete protons into the urine. Lastly, it should be noted that during a fast, protein is broken down to free amino acids which are used for gluconeogenesis. The increase in protein degradation during fasting results in increased urea synthesis and excretion, a mechanism to dispose of the released nitrogen.

Clinical test box Screening for amino acid metabolic defects in the newborn

In most developed countries today, a spot of the blood of newborn infants is routinely collected on filter paper and tested for a series of compounds which are markers of inherited metabolic disease. The number of markers tested for may vary from state to state in the USA, but generally ranges from 10 to 30. Because of the need for rapid screening, small sample size and reduced cost, older methodology is rapidly being replaced by technology (see Chapter

36) which uses gas or liquid chromatography–mass spectrometry to measure the level of multiple markers simultaneously. The speed and high throughput capacity of this **metabolomics** technology allow rapid screening of 20 or more markers from dried blood spots and the identification of infants who are potential victims of these inborn errors of metabolism. This technology is also applied to analysis of urine samples.

Defects in any of the enzymes of the urea cycle have serious consequences. Infants born with defects in any of the first four enzymes in this pathway may appear normal at birth, but rapidly become lethargic and hypothermic, and may have difficulty breathing. Blood concentrations of ammonia increase quickly, followed by cerebral edema. The symptoms are most severe when early steps in the cycle are affected. However, a defect in any of the enzymes in this pathway is a serious issue and may cause hyperammonemia and lead rapidly to CNS edema, coma and death. Ornithine transcarbamoylase is the most common of these urea cycle defects and shows an X-linked inheritance pattern. The remainder of the known defects associated with the urea cycle are autosomal recessive. A deficiency of arginase, the last enzyme in the cycle, produces less severe symptoms but is nevertheless characterized by increased concentrations of blood arginine and at least a moderate increase in blood ammonia. In individuals with high blood concentrations of ammonia, hemodialysis must be used, often followed by intravenous administration of sodium benzoate and phenyllactate. These compounds are conjugated with glycine and glutamine, respectively, to form water-soluble adducts, trapping the ammonia in a nontoxic form that can be excreted in the urine.

The concept of nitrogen balance

A careful balance is maintained between nitrogen ingestion and secretion

Because there is no significant storage form of nitrogen or amino compounds in humans, nitrogen metabolism is quite dynamic. In an average, healthy diet, the

protein content exceeds the amount required to supply essential and nonessential amino acids for protein synthesis, and the amount of nitrogen excreted is approximately equal to that taken in. Such a healthy adult would be said to be ‘in neutral nitrogen balance’. When there is a need to increase protein synthesis, such as in recovering from trauma or in a rapidly growing child, the amount of nitrogen excreted is less than that consumed in the diet, and the individual would be in ‘positive nitrogen balance’. The converse is true in protein malnutrition: because of the need to synthesize essential body proteins, other proteins, particularly muscle proteins, are degraded and more nitrogen is lost than is consumed in the diet. Such an individual would be said to be in ‘negative nitrogen balance’. Fasting, starvation and poorly controlled diabetes are also characterized by negative nitrogen balance, as body protein is degraded to amino acids and their carbon skeletons are used for gluconeogenesis. The concept of nitrogen balance is clinically important because it reminds us of the continuous turnover of amino acids and proteins in the body (see Chapter 22).

Metabolism of the carbon skeletons of amino acids

Metabolism of amino acids interfaces with carbohydrate and lipid metabolism

When one examines the metabolism of the carbon skeletons of the 20 common amino acids, there is an obvious interface with carbohydrate and lipid metabolism. Virtually all the carbons can be converted into intermediates in the glycolytic pathway, the TCA cycle or lipid metabolism. The first step in this process is the transfer of an α-amino group by transamination to α-ketoglutarate or oxaloacetate, providing glutamate and aspartate, the sources for the nitrogen atoms of the urea cycle (Fig. 19.9). The single exception to this is lysine, which does not undergo transamination. Although the details of pathways for the various amino acids vary, the general rule is that there is loss of the amino group, followed by either direct metabolism in a central pathway (glycolysis, the TCA cycle or ketone body metabolism), or one or more intermediate conversions to yield a metabolite in one of the central pathways. Examples of amino acids that follow the former scheme include alanine, glutamate and aspartate, which yield pyruvate, α-ketoglutarate and oxaloacetate, respectively. The branched-chain amino acids, leucine, valine and isoleucine, and the aromatic amino acids, tyrosine, tryptophan and phenylalanine, are examples of the latter, more complex pathways.

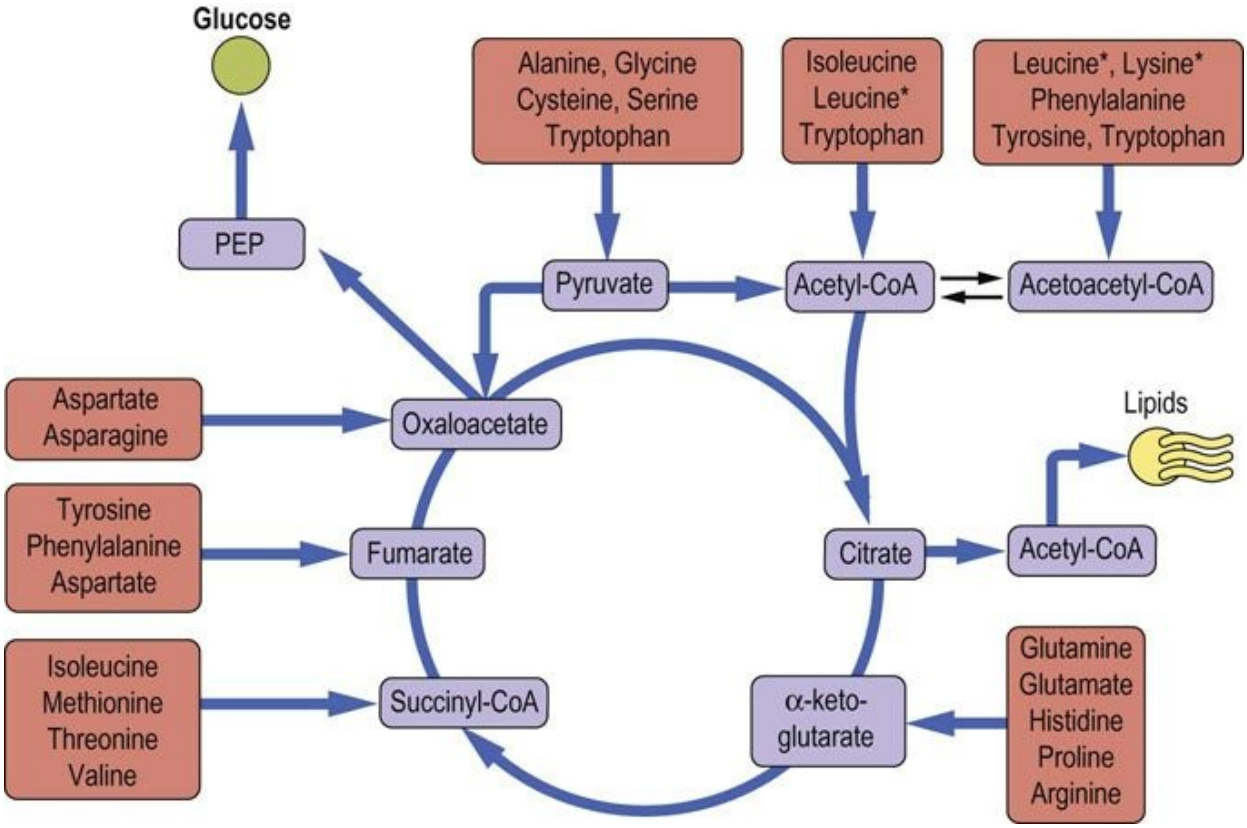


FIG. 19.9 Amino acid metabolism and central metabolic pathways.
This figure summarizes the interactions between amino acid metabolism and central metabolic pathways. *The amino acids marked with an asterisk are ketogenic only. PEP, phosphoenolpyruvate.

Clinical box Homocystinuria

A 21-year-old male was admitted to hospital following an episode of loss of speech and severe weakness on his right side. A diagnosis of ischemic stroke was made and the patient was treated with anticoagulant therapy and improved. Laboratory results indicated substantially elevated levels of blood homocysteine. The patient made a significant recovery and was discharged on a modified diet along with supplements of vitamin B_6, folic acid and vitamin B_{12}.

Comment.

Homocystinuria is a relatively rare autosomal recessive condition (1 in 200,000 births) which results in a variety of symptoms

including mental retardation, vision problems, and thrombotic strokes and coronary artery disease at a young age. The condition is caused by lack of an enzyme which catalyzes the transfer of sulfur from homocysteine to serine through the formation of a cystathionine intermediate. Some of these patients respond to vitamin supplementation.

■ Moderately elevated levels of homocysteine in plasma are implicated in the development of cardiovascular disease and cerebrovascular ischemic episodes (stroke). Cross-sectional and retrospective studies suggest that even moderately elevated levels of homocysteine may be correlated with increased incidence of heart disease and stroke, but the jury is still out as to whether lowering homocysteine levels will reduce the development of these serious illnesses.

Amino acids may be either glucogenic or ketogenic

Depending on the point at which the carbons from an amino acid enter central metabolism, that amino acid may be considered to be either **glucogenic** or **ketogenic**, *i.e.* possessing the ability to increase the concentrations of either glucose or ketone bodies, respectively, when fed to an animal. Those amino acids that feed carbons into the TCA cycle at the level of α-ketoglutarate, succinyl-CoA, fumarate or oxaloacetate, and those that produce pyruvate can all give rise to the net synthesis of glucose via gluconeogenesis and are hence designated glucogenic. Those amino acids that feed carbons into central metabolism at the level of acetyl-CoA or acetoacetyl-CoA are considered ketogenic. Because of the nature of the TCA cycle, no net flow of carbons can occur between acetate or its equivalent (e.g. butyrate or acetoacetate) from ketogenic amino acids to glucose via gluconeogenesis (see Chapter 13).

Several amino acids, primarily those with more complex or aromatic structures, can yield both glucogenic and ketogenic fragments (see Fig. 19.9). Only the amino acids leucine and lysine are regarded as being exclusively ketogenic and, because of its complex metabolism and lack of ability to undergo transamination, some authors do not consider lysine to be exclusively ketogenic. These classifications may be summarized as follows:

■ **Glucogenic amino acids**: alanine, arginine, asparagine, aspartic acid, cysteine, cystine, glutamine, glutamic acid, glycine, histidine, methionine, proline, serine, valine.
■ **Ketogenic amino acids**: leucine, lysine.
■ **Both glucogenic and ketogenic amino acids**: isoleucine, phenylalanine, threonine, tryptophan, tyrosine.

Metabolism of the carbon skeletons of selected amino acids

The 20 amino acids are metabolized by complex pathways to various intermediates in carbohydrate and lipid metabolism

Alanine, aspartate and glutamate are examples of glucogenic amino acids. In each case, through either transamination or oxidative deamination, the resulting α-keto acid is a direct precursor of oxaloacetate via central metabolic pathways. Oxaloacetate can then be converted to PEP, and subsequently to glucose via gluconeogenesis. Other glucogenic amino acids reach the TCA cycle or related metabolic intermediates through several steps, after the removal of the amino group (Fig. 19.10).

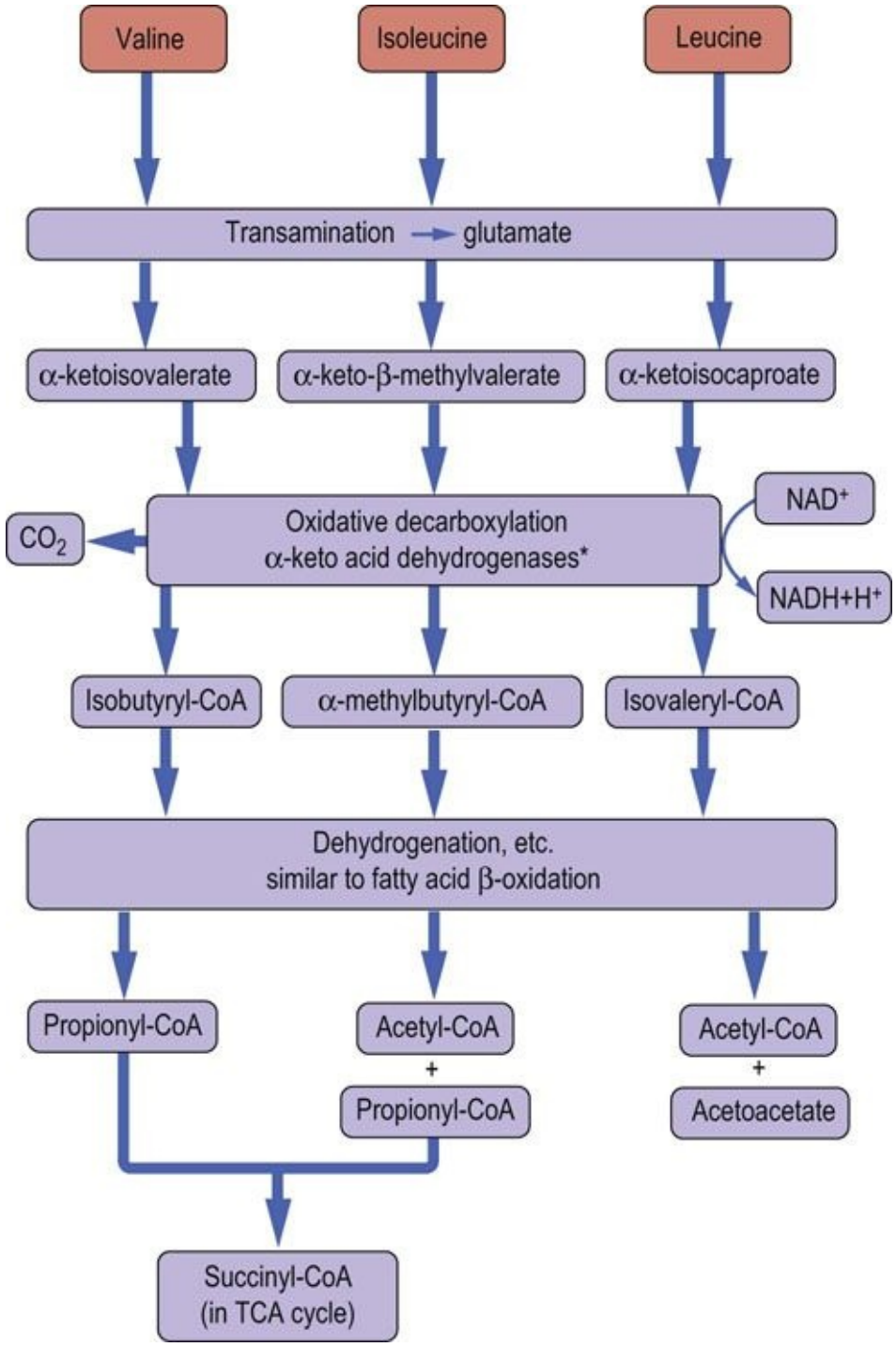


FIG. 19.10 Degradation of branched-chain amino acids.
Metabolism of the branched-chain amino acids produces acetyl-CoA and acetoacetate. In the case of valine and isoleucine, propionyl-CoA is produced and metabolized, in two steps, to succinyl-CoA (Fig. 15.5). *The branched-chain amino acid dehydrogenases are structurally related to pyruvate dehydrogenase and α-ketoglutarate dehydrogenase, and use the cofactors thiamine pyrophosphate, lipoic acid, flavin adenine dinucleotide, NAD^+ and CoA.

Leucine is an example of a ketogenic amino acid. Its catabolism begins with transamination to produce 2-keto-isocaproate. The metabolism of 2-ketoisocaproate requires oxidative decarboxylation by a dehydrogenase complex to produce isovaleryl-CoA. Further metabolism of isovaleryl-CoA leads to formation of 3-hydroxy-3-methylglutaryl-CoA, a precursor of both acetyl-CoA and the ketone bodies. The metabolism of leucine and the other branched-chain amino acids is summarized in Figure 19.10. Propionyl-CoA derived from either amino acid degradation or odd-chain fatty acid metabolism is converted to succinyl-CoA (see Fig. 15.5).

Tryptophan is a good example of an amino acid that yields both glucogenic and ketogenic precursors. After cleavage of its heterocyclic ring and a complex set of reactions, the core of the amino acid structure is released as alanine (a glucogenic precursor), while the balance of the carbons are ultimately converted to glutaryl-CoA (a ketogenic precursor). Figure 19.11 summarizes key points in the catabolism of the aromatic amino acids.

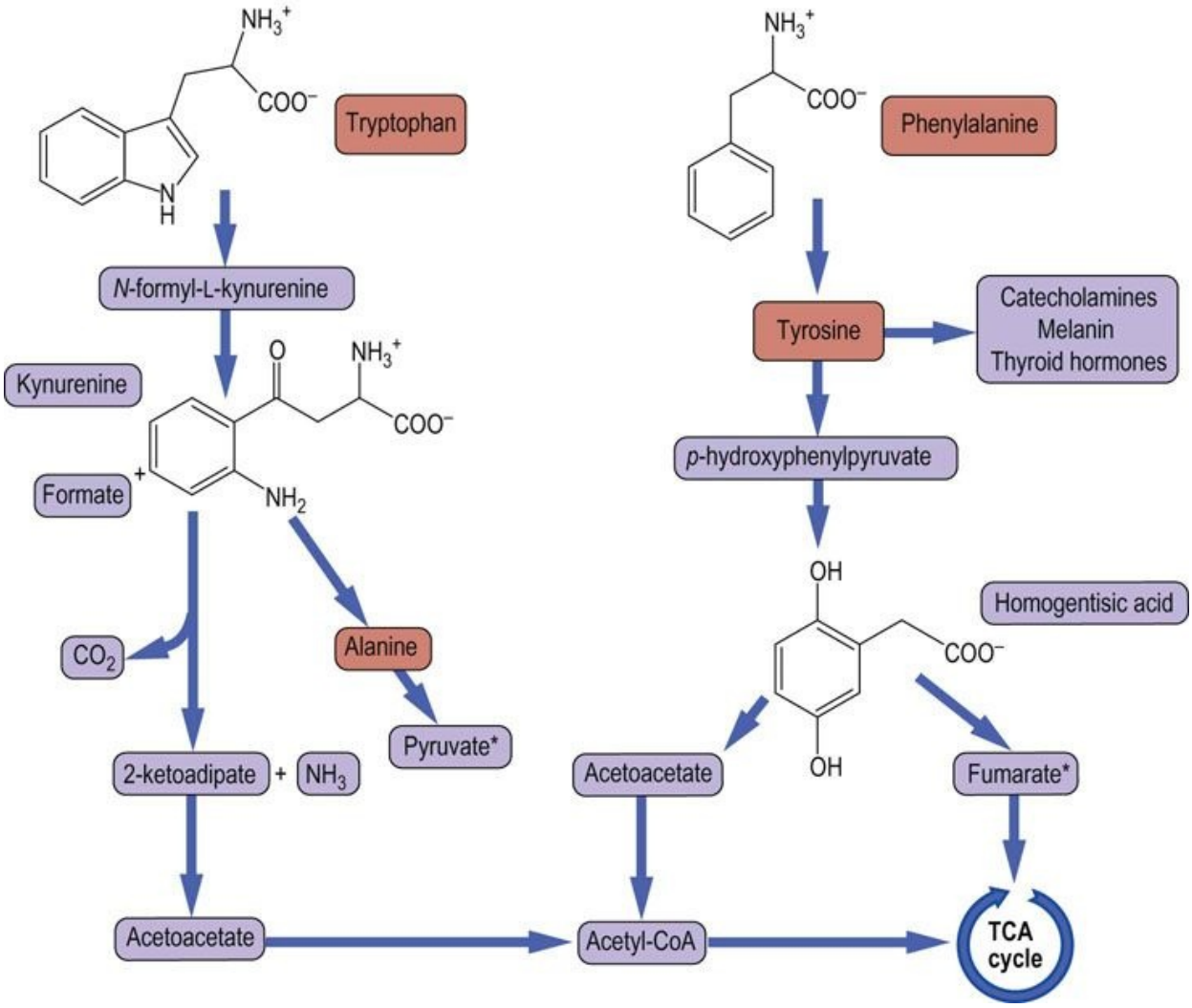


FIG. 19.11 Catabolism of aromatic amino acids.
This figure summarizes the catabolism of the aromatic amino acids, illustrating the pathways that lead to ketogenic and glucogenic precursors derived from both tyrosine and tryptophan. *Both pyruvate and fumarate can lead to net glucose synthesis. They constitute the gluconeogenic portions of the metabolism of these amino acids.

Clinical box Histamine, antihistamines and allergy

An 8-year-old, male child was referred to an allergy clinic due to repeated bouts of eczema with intense itching. He had no other known health issues. Previous treatment consisted of oral antihistamine medication which provided some relief but did not prevent recurrence of the problem. Following extensive testing, it

was found that while he was marginally positive for allergic reaction to dog and cat dander and to house dust mites, he was strongly positive for allergy to tomato. Examination of the boy's diet (he was quite fond of both pizza and spaghetti with tomato-based sauce) yielded a correlation of his bouts of eczema with his consumption of tomato-containing products. A dietary modification to avoid tomato products was initiated and his symptoms became immediately less frequent and were well managed with oral antihistamine medication and the occasional use of topical steroid creams.

This is a good example of both the importance of appropriate allergy testing and the importance of antihistamines in the treatment of allergic reactions. This class of medication (there are many currently available) acts by interfering with the interaction of histamine with its receptor or inhibiting the production of histamine from its precursor, the amino acid histidine.

Biosynthesis of amino acids

Evolution has left our species without the ability to synthesize almost half the amino acids required for synthesis of proteins and other biomolecules

Humans use 20 amino acids to build peptides and proteins that are essential for the many functions of their cells. Biosynthesis of the amino acids involves synthesis of the carbon skeletons for the corresponding α-keto acids, followed by addition of the amino group via transamination. However, humans are capable of carrying out the biosynthesis of the carbon skeletons of only about half of those α-keto acids. Amino acids that we cannot synthesize are termed **essential amino acids**, and are required in the diet. While almost all the amino acids can be classified as clearly essential or nonessential, a few require further qualification. For example, although cysteine is not generally considered an essential amino acid because it can be derived from the nonessential amino acid serine, its sulfur must come from the required or essential amino acid methionine. Similarly, the amino acid tyrosine is not required in the diet, but must be derived from the essential amino acid phenylalanine. This relationship between phenylalanine and tyrosine will be discussed further in considering the inherited disease phenylketonuria (PKU). Tables 19.5 and 19.6 list the nonessential and essential amino acids, and the source of the carbon skeleton in the case of those not required in the diet.

Table 19.5

Origins of nonessential amino acids

Amino acid	Source in metabolism, *etc.*
Alanine	From pyruvate via transamination
Aspartic acid, asparagine, arginine, glutamic acid, glutamine, proline	From intermediates in the citric acid cycle
Serine	From 3-phosphoglycerate (glycolysis)
Glycine	From serine
Cysteine*	From serine; requires sulfur derived from methionine
Tyrosine*	Derived from phenylalanine via hydroxylation

*These are examples of nonessential amino acids that depend on adequate amounts of an essential amino acid.

Table 19.6

Essential dietary amino acids

Mnemonic	Amino acid*	Notes or comments
P	Phenylalanine	Required in the diet also as a precursor of tyrosine
V	Valine	One of three branched-chain amino acids
T	Threonine	Metabolized like a branched-chain amino acid
T	Tryptophan	Its heterocyclic indole side chain cannot be synthesized in humans
I	Isoleucine	One of three branched-chain amino acids
M	Methionine	Provides the sulfur for cysteine and participates as a methyl donor in metabolism; the homocysteine is recycled
H	Histidine	Its heterocyclic imidazole side chain cannot be synthesized in humans
A	Arginine	Whereas arginine can be derived from ornithine in the urea cycle in amounts sufficient to support the needs of adults, growing animals require it in the diet
L	Leucine	A pure ketogenic amino acid
L	Lysine	Does not undergo direct transamination

*The mnemonic PVT TIM HALL is useful for recalling the names of the essential amino acids.

Amino acids are precursors of many essential compounds

In addition to their role as the building blocks for peptides and proteins, amino acids are essential precursors of a number of neurotransmitters, hormones, inflammatory mediators, and carrier and effector molecules (Table 19.7). Examples of this include histidine, which serves as the precursor of histamine (the mediator of inflammation released by mast cells and lymphocytes), glutamate, glycine and aspartate, which serve directly as a neurotransmitters. Additional examples include gamma amino butyric acid (GABA), which is derived from glutamate, and tyrosine, which is derived from phenylalanine. Tyrosine is then the precursor of the neurotransmitters 1,3-dihydroxyphenylalanine (DOPA), dopamine, epinephrine, the thyroid hormones triiodothyronine and thyroxine, and melanin.

Table 19.7

Examples of amino acids as effector molecules or precursors

Amino acid	Effector molecule or prosthetic group
Arginine	Immediate precursor of urea, precursor of nitric oxide
Aspartate	An excitatory neurotransmitter
Glycine	An inhibitory neurotransmitter; precursor of heme
Glutamate	Excitatory neurotransmitter; precursor of γ-amino butyric (GABA), an inhibitory neurotransmitter
Histidine	Precursor of histamine, a mediator of inflammation and a neurotransmitter
Tryptophan	Precursor of serotonin, a potent smooth muscle contraction stimulator; precursor of melatonin, a regulator of circadian rhythm
Tyrosine	Precursor of the hormones and neurotransmitters catecholamines, dopamine, epinephrine and norepinephrine, thyroxine

Inherited diseases of amino acid metabolism

In addition to deficiencies in the urea cycle, defects in the metabolism of the carbon skeletons of various amino acids were among the first disease states to be associated with simple inheritance patterns. These observations gave rise to the concept of the genetic basis of inherited metabolic disease states, also known as **inborn errors of metabolism**. Garrod considered a number of disease states that appeared to be inherited in a Mendelian pattern, and proposed a correlation between these abnormalities and specific genes, in which the disease state could be either dominant or recessive. Dozens of inborn errors of amino acid metabolism have now been described, and the molecular defect has been described for many of them. Three classic inborn errors of metabolism will be discussed in some detail here.

Phenylketonuria (PKU)

The common form of PKU results from a deficiency of the enzyme phenylalanine hydroxylase. The hydroxylation of phenylalanine is a required step in both the normal degradation of the carbon skeleton of this amino acid and the synthesis of tyrosine (Fig. 19.12). When untreated, this metabolic defect leads to excessive urinary excretion of phenylpyruvate and phenyllactate, and severe mental retardation. In addition, individuals with PKU tend to have very light skin pigmentation, unusual gait, stance, and sitting posture, and a high frequency of epilepsy. In the USA, this autosomal recessive defect occurs in about 1 in 30,000 live births. Because of its frequency, and the ability to prevent the most serious consequences of the defect by a low-phenylalanine diet, newborns in most developed countries are routinely tested for blood concentrations of phenylalanine. Fortunately, with early detection and the use of a diet restricted in phenylalanine but supplemented with tyrosine, most of the mental retardation can be avoided. Mothers who are homozygous for this defect have a very high probability of bearing children with congenital defects and mental retardation unless their blood phenylalanine concentrations can be controlled by diet. The developing fetus is very sensitive to the toxic effects of high concentrations of phenylalanine and related phenylketones. Not all hyperphenylalaninemias are caused by a defect in phenylalanine hydroxylase. In

some cases, there is a defect in biosynthesis or reduction of a required tetrahydrobiopterin cofactor.

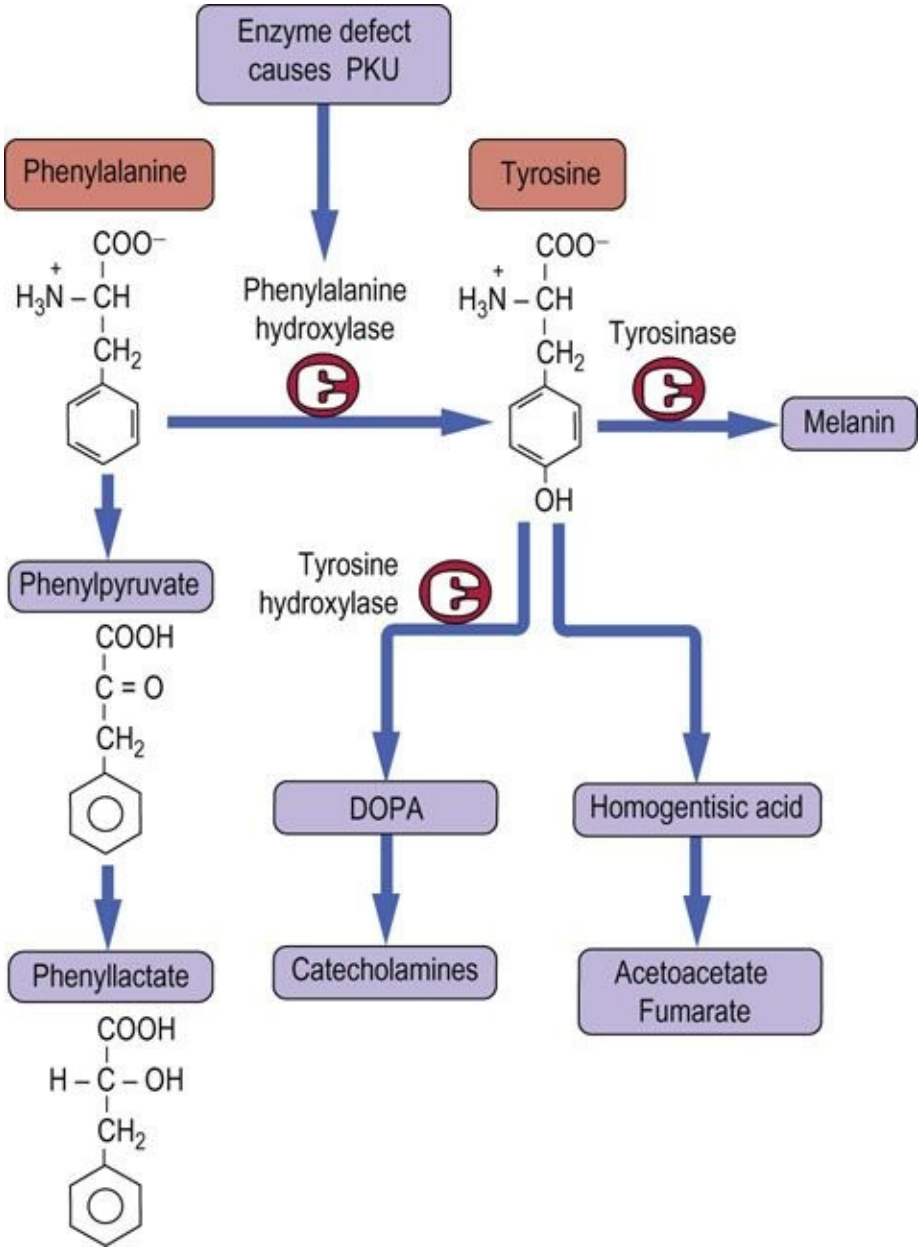


FIG. 19.12 Degradation of phenylalanine.
In order to enter normal metabolism, phenylalanine must be hydroxylated by the enzyme phenylalanine hydroxylase. A defect in this enzyme leads to phenylketonuria (PKU). Tyrosine is a precursor of acetyl-CoA and fumarate, catecholamine hormones, the neurotransmitter dopamine, and the pigment melanin. DOPA, dihydroxyphenylalanine.

Clinical box Albinism

A full-term infant, born to a normal and healthy mother and father, was observed to have a marked lack of pigmentation. The infant, who appeared to be otherwise normal, had blue eyes and very light blond, almost white, hair. This lack of pigmentation was confirmed as classic albinism on the basis of a family history and the establishment of a lack of the enzyme tyrosinase, which is responsible for a two-step hydroxylation of tyrosine to dihydroxyphenylalanine (DOPA) and a subsequent further oxidation to a quinone, a precursor of melanin in melanocytes.

Comment.

The primary cause of albinism is a homozygous defect in either tyrosinase or an accessory P protein. A separate DOPA-producing enzyme, tyrosine hydroxylase, is involved in biosynthesis of the catecholamine neurotransmitters, so albinos do not appear to have neurologic deficits. As a result of their lack of pigmentation, however, they are quite sensitive to damage from sunlight and must take added precautions against ultraviolet radiation from the sun. Albinos are generally very sensitive to bright light. They may have normal eyesight, in spite of the lack of pigmentation – retinal pigments are derived from carotene (vitamin A), rather than tyrosine.

Alkaptonuria (black urine disease)

A second inherited defect in the phenylalanine–tyrosine pathway involves a deficiency in the enzyme that catalyzes the oxidation of homogentisic acid, an intermediate in catabolism of tyrosine and phenylalanine. In this condition, which occurs in 1 in 1,000,000 live births, homogentisic acid accumulates and is excreted in urine. This compound oxidizes to alkaptone on standing or on treatment with alkali, and gives the urine a dark color. Individuals with **alkaptonuria** ultimately suffer from deposition of dark (ochre-colored) pigment

in cartilage tissue, with subsequent tissue damage, including severe arthritis; the onset of these symptoms is generally in the third or fourth decade of life. This autosomal recessive disease was the first of several that Garrod considered in proposing his initial hypothesis for inborn errors of metabolism. Although alkaptonuria is relatively benign compared with PKU, little is available in the way of treatment, other than symptomatic relief.

Advanced concept box

Selenocysteine

In addition to the 20 common amino acids found in proteins, a 21st amino acid has been discovered and shown to be an active site amino acid in several enzymes, including the antioxidant enzyme glutathione peroxidase (Chapter 37) and 5′-deiodinases (important in the metabolism of thyroid hormones) (Fig. 39.8). Selenocysteine is derived from serine and has unique chemical properties. It is because of the need for selenocysteine that trace amounts of selenium are required in the diet. It should be noted that while selenocysteine is incorporated per se into the enzymes in which it functions, a number of other unusual amino acids may be found in some proteins due to post-translational modification. Examples of this can be seen in the collagens and connective tissue proteins which contain hydroxylated forms of proline and lysine, which are formed following the incorporation of proline and lysine into the protein polypeptide (see Chapter 29).

Maple syrup urine disease (MSUD)

The normal metabolism of the branched-chain amino acids leucine, isoleucine and valine involves loss of the α-amino group, followed by oxidative decarboxylation of the resulting α-keto acid. This decarboxylation step is catalyzed by branched-chain keto acid decarboxylase, a multienzyme complex associated with the inner membrane of the mitochondrion. In approximately 1 in 300,000 live births, a defect in this enzyme leads to accumulation of the keto acids corresponding to these branched-chain amino acids in the blood, and then

to branched-chain ketoaciduria. When untreated or unmanaged, this condition may lead to both physical and mental retardation of the newborn and a distinct maple syrup odor of the urine. This defect can be partially managed with a low-protein or modified diet, but not in all cases. In some instances, supplementation with high doses of thiamine pyrophosphate, a cofactor for this enzyme complex, has been helpful.

Clinical box Cystinuria

A 21-year-old man came to the emergency room with severe pain in his right side and back. Subsequent investigation indicated a kidney stone, and increased concentrations of cystine, arginine, and lysine in the urine. This patient exhibited the characteristic symptoms of cystinuria.

Comment.

Cystinuria is an autosomal recessive disorder of intestinal absorption and proximal tubular reabsorption of dibasic amino acids; it does not result from a defect in cysteine metabolism per se. Because of the transport deficiency, cysteine, which is normally reabsorbed in the proximal renal tubule, remains in the urine. The cysteine spontaneously oxidizes to its disulfide form, cystine. Cystine is relatively insoluble and tends to precipitate in the urinary tract, forming kidney stones. The condition is generally treated by restricting the dietary intake of methionine (a biosynthetic precursor of cysteine), encouraging high fluid intake to keep the urine dilute and, more recently, with drugs that convert urinary cysteine to a more soluble compound that will not precipitate.

Summary

In this chapter, we have seen that the metabolism of amino acids is integrally related to the mainstream of metabolism.

■ The catabolism of amino acids generally begins with the removal of the α-amino group, which is transferred to α-ketoglutarate and oxaloacetate, and ultimately excreted in the form of urea.

■ The resulting carbon skeletons are converted to intermediates that enter central metabolism at various points.

■ Because carbon skeletons corresponding to the various amino acids can be derived from or fed into the glycolytic pathway, the TCA cycle, fatty acid biosynthesis and gluconeogenesis, amino acid metabolism should not be considered as an isolated pathway.

■ Although amino acids are not stored like glucose (glycogen) or fatty acids (triglycerides), they have an important and dynamic role, not only in providing the building blocks for the synthesis and turnover of protein but also in normal energy metabolism, providing a carbon source for gluconeogenesis when needed and an energy source of last resort in starvation.

■ Amino acids provide precursors for the biosynthesis of a variety of small signaling molecules, including hormones and neurotransmitters.

■ The severe consequences of inherited diseases such as phenylketonuria and maple syrup urine disease illustrate the effects of abnormal amino acid metabolism.

Active learning

1. Tyrosine is included as a supplement in the diet plan for individuals with phenylketonuria. What is the rationale for this supplement? Compare the therapeutic approaches used for treatment of the various forms of PKU in which phenylalanine hydroxylase is not affected.
2. Review the rationale for the use of levodopa, catechol-*O*-methyltransferase inhibitors and monoamine oxidase inhibitors for treatment of Parkinson's disease.
3. Review the pathways for biosynthesis of the neurotransmitters serotonin, melatonin, dopamine, and the catecholamines. What enzymes are involved in the inactivation of these compounds?

Further reading

Dietzen, DJ, Rinaldo, P, Whitley, RJ, *et al*. National academy of clinical biochemistry laboratory medicine practice guidelines: follow-up testing for metabolic disease identified by expanded newborn screening using tandem mass spectrometry; executive summary. *Clin Chem*. 2009; 55:1615–1626.

Gropman, AL, Summar, M, Leonard, JV. Neurological implications of urea cycle disorders. *J Inherit Metab Dis*. 2007; 30:865–879.

Kuhara, T. Noninvasive human metabolome analysis for differential diagnosis of inborn errors of metabolism. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2007; 855:42–50.

Mitchell, JJ, Trakadis, YJ, Scriver, CR. Phenylalanine hydroxylase deficiency. *Genet Med*. 2011; 13:697–707.

Morris, SM, Jr. Regulation of enzymes of the urea cycle and arginine metabolism. *Annu Rev Nutr*. 2002; 22:87–105.

Morris, SM, Jr. Arginine: beyond protein. *Am J Clin Nutr*. 2006; 83(suppl):508S–512S.

Lindemann, B, Ogiwara, Y, Ninomiya, Y. The discovery of Umami. *Chem Senses*. 2002; 27:843–844.

Ogier de Baulny, H, Saudubray, JM. Branched-chain organic acidurias. *Semin Neonatal*. 2002; 7:65–74.

Saudubray, JM, Nassogne, MC, de Lonlay, P, *et al*. Clinical approach to inherited metabolic disorders in neonates: an overview. *Semin Neonatal*. 2002; 7:3–15.

Singh, RH. Nutritional management of patients with urea cycle disorders. *J Inherit Metab Dis*. 2007; 30:880–887.

Summar, ML, Dobbelaere, D, Brusilow, S, *et al*. Diagnosis, symptoms, frequency and mortality of 260 patients with urea cycle disorders from a 21-year, multicentre study of acute hyperammonaemic episodes. *Acta Paediatr*. 2008; 97:1420–1425.

Wilcken, B. Screening for disease in the newborn: the evidence base for blood-spot screening. *Pathology*. 2012; 44:73–79.

Websites

Society for the Study of Inborn Errors of Metabolism (SSIEM) www.ssiem.org

Urea cycle disorders:

www.uptodate.com/contents/ureacycle-disorders-clinical-features-and-diagnosis

www.ncbi.nlm.nih.gov/books/NBK1217/

www.ureacycle.com

Nitrogen metabolism. http://themedicalbiochemistrypage.org/nitrogen-metabolism.html.

Maple syrup urine disease. www.msud-support.org/overview.htm.

Parkinson's disease. www.enotes.com/parkinsons-disease-reference/parkinsons-disease-172267.

Phenylketonuria. www.nlm.nih.gov/medlineplus/phenylketonuria.html.

CHAPTER 20

Muscle

Energy Metabolism and Contraction

John W. Baynes and Matthew C. Kostek

Learning objectives

After reading this chapter you should be able to:

■ Describe muscle structure and its function in mechanical force production, including differences among skeletal, cardiac and smooth muscle types that are related to their physiologic functions.

■ Describe the structure and protein composition of the sarcomere, the sliding filament model of muscle contraction, and the source of the banding pattern in striated muscle. Describe the sequence of events in excitation–contraction coupling, including the roles of membrane depolarization, the sarcoplasmic reticulum, and calcium triggering.

■ Identify the key sites of energy utilization during muscle contraction, the role of creatine phosphate in skeletal muscle, and the impact of skeletal muscle fiber type on substrate utilization and muscle function.

■ Describe the changes in skeletal muscle mass and metabolism with age, in response to acute and prolonged exercise, and in diseases such as sarcopenia, metabolic syndrome, and wasting conditions.

Introduction

There are three types of muscle: – skeletal, cardiac, and smooth muscle – each with a unique physiologic role

All muscles function to convert chemical energy to mechanical energy, but the various types of muscle differ in their mechanism of initiation of contraction, rate of force development, duration of contraction, ability to adapt to their environment, and substrate utilization. Muscle accounts for about 40% of total body mass, and muscle metabolism is a major determinant of whole-body metabolic rate in both the basal and active state. Changes in skeletal muscle metabolism occur with physical activity and are directly related to the required force output and duration of activity. These factors also affect the muscle's relative utilization of glucose and fatty acids for fuel. Besides locomotion, skeletal muscle is also a source of body heat, provides amino acids for hepatic gluconeogenesis during fasting, and is a major site of glucose and triglyceride disposal following a meal. Because of its critical role in the regulation of systemic fuel flux and metabolism, loss of muscle mass has a profound effect on overall metabolism. Advancing age, sepsis, and wasting diseases, such as AIDS and cancer, are conditions associated with loss of muscle mass, and this loss is associated with increased morbidity and mortality.

The primary focus of this chapter will be on skeletal muscle, supplemented by discussion of similarities and differences in skeletal, cardiac and smooth muscle structure, function and metabolism. The chapter will begin with a discussion of the mechanism of muscle contraction, proceed to the signaling that initiates the contractile process, and then examine energy metabolism essential for contraction.

Clinical box Muscular dystrophies

A young boy was brought to the clinic because his mother had noticed that he walked with a waddling gait. Physical evaluation confirmed muscle weakness especially in the legs, although his calf muscles were large and firm. There was a 20-fold elevation in serum creatine (phospho) kinase (CK) activity, identified as the MM (muscle) isozyme. Histology revealed muscle loss, some

necrosis, and increased connective tissue and fat volume in muscle. A tentative diagnosis of Duchenne muscular dystrophy (DMD) was confirmed by immunoelectrophoretic (Western blot) analysis showing the lack of the cytoskeletal protein dystrophin in muscle.

Comment.

Though there are many forms of muscular dystrophy, some genetic and some acquired, DMD is the most common genetic dystrophy and is lethal. **Dystrophin** is a high-molecular-weight cytoskeletal protein that reinforces the plasma membrane of the muscle cell and mediates interactions with the extracellular matrix. In its absence, the plasma membrane of muscle cells shears during the contractile process, leading to muscle cell death.

The dystrophin gene is located on the X-chromosome and is nearly 2.5×10^6 base pairs in length. Spontaneous mutations in this gene are relatively common, the frequency of DMD being approximately 1 in 3500 male births. DMD is a progressive myodegenerative disease, commonly leading to confinement to a wheelchair by puberty, with death by age 30 years from respiratory or cardiac failure. Dystrophin is completely absent in DMD patients, though a variant of the disease, known as Becker muscular dystrophy, has milder symptoms and is characterized by expression of an altered dystrophin protein and survival into the fifth decade. Although there is currently no treatment for DMD, gene therapy still holds some promise and newer technologies that utilize ‘exon skipping’ are allowing cells to skip over mutated exons and thereby translate a slightly smaller, but still functional, protein product. The smaller dystrophin protein produces Becker-like symptoms in animal experiments and thus could translate to a doubling of human life span if the results are reproducible in humans.

Muscle structure

The sarcomere: the functional contractile unit of muscle

A common characteristic of cardiac myocytes, smooth muscle cells and skeletal myofibers is that their cytoplasm is packed full of contractile protein. The contractile protein is organized in linear arrays of sarcomere units in skeletal myofibers and cardiac myocytes, giving these muscles a striated appearance; thus, the term **striated muscle**. Contractile protein in **smooth muscle** cells is not organized into a sarcomeric structure, and this tissue is described as nonstriated muscle. Skeletal muscle's hierarchic structure (Fig. 20.1) consists of bundles (fasciculi) of elongated, multinucleated fiber cells (myofibers). The myofiber cells contain bundles of myofibrils which are, in turn, composed of myofilament proteins, primarily myosin and actin, that form the sarcomere (Table 20.1). Electron microscopic analysis of muscle reveals a repeating pattern of light-and dark-staining regions in the myofibril (Fig. 20.2). These regions are known as the I (isotropic)- and A (anisotropic)-bands, respectively. At the center of the I-band is a discrete, darker staining Z-line, while the center of the A-band has a lighter-staining H-zone with a central M-line. The contractile unit, the **sarcomere**, is centered on the M-line, extending from one Z-line to the next. Smooth muscle lacks a defined Z-line.

Table 20.1

The structural elements of skeletal muscle arranged in descending order of size

Microscopic unit	Fasciculus: bundle of muscle cells
Cellular unit	Myofiber cell: long, multinucleated cell
Subcellular unit	Myofibril: composed of myofilament proteins
Functional unit	Sarcomere: contractile unit, repeating unit of the myofibril
Myofilament components	Proteins: primarily actin and myosin

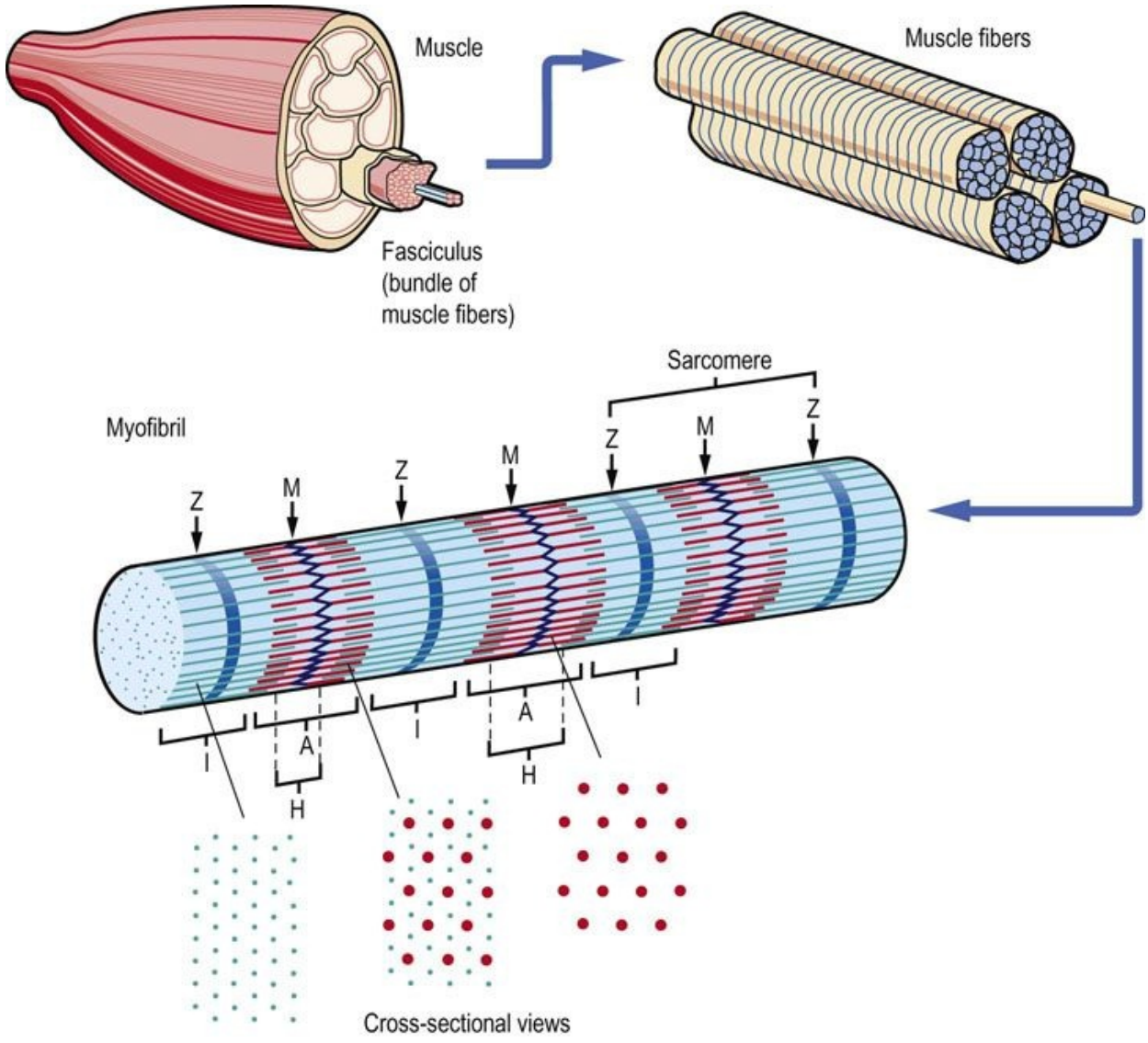


FIG. 20.1 Hierarchic structure of muscle.
Hierarchic structure of skeletal muscle, showing an exploding view of fasciculi, myofibers, myofibrils and myofilament proteins. The location of the I-band (thin, actin filaments extending from the Z-line) and the A-band (thick, myosin filaments, extending from the M-line), with darker-staining regions of the A-band corresponding to the region of overlap of actin and myosin filaments.

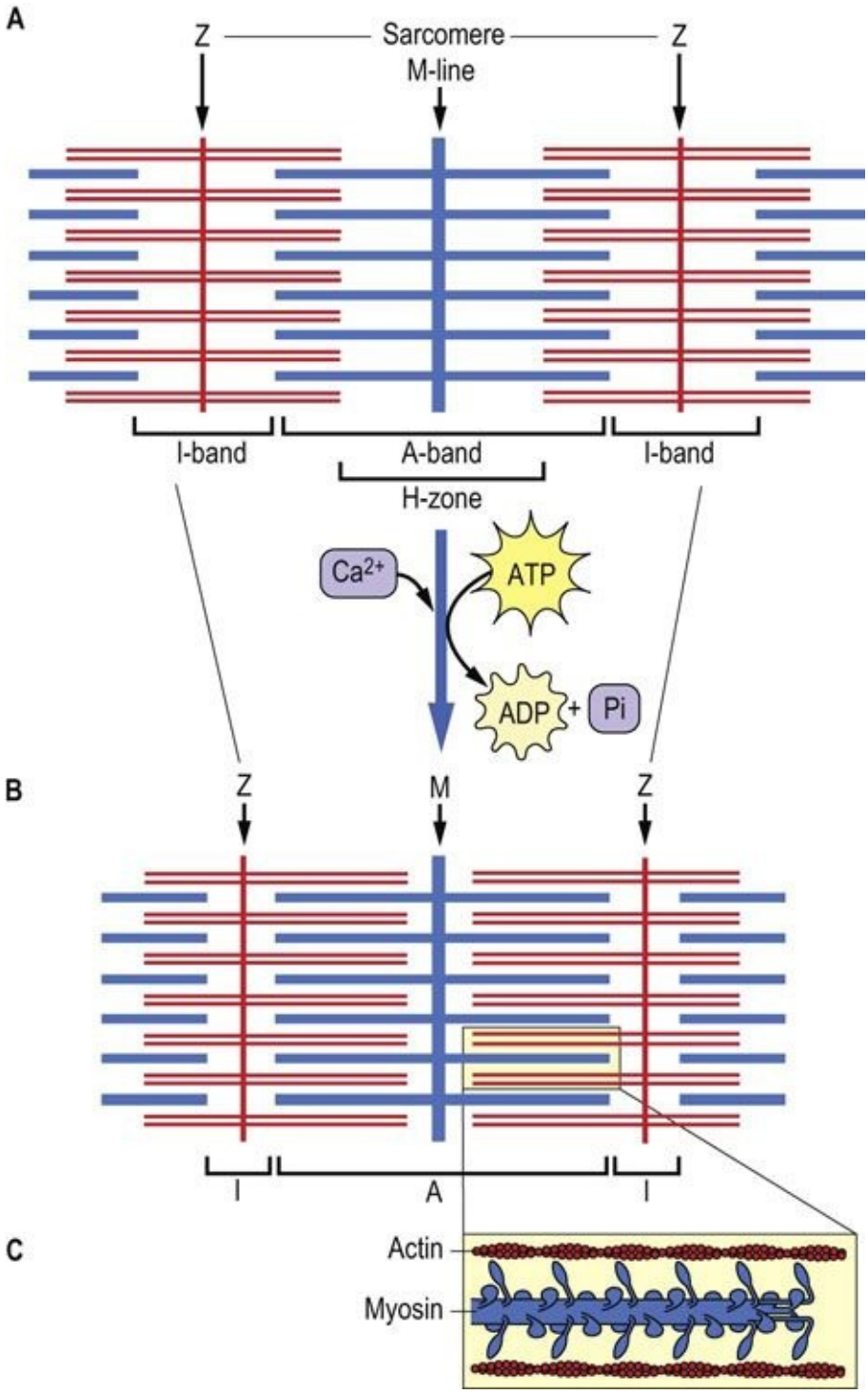


FIG. 20.2 Schematic structure of the sarcomere, indicating the distribution of actin and myosin in the A-and I-bands.
(A) Relaxed sarcomere. **(B)** Contracted sarcomere. **(C)** Magnification of contracted sarcomere, illustrating the polarity of the arrays of myosin molecules. Increased overlap of actin and myosin filaments during contraction, accompanied by a decrease in the length of the H-zones and I-bands, illustrates the sliding filament model of muscle

contraction.

The thick and thin filaments

Actin and Myosin account for over 75% of muscle protein.

The sarcomere may shorten by as much as 70% in length during muscle contraction (see Fig. 20.2). The components effecting the contraction are the thick and thin filaments. **The thick filament is composed of myosin and titin protein, and the thin filament is mainly made up of actin, with associated proteins, tropomyosin and troponins.** The thin filament also has some interaction with titin. Thick and thin filaments extend in opposite directions from both sides of the M-and Z-lines, respectively, and overlap and slide past one another during the contractile process (see Fig. 20.2). The M-and Z-lines are, in effect, base plates for anchoring the myosin and actin filaments. In striated muscle, increased thick–thin filament intercalate during contraction, causing the H-zone (myosin only) and I-bands (actin only) to shrink. In smooth muscle, thick and thin filaments are anchored at structures called dense bodies that are further anchored by intermediate filaments. Although all three muscle types contain the same proteins, each muscle type expresses tissue-specific isoforms; the cardiac actin and troponins, for example, differ slightly from those in skeletal muscle.

Sarcomere proteins

Myosin

Interaction between actin and myosin during muscle contraction is dependent on cytoplasmic Ca^{++} concentration

Myosin is one of the largest proteins in the body, with a molecular mass of approximately 500 kDa, and accounts for more than half of muscle protein (Table 20.2). Under the electron microscope, myosin appears as an elongated protein with two globular heads. It is the primary component of the thick filament in muscle. Each myosin molecule is made up of two heavy chains (approx. 200 kDa) and four light chains (approx. 20 kDa). The heavy chain can

be subdivided into the helical tail and globular head regions; the four light chains are bound to the globular heads. Structural analysis by limited proteolysis indicates that there are two flexible hinge regions in the myosin molecule (Fig. 20.3): one where the globular head attaches to the helical region and the other further into the helical region. The myosin filaments are associated through their helical regions and extend outward from the M-line toward the Z-line of each myofibril (see Figs 20.2 and 20.3). The hinge regions allow the myosin heads to interact with actin and provide the flexibility needed for reversible interactions and conformational changes during muscle contraction.

Table 20.2

Muscle proteins and their functions

Protein	Function
Myosin	Ca^{2+}-dependent ATPase activity
C-protein	assembly of myosin into thick filaments
M-protein	binding of myosin filaments to M-line
Actin	G-actin polymerizes to filamentous F-actin
tropomyosin	stabilization and propagation of conformational changes of F-actin
troponins-C, I and T	modulation of actin–myosin interactions
α-and β-actinins	stabilization of F-actin and anchoring to Z-line
nebulin	possible role in determining length of F-actin filaments
titin	control of resting tension and length of the sarcomere
desmin	organization of myofibrils in muscle cells
dystrophin	reinforcement of cytoskeleton and muscle cell plasma membrane

Actin and myosin account for over 90% of muscle proteins, but several associated proteins are required for assembly and function of the actomyosin complex.

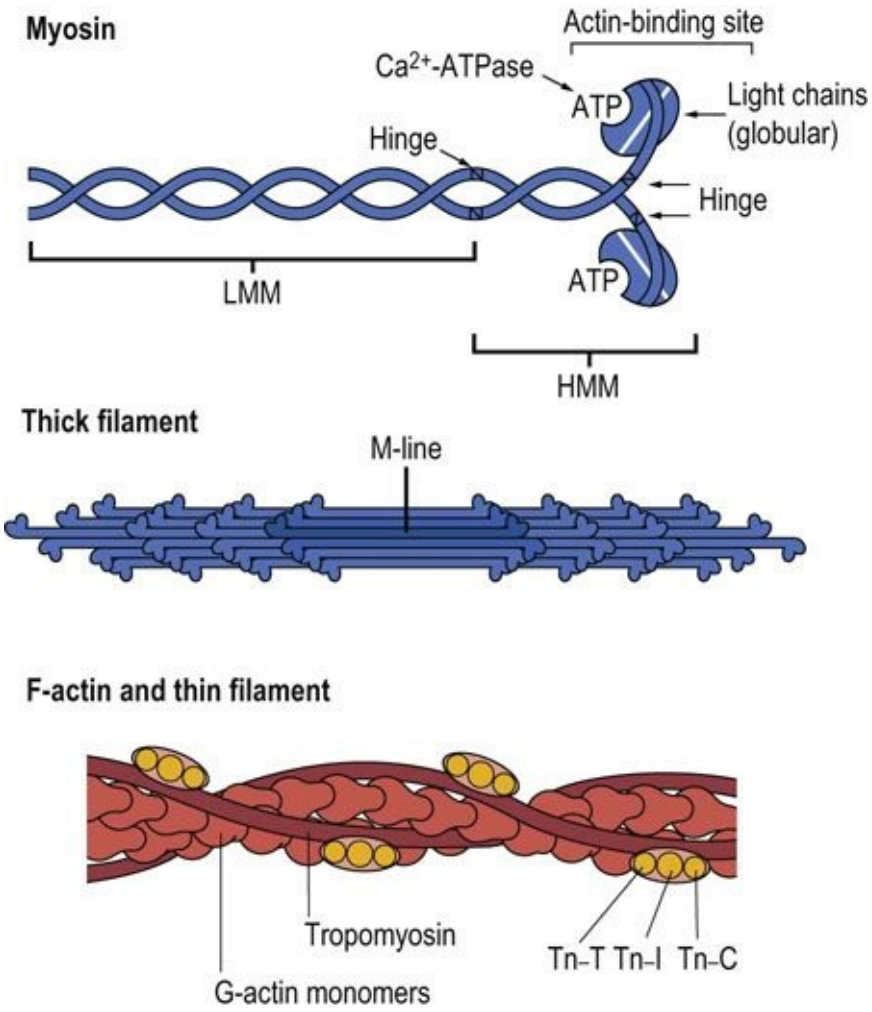


FIG. 20.3 Polymerization of myosin and actin into thick and thin filaments. Tn-C, calcium-binding troponin; Tn-I, troponin inhibitory subunit; Tn-T, tropomyosin-binding troponin. LMM, light meromyosin; HMM, heavy meromyosin.

There are several features of myosin that are essential for muscle contraction.

■ The myosin globular heads have binding sites for ATP and its hydrolysis products, ADP and phosphate (Pi).

■ The myosin globular heads have a Ca^{2+}-dependent ATPase activity.

■ Myosin binds reversibly to actin as a function of Ca^{2+}, ATP, and ADP + Pi concentrations.

■ The binding of calcium and hydrolysis of ATP lead to major changes in the conformation of the myosin molecule and its interaction with actin.

■ **Myosin-ATPase** activity, myosin–actin interactions and conformational changes are integrated into the sliding filament model of muscle contraction (below). They also explain the development of rigor mortis. The increase in Ca^{2+}

in the muscle cytoplasm (sarcoplasm) and decrease in ATP after death lead to tight binding between myosin and actin, forming rigid muscle tissue.

Actin

Actin is composed of 42 kDa subunits, known as **G-actin** (globular), which polymerize into a filamentous array (**F-actin**). Two polymer chains coil around one another to form the F-actin myofilament (see Fig. 20.3). F-actin is the major component of the thin filament and interacts with myosin in the actomyosin complex. The F-actin chains extend in opposite directions from the Z-line, overlapping with the myosin chains extending from the M-line. Each myosin-containing thick filament is surrounded by six actin molecule containing thin filaments. Each thin filament interacts with three myosin-containing thick filaments (see Fig. 20.1 for a cross-sectional view).

Tropomyosin and troponins

Troponins modulate the interaction between actin and myosin

Calcium activation of muscle contraction in striated muscle involves thin filament-associated proteins, tropomyosin and the troponins. **Tropomyosin** is a fibrous protein that extends along the grooves of F-actin, each molecule contacting about seven G-actin subunits. Tropomyosin has a role in stabilizing F-actin and coordinating conformational changes among actin subunits during contraction. In the absence of Ca^{2+}, tropomyosin blocks the myosin-binding site on actin.

A complex of **troponin** proteins is bound to tropomyosin: Tn-T (tropomyosin-binding), Tn-C (calcium-binding) and Tn-I (inhibitory subunit). Calcium binding to Tn-C, a calmodulin-like protein, induces changes in Tn-I which shift the interaction between tropomyosin and actin, exposing the myosin-binding site on F-actin and permitting actin–myosin interactions. For a description of the diagnostic use of cardiac troponin measurements see Box at end of this chapter.

Clinical box Muscle loss during sepsis

Skeletal muscle affects both the morbidity and mortality of sepsis. Sepsis is the leading cause of death in noncoronary intensive care

units (ICUs) and the 10th leading cause of death in the United States overall. The incidence of severe sepsis in the United States is approximately 700,000 cases per year and is increasing by 8% each year. Numerous factors are contributing to this increase (e.g. increasing population age, overuse of antibiotics). Sepsis can be defined as an inappropriate regulation of the immune and physiologic response to a pathogen. Normally, the body responds to an infection by eradicating the pathogen when it first comes in contact with immune cells. When this system becomes overwhelmed it disturbs whole body homeostasis and often results in multiple organ failure.

Indeed, the mortality rate of the more severe forms of sepsis is >50%. While controlling the inflammatory response and eradicating the pathogen is the primary therapeutic goal, skeletal muscle plays an important role in patient prognosis. Sepsis results in severe muscle mass loss and patients with low muscle mass are more likely to die of sepsis. The muscle degradation pathway is activated by the inflammatory cytokines such as IL-6 and TNF-α caused by the uncontrolled immune response. Muscle protein breakdown and amino acids release into the blood, are similar to starvation. However, different molecular pathways are involved; increasing nutrition or protein intake is not effective as anabolic signaling pathways are not responsive. As muscle loss is a contributing factor to morbidity and mortality of patients and affects long-term recovery outcomes, it becomes of paramount importance to consider methods of attenuating the muscle loss associated with sepsis. There are several protein degradation pathways related to muscle loss, but the ubiquitin-proteasome (UbP) pathway (see Chapter 34 for details) appears to be the primary pathway activated during sepsis. The myofibrillar proteins are most susceptible and 3-Methylhistidine, a post-translationally modified amino acid found in actin and myosin, is increased in blood and urine as a result of protein turnover. The UbP pathway is currently a target for drugs in development that might be used specifically to treat muscle loss during sepsis and thus improve mortality rates.

Titin

Titin modulates the passive tension of muscle

Titin is the largest protein in the human body, with more than 34,000 amino acids and a mass of 3800 kDa. Structurally, titin spans half the length of the sarcomere, with its *N*-terminus anchored to the Z-line, and its *C*-terminus to the thick filament at the M-line. Titin has an elastic, extensible **PEVK domain** (rich in Pro, Glu, Val and Lys) that contributes to passive myocardial and skeletal muscle tension, and a kinase domain that participates in intracellular signaling. Depending on the skeletal muscle, titin may account for more than half of the passive tension of the muscle, and contributes a spring-like property to the sarcomere – when a muscle is stretched, potential energy is stored in the PEVK domain, which re-coils during relaxation. Mutations in one region of titin may cause a genetic disease of the heart (e.g. hypertrophic cardiomyopathy), while a mutation elsewhere in the gene causes a disease of skeletal muscle only (e.g. limb girdle muscular dystrophy).

The contractile process

The sliding filament model of muscle contraction

The sliding filament model describes how a series of chemical and structural changes in the actomyosin complex can induce sarcomere shortening

The contractile response depends on reversible, Ca^{2+}-dependent **cross-bridge** formation between the myosin head and its binding site on actin. A conformational change in the hinge regions of myosin occurs after cross-bridge formation, providing the **power stroke** for muscle contraction (Fig. 20.4). This conformational change, the relaxation of the high-energy form of myosin, is accompanied by dissociation of ADP and Pi. After the stroke is completed, the binding and hydrolysis of ATP restore the high-energy conformation. The stability of the contracted state is maintained by multiple and continuous Ca^{2+}-dependent actin–myosin interactions, so that slippage is minimized until calcium is removed from the sarcoplasm, allowing dissociation of the actomyosin complex and muscle relaxation.

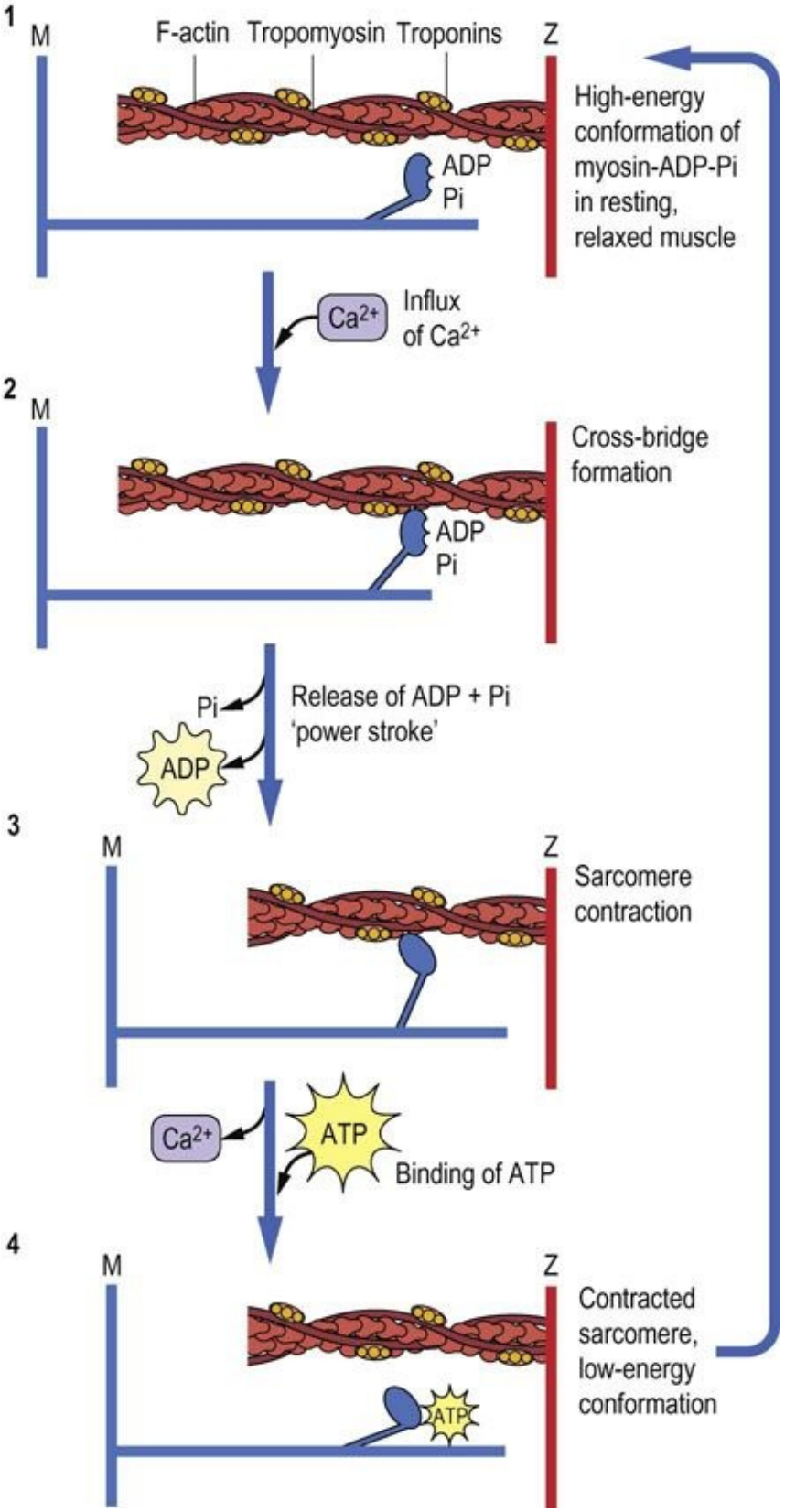

1
M
F-actin
Tropomyosin
Troponins
Z
High-energy conformation of myosin-ADP-Pi in resting, relaxed muscle
ADP
Pi
Ca2+
Influx of Ca2+
2
M
Cross-bridge formation
ADP
Pi
Pi
ADP
Release of ADP + Pi 'power stroke'
3
M
Z
Sarcomere contraction
Ca2+
ATP
Binding of ATP
4
M
Z
Contracted sarcomere, low-energy conformation
ATP

FIG. 20.4 Proposed stages in muscle contraction, according to the sliding filament model.

(1) In resting, relaxed muscle, calcium concentration is~10^{7} mol/L. The head group of myosin chains contains bound ADP and Pi, and is extended from the axis of the myosin helix in a high-energy conformation. Although the myosin-ADP-Pi complex has a high affinity for actin, binding of myosin to actin is inhibited by tropomyosin, which blocks the myosin-binding site on actin at low calcium concentration. (2) When muscle is stimulated, calcium enters the sarcoplasm through voltage-gated calcium channels (see Chapter 8). Calcium binding to Tn-C causes a conformational change in Tn-I, which is transmitted through Tn-T to tropomyosin. Movement of tropomyosin exposes the myosin-binding site on actin. Myosin-ADP-Pi binds to actin, forming a cross-bridge. (3) Release of Pi, then ADP, from myosin during the interaction with actin, is accompanied by a major conformational change in myosin, producing the 'power stroke', which moves the actin chain about 10 nm (100 Å) in the direction opposite the myosin chain, increasing their overlap and causing muscle contraction. (4) The uptake of calcium from the sarcoplasm and binding of ATP to myosin leads to dissociation of the actomyosin cross-bridge. The ATP is hydrolyzed, and the free energy of hydrolysis of ATP is conserved as the high-energy conformation of myosin, setting the stage for continued muscle contraction in response to the next surge in Ca^{2+} concentration in the sarcoplasm.

Higher myosin-ATPase activity increases cross-bridge cycling, which allows for increased rate of contraction. Different myosin isoforms have varying levels of ATPase activity, with fast muscles having higher rates of myosin-ATPase activity. Isoforms of actin and myosin are also found in the cytoskeleton of nonmuscle cells, where they have roles in diverse processes such as cell migration, vesicle transport during endocytosis and exocytosis, maintenance or changing of cell shape, and anchorage of intracellular proteins to the plasma membrane.

Excitation–contraction coupling: muscle membrane depolarization

T-tubules transmit electrochemical signals for efficient muscle contraction

Skeletal muscle contraction is initiated by neuronal stimulation at the neuromuscular endplate. As described previously (see Fig. 8.4), this stimulus leads to depolarization of the electrochemical gradient across the muscle plasma membrane (sarcolemma). The depolarization, caused by an influx of Na^{+}, propagates rapidly along the **sarcolemma membrane** and signals a voltage-gated calcium release from the **sarcoplasmic reticulum** (SR), a membrane-

bound, calcium-sequestering compartment inside the muscle cell. The influx of Ca^{2+} from the SR to the sarcoplasm initiates cross-bridge formation and excitation–contraction coupling (see Fig. 20.4). In striated muscle, depolarization is transmitted into the muscle fiber by invaginations of the plasma membrane, called **transverse tubules (T tubule)** (Fig. 20.5). The transmission of depolarization through the highly branched T-tubule network, which interacts closely with the SR, leads to rapid, concerted release of calcium from the SR into the sarcoplasm. In order for depolarization to occur again, sodium must be actively pumped out of the cytosol, by Na^+/K^+-ATPase pumps located in the sarcolemma. The rate of muscle repolarization is affected by both the rate and density of these pumps. Higher levels of Na^+/K^+-ATPase activity are found in fast contracting muscles, and increased Na^+/K^+-ATPase pump density is an important adaptation to exercise.

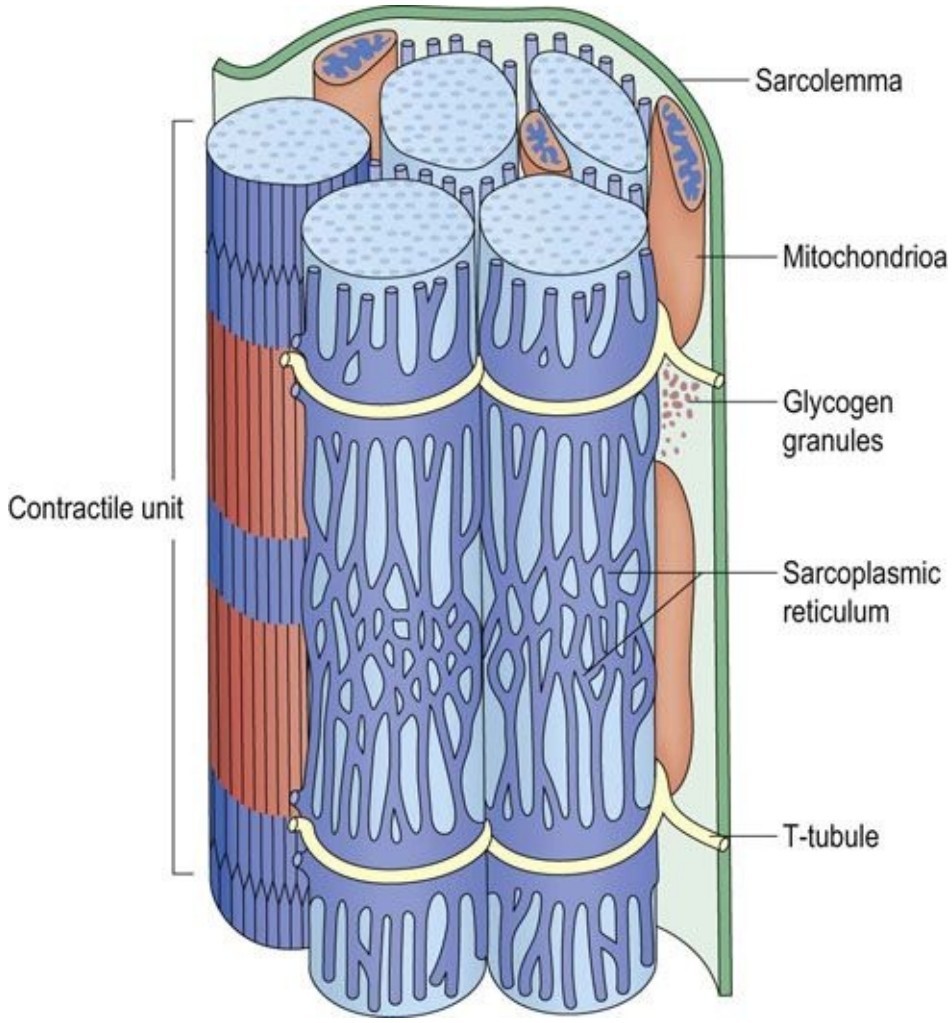


FIG. 20.5 Side view of the transverse tubular network in skeletal muscle cells. Transverse tubules are invaginations of the sarcolemma, which are connected to the sarcoplasmic reticulum (SR) by protein channels. The SR is a continuous, tubular compartment in close association with the myofibrils. The transverse tubules are extensions of the sarcolemma around the Z-line. They transmit the depolarizing nerve impulse to terminal regions of the SR, coordinating calcium release and contraction of the myofibril.

Skeletal, cardiac, and smooth muscle differ in their mechanism of neural stimulation, and have different structural adaptations for propagating the depolarization. Skeletal muscle contraction is volitional and fibers are innervated by motor nerve endplates that originate in the spinal cord; acetylcholine functions as the neurotransmitter (see Chapter 41). The **neuromuscular junction** is a special structural feature of skeletal muscle that is not found in cardiac or smooth muscle. Each individual fiber is innervated by only one motor nerve, and all the fibers innervated by one nerve are defined as a **motor unit**.

Motor unit control and synchronization is the basis for coordinated whole-muscle contraction. Skeletal muscle cramping is a nonvoluntary muscle contraction resulting from alterations in neuromuscular control and/or from electrolyte imbalances after excessive fluid loss typically during intense exercise in hot, humid conditions.

Cardiac muscle is striated and contracts rhythmically under involuntary control. The general mechanism of contraction of heart muscle is similar to that in skeletal muscle; however, the sarcoplasmic reticulum is less developed and the transverse tubule network is more developed in the heart. The heart is more dependent on, and actually requires, extracellular calcium for its contractile response (see Fig. 8.4); the influx of extracellular calcium enhances Ca^{2+} release from the SR. Lacking direct neural contact, cardiac myocytes propagate depolarization from a single node, the SA node, throughout the myocardium. The depolarization is passed cell to cell along specialized membrane structures called **intercalated disks**. Cardiac muscle is also more responsive to hormonal regulation. For example, cAMP-dependent protein kinases phosphorylate transport proteins and Tn-I, mediating changes in the force of contraction in response to epinephrine.

Smooth muscle can respond to both neural and circulating factors. Unlike skeletal muscle, neural input to smooth muscle innervates bundles of smooth muscle cells that cause both phasic (rhythmic) and tonic (sustained) contractions of the tissue. Smooth muscle can also be induced to depolarize by ligand–receptor interactions at the sarcolemma. This is called pharmacomechanical coupling, and is the basis for many drugs that target smooth muscle contraction or relaxation. **Nitric oxide donors**, such as amyl nitrite and nitroglycerin, used for treatment of angina, relax vascular smooth muscles, increasing the flow of blood to cardiac muscle.

Excitation–contraction coupling: the calcium trigger

The calcium content of the sarcoplasm is normally very low, 10^{-7} mol/L or less, but increases rapidly by over 100-fold in response to neural stimulation. The sarcoplasmic reticulum, a specialized organelle derived from the smooth endoplasmic reticulum, is rich in a Ca^{2+}-binding protein, **calsequestrin**, and serves as the site of calcium sequestration inside the cell. In striated muscle, T-tubule depolarization opens the Ca^{2+} channels in the SR (see Fig. 20.5). **The**

influx of Ca^{2+} into the sarcoplasm triggers both actin–myosin interactions and myosin-ATPase activity, leading to muscle contraction. Troponins are not expressed in smooth muscle. In this case, calcium triggers contraction by binding to calmodulin and activating myosin light chain kinase, which then phosphorylates myosin. Myosin phosphorylation enhances myosin–actin interaction.

Increased intracellular calcium activates more cross-bridges and causes sarcomere shortening through activation of myosin-ATPase. Thus, higher calcium levels increase muscle contractile force until saturation is reached. Calcium channel blockers used for treatment of **hypertension**, such as nifedipine, inhibit the flow of Ca^{2+} into the SR, thereby limiting the force of contraction of cardiac myocytes. While muscle contraction is triggered by increased calcium, muscle relaxation is dependent on calcium being actively pumped back into the SR. The rate of muscle relaxation is directly related to SR Ca^{2+}-ATPase activity. The SR is rich in Ca^{2+}-ATPase, which maintains cytosolic calcium in the sarcoplasm at submicromolar ($\sim 10^{-7}$ mol/L) concentrations. As intracellular calcium levels decrease, the number of active cross-bridges also decreases, and muscle contractile force declines.

Clinical box Malignant hyperthermia

About 1 in 150,000 patients treated with halothane (gaseous halocarbon) anesthesia or muscle relaxants responds with excessive skeletal muscle rigidity and severe hyperthermia with a rapid onset, up to 2°C (4°F) within 1 hour.

Unless treated rapidly, cardiac abnormalities may be life-threatening; mortality from this condition exceeds 10%. This genetic disease results from excessive or prolonged release of Ca^{2+} from the SR, most commonly the result of mutations in gene(s) that code for the Ca^{2+}-release channels within the SR. Excessive release of Ca^{2+} leads to a prolonged increase in sarcoplasmic Ca^{2+} concentration. Muscle rigidity results from Ca^{2+}-dependent consumption of ATP, and hyperthermia results from increased metabolism to replenish the ATP.

As muscle metabolism becomes anaerobic, lacticacidemia and acidosis may develop. The cardiac abnormalities result from

hyperkalemia, caused by release of potassium ions from muscle; as supplies of ATP are exhausted, muscle is unable to maintain ion gradients across its plasma membrane. Treatment of malignant hyperthermia includes use of muscle relaxants, *e.g.* dantrolene, an inhibitor of the ryanodine-sensitive Ca^{2+}-channel, to inhibit Ca^{2+} release from the SR. Supportive therapy involves cooling, administration of oxygen, correction of blood pH and electrolyte imbalances and also treatment of cardiac abnormalities.

Muscle energy metabolism

Energy resources in the muscle cell

Muscle is the primary site of **glucose disposal** (uptake from the circulation) in the body and is thus a natural target for treatment of the hyperglycemia of diabetes. The glucose transporter GLUT-4 is transported to the cell surface not only in response to insulin or pharmaceuticals but also in response to cellular energy status and by muscle contractions. Thus, exercise serves as a regulator of blood glucose levels and, in effect, a treatment for diabetes – exercise, in this context, is good medicine. The muscle content and activity of hexokinase also increases with exercise, both acutely (~3 hours after one session) and chronically (after several weeks of training). Studies in animal models have shown that exercise and drugs activate both similar and distinct biochemical signaling cascades in muscle; thus, the two treatments appear to be complementary.

ATP is used for muscle contraction

Three ATPases are required for muscle contraction: Na^+/K^+-ATPase, Ca^{2+}-ATPase and myosin-ATPase. Decreased ATP availability or inhibition of any of these ATPases will cause a decrease in muscle force production. However, the intracellular concentration of ATP does not change dramatically during exercise. Actively contracting muscle relies on the rapid resynthesis of ATP from ADP. Energy systems that synthesize ATP for muscle contraction include the creatine phosphate shuttle, anaerobic glycolysis from plasma glucose or glycogen, and aerobic metabolism of glucose and fatty acids via oxidative phosphorylation. The energy systems that synthesize ATP are not equivalent, and directly affect the amount and duration of power output from the contracting muscle.

Short-duration, high-power output contractions

Creatine phosphate is a high energy phosphate buffer used for rapid regeneration of ATP in muscle

A metabolic reality for skeletal muscle is that high force output can only be maintained for a short period of time. Contractions at or near maximal power

levels depend on high myosin-ATPase activity and rapid ATP resynthesis by substrate-level phosphorylation using the high-energy compound creatine phosphate (creatine-P). Creatine (see Table 9.2) is synthesized from arginine and glycine and is phosphorylated reversibly to creatine-P by the enzyme **creatine (phospho)kinase (CK or CPK)** (Fig. 20.6). CK is a dimeric protein and exists as three isozymes: the MM (skeletal muscle), BB (brain) and MB isoforms. The MB isoform is enriched in cardiac tissue.

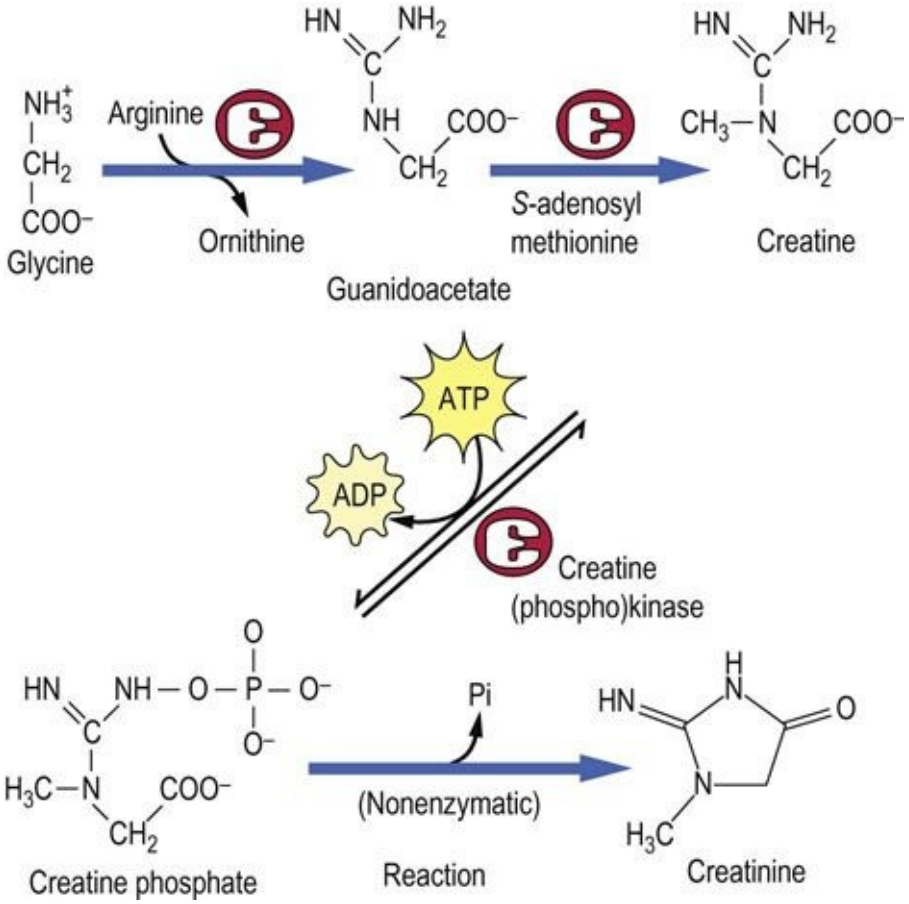


FIG. 20.6 Synthesis and degradation of creatine phosphate (creatine-P). Creatine is synthesized from glycine and arginine precursors. Creatine-P is unstable and undergoes slow, spontaneous degradation to Pi and creatinine, the cyclic anhydride form of creatine, which is excreted from the muscle cell into plasma and then into urine.

The level of creatine-P in resting muscle is several-fold higher than that of ATP (Table 20.3). Thus, ATP concentration remains relatively constant during the initial stages of exercise. It is replenished not only by the action of CK but

also by adenylate kinase (**myokinase**) as follows:

Table 20.3

Changes in energy resources in working muscle: concentrations of energy metabolites in human leg muscle during bicycle exercise

Metabolite	Metabolite concentration (mmol/kg dry weight)		
	Resting	3 minutes	8 minutes
ATP	27	26	19
Creatine-P	78	27	7
Creatine	37	88	115
Lactate	5	8	13
Glycogen	408	350	282

These experiments were conducted during ischemic exercise, which exacerbates the decline in ATP concentration. They illustrate the rapid decline in creatine-P and the increase in lactate from anaerobic glycolysis of muscle glycogen.

Data are adapted from Timmons JA, *et al:* J Clin Invest101:79–85, 1998.

$$\text{Creatine (Cr) phosphokinase}: \ \text{Cr-P} + \text{ADP} \rightarrow \text{Cr} + \text{ADP}$$

$$\text{Adenylate kinase}: \ 2\ \text{ADP} \rightleftharpoons \text{ATP} + \text{AMP}$$

Creatine phosphate stores decline rapidly during the first minute of high power output muscle contraction. As creatine phosphate stores are depleted, the muscle becomes unable to sustain the high force output, and contractile force rapidly declines. At this point, muscle glycogenolysis becomes a major source of energy. Calcium entry into muscle, in addition to its role in activation of myosin-ATPase-dependent contraction, also leads to formation of a Ca^{2+}–calmodulin

complex, which activates phosphorylase kinase, catalyzing the conversion of phosphorylase *b* to phosphorylase *a*. AMP also allosterically activates muscle phosphorylase and phosphofructokinase-1, accelerating glycolysis from muscle glycogen (see Chapter 13).

A further decline in force occurs as pyruvate and lactate gradually accumulate in the contracting muscle, resulting in a decrease in muscle pH. Force will then decline to a level that can be maintained by aerobic metabolism of fatty acids. Maximal aerobic power is about 20% of maximal power output, and about 50–60% of maximal aerobic power can be sustained for long periods of time.

Low-intensity, long-duration contractions

Fatty acids are the major source of energy in muscle during prolonged exercise

The availability and utilization of oxygen in working muscle are major limitations for maintaining continuous physical activity. Long-duration contractile activity requires adequate oxygen delivery and the capacity for the muscle to utilize the oxygen delivered. Oxygen delivery to muscle is affected by the red blood cell and hemoglobin concentrations in blood, the number of capillaries within the muscle and heart pump capacity. Highly oxidative muscle has a higher capillary density than glycolytic muscle, and muscle capillary density increases with endurance exercise training. Muscle oxygen utilization is also directly related to the number and size of muscle mitochondria. Muscles subjected to continual contractile activity, such as postural muscles, have more mitochondria than infrequently contracted muscle. A standard observation in muscle subjected to increased contractile demands is an elevation in oxidative enzyme activity.

At rest or at low intensities of physical work, oxygen is readily available and the aerobic oxidation of lipid predominates as the main source of ATP synthesis. However, at higher work intensities, oxygen availability for lipid catabolism can become limiting, and subsequently the muscle work rate decreases. During the first 15–30 minutes of exercise, there is a gradual shift from glycogenolysis and glycolysis to aerobic metabolism of fatty acids. Perhaps this is an evolutionary response to deal with the fact that lactate, produced by glycolysis, is more acidic and less diffusible than CO_2. As exercise continues, epinephrine contributes to activation of hepatic gluconeogenesis, providing an exogenous source of glucose

for muscle. Lipids gradually become the major source of energy in muscle during long-term, lower-intensity exercise when oxygen is not limiting.

Long-term muscle performance (stamina) depends on levels of muscle glycogen

Fats burn in the flame of carbohydrates; glycogen is required for efficient metabolism of lipids in muscle

Marathon runners typically 'hit the wall' when muscle glycogen reaches a critically low level. Glycogen is the storage form of glucose in skeletal muscle and its muscle concentration can be manipulated by diet, *e.g.* by **carbohydrate loading** prior to a marathon run. Fatigue, which can be defined as an inability to maintain the desired power output, occurs when the rate of ATP utilization exceeds its rate of synthesis. For efficient ATP synthesis, there is a continuing, but poorly understood, requirement for a basal level of glycogen metabolism, even when glucose is available from plasma and when fats are the primary source of muscle energy. Carbohydrate metabolism is important as a source of pyruvate, which is converted to oxaloacetate by the anaplerotic, pyruvate carboxylase reaction. Oxaloacetate is required to maintain the activity of the TCA cycle, for condensation with acetyl-CoA derived from fats. To some extent, muscle glycogen can be spared and performance time increased during long-term vigorous physical activity by increasing the availability of circulating glucose, either by gluconeogenesis or by carbohydrate ingestion, *e.g.* bread or Gatorade®. Increased utilization of fatty acids during early stages of exercise is an important training adaptation to regular vigorous physical activity – it serves to spare glycogen stores.

Muscle consists of two types of striated muscle cells: fast-glycolytic and slow-oxidative fibers

Striated muscle cells are generally classified by their physiologic contractile properties (fast versus slow) and primary type of metabolism (oxidative versus glycolytic). The muscle type is closely related to muscle function in skeletal muscle, and this comparison can easily be seen with muscles whose contraction is for infrequent-burst activities versus muscles used continuously for

maintaining posture (antigravity). The coloring of the two striated muscle types readily distinguishes them. **Fast-glycolytic muscle** used for burst activity is white in appearance (like chicken breast – chickens squawk a lot, but cannot fly far!) because of less blood flow, lower mitochondrial density and decreased myoglobin content compared with slow-twitch oxidative muscle, which is red. Fast-glycolytic fibers also have increased glycogen stores and lower fat content; they rely on glycogen and anaerobic glycolysis for short bursts of contraction when additional muscle force is required such as in the 'fight or flight' stress response. These muscle fibers are not capable of sustaining contraction for long periods. In contrast, **slow-oxidative fibers** in postural muscles (and in goose breast – geese are migratory birds) are well perfused with blood, rich in mitochondria and myoglobin. This muscle type has the ability to sustain low-intensity contractions for long periods. Slow muscle uses fatty acid oxidation for ATP synthesis, which requires mitochondria. Cardiac muscle, which is continuously contracting, has many contractile and metabolic characteristics that are similar to slow-oxidative skeletal muscle. Cardiac muscle is well perfused with blood, rich in mitochondria, and relies largely on oxidative metabolism of circulating fatty acids. Goose breast, which powers long, migratory flights, is a fairly fatty and dark meat, compared to chicken breast, and has many characteristics of cardiac muscle.

Clinical box Muscle wasting syndromes

Many patients with conditions that include HIV and many cancers experience severe body weight loss, a condition known as **cachexia**. Patients with cachexia are frequently unable to tolerate radiation or chemotherapy and have higher morbidity and mortality. The loss of body weight is often independent of caloric intake, and not just akin to starvation. Appetite stimulants alone are often not effective. The weight loss is associated with the loss of both muscle and adipose tissue. Health problems may be magnified in cachectic individuals due to the metabolic dysregulation that can accompany the loss of adipose and muscle tissue. There also appears to be an effect of fiber type on muscle loss, which may be related to metabolism. Fast-glycolytic muscle fibers undergo more protein loss than slow-oxidative muscle fibers. This preferential

loss of fast-glycolytic fibers with wasting is the opposite of what is seen in muscle with extended periods of disuse (disuse atrophy).

Slow-oxidative fibers atrophy preferentially during muscle disuse. Although the exact mechanisms inducing wasting are not certain, prime candidates with many wasting syndromes involve systemic inflammatory signaling by cytokines, such as TNF-α and IL-6. Inflammatory signaling induced by the disease process can activate muscle protein degradation, inhibit muscle protein synthesis, and induce adipose tissue lipolysis. Maintaining or preventing severe body weight loss in many disease states can improve patient treatment options, survival and quality of life. Anabolic agents, like testosterone, have proven beneficial in maintaining muscle mass in AIDS patients and are widely used clinically. With other wasting diseases, research in animal models has demonstrated that inhibition of inflammatory signaling can inhibit wasting. Further research is necessary before this approach is widely applied to human populations.

Clinical test box Assay of creatine to assess renal function and urine dilution

Since creatine phosphate concentration is relatively constant per unit muscle mass, the production of creatinine (see Fig. 20.6) is relatively constant during the day. Creatinine is eliminated in urine at a relatively constant amount per hour, primarily by glomerular filtration and to a lesser extent by tubular secretion. Since its concentration in urine varies with the dilution of the urine, levels of metabolites in random urine samples are often normalized to the urinary concentration of creatinine. Otherwise, a 24 h collection would be required to assess daily excretion of a metabolite. Normal creatinine concentration in plasma is about 20–80 mmol/L (0.23–0.90 mg/dL). Increases in plasma creatinine concentration

are commonly used as an indicator of renal failure. The albumin : creatinine ratio in a random urine sample, an indicator of protein filtration selectivity of the glomerulus, is used as a measure of the microalbuminuria to assess the progression of diabetic nephropathy (see also Chapter 24.).

Advanced concept box Sarcopenia

Sarcopenia, the loss of skeletal muscle mass, develops gradually in humans after the fifth decade of life and can lead to frailty and loss of functional capacity. Besides the basic erosion of quality of life, loss of skeletal muscle mass also increases the risk of mortality and morbidity. The cause of sarcopenia appears to be related to gradual decreases in physical activity and loss of regenerative capacity. Muscle fiber innervation by spinal motor neurons is critical to both development and maintenance of muscle fibers (cells). Spinal motor neurons decrease in number with advancing age, possibly because of cumulative oxidative damage to these postmitotic cells. The loss of motor neurons appears to cause the substantial (>40%) loss in the muscle fiber number, which is the primary determinant of **age-dependent sarcopenia**, and is accompanied by an increase in motor unit size, and decrease in fine motor skill. Sarcopenia has also been linked to age-induced systemic changes to the endocrine, cardiovascular, and immune systems, whose functions are all critical for the maintenance of skeletal muscle mass.

Comment.

The scientific evidence is clear that most older individuals can increase muscle strength and mass with a regular resistance exercise program. Pharmaceutical treatments have also been examined for individuals who cannot regularly exercise. Currently there is no treatment for spinal motor neuron loss. Pharmaceutical treatments targeting muscle have had varying degrees of success and are usually limited by side effects. The treatments include

endocrine interventions with male or female sex hormone replacement therapy, and growth hormone therapy. Anti-inflammatory medication is also employed to allow individuals to participate in physical activity programs. One of the best defenses against sarcopenia is regular exercise in order to maintain muscle mass during middle age.

Tissue engineering and replacement of muscle

As the field of tissue engineering advances, muscle tissue is at the forefront of experiments to grow an organ outside the human body. Muscle is derived from proliferating cells that originate in the mesenchyme germ layer in the developing embryo. These cells are ‘determined’ into the muscle lineage and then become myoblasts. **Myoblasts** exit the cell cycle and differentiate into a mature multinucleated muscle cell. Skeletal muscle cells (SkMC) are terminally differentiated, but skeletal muscle contains a small population (<5% of myonuclei) of undifferentiated muscle precursor cells, satellite cells. **Satellite cell** proliferation and differentiation are critical events for postnatal muscle growth and repair, *e.g.* in response to exercise, and for regeneration after damage. Skeletal muscle is one of the few human tissues that can largely regenerate itself after extensive injury. Muscle is also an ideal candidate for tissue ‘replacement’ following severe injury, as it quickly adapts to its mechanical and chemical environment. Skeletal muscles are highly specialized to their location and particular job. Muscle replacement surgeries using donor muscle for the hand have shown that morphological and biochemical differences must be taken into account when selecting muscle tissue for transplantation. Yet as muscle is highly adaptive (plastic fantastic), it is likely that skeletal muscle will be one of the first tissues (along with skin) to be completely engineered ex vivo for in vivo transplantation.

Terminally differentiated myoblasts in the heart are called cardiac myocytes; these cells remain single or binucleated throughout life. The heart has very limited regenerative capacity, so that the effects of myocardial infarction are long-lasting. Myoblasts of smooth muscle differentiate into their mature smooth muscle cells (SMC). But SMC, unlike heart and skeletal muscle, is not terminally differentiated. SMC phenotype also varies, based on the cell's location and function. SMCs are found throughout the body in the vascular wall, and retain the ability to proliferate, *e.g.* in response to hypertension or during angiogenesis.

Clinical test box Diagnosis of

myocardial infarction

Myocardial infarction (MI) is the result of blockage of blood flow to the heart. Tissue damage results in leakage of intracellular enzymes into blood (Fig. 20.7). Among these are glycolytic enzymes, such as LDH (Chapter 12); however, measurements of myoglobin, total plasma CK and CK-MB isozymes are more commonly used for diagnosis and management of MI. Myoglobin, a small protein (17 kDa), rises rapidly in plasma, within 2 hours following MI. Although it is sensitive, it lacks specificity for heart tissue. It is cleared rapidly by renal filtration and returns to normal within 1 day. Since plasma myoglobin also increases following skeletal muscle trauma, it would not be useful for diagnosis of MI, *e.g.* following an automobile accident. Total plasma CK and the CK-MB isozyme begin to rise within 3–10 hours following an MI, and reach a peak value of up to 25 times normal after 12–30 hours; they may remain elevated for 3–5 days. Total CK may also increase as a result of skeletal muscle damage but the measurement of CK-MB provides specificity for cardiac damage.

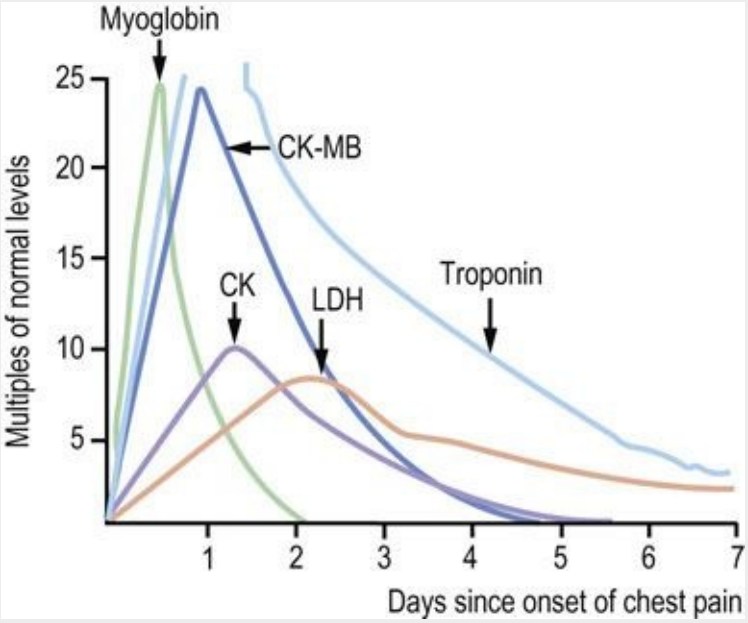


FIG. 20.7 Serum enzyme changes after myocardial infarction (MI). Various marker enzymes increase in plasma following MI. These are still used for the diagnosis of MI, but the currently recommended test is the

measurement of serum troponin concentration. CK, creatine phosphokinase; CK-MB, cardiac isozyme of CK; LDH, lactate dehydrogenase. Adapted from Pettigrew AR, Pacanis A: Diagnosis of myocardial infarction. In Dominiczak MH, editor: *Seminars in clinical biochemistry*, Glasgow, 1997, University of Glasgow Computer Publishing Unit.

Comment.

Enzyme-linked immunosorbent assays (ELISA) for the myocardial troponins are now recommended for use in diagnosis and management of MI. These assays depend on the presence of unique isoforms of troponin subunits in the adult heart. Plasma Tn-T concentration increases within a few hours after a heart attack, peaks at up to 300 times normal plasma concentration, and may remain elevated for 1–2 weeks. An assay for a specific isoform in an adult heart, Tn-T_2, is essentially 100% sensitive for diagnosis of MI and yields fewer than 5% false-positive results. Significant increases in plasma Tn-T_2 are detectable even in patients with unstable angina and transient episodes of ischemia in the heart. Troponins are commonly used as a component of an algorithm to differentiate high-risk from low-risk patients in terms of need for immediate invasive intervention. The recent definition of myocardial infarction is based on observed serum troponin concentrations.

Effect of exercise

Strength or resistance training increases muscle mass

A change in daily use of skeletal muscle has a profound effect upon its functional capacity. Both an increase and a decrease in daily activity level can change muscle structure, force production capacity, and fatigability. From a biochemical point of view, these changes are primarily caused by changes in tissue perfusion and metabolic enzymes, and thus the muscle's ability to take up glucose, utilize fat as an energy source, and generate ATP. The amount and intensity of daily physical activity occurs on a continuum, and muscle adaptation to this occurs in response to the specific stress placed upon it. For the sake of simplicity, and because this is how most research studies are designed, we can separate increased use (exercise training) into two categories: strength and aerobic training. The primary purpose of strength training, also called resistance training, is to increase the ability of a specific muscle, or group of muscles, to produce force. This is typically conducted through a small number of repetitions of one exercise movement against a resistance that only allows the muscle to contract through a full range of motion a very limited number of times (e.g. 6 to 8 repetitions of a bicep curl). The primary purpose of aerobic training, also called endurance training, is to increase endurance and decrease fatigue during prolonged, lower-intensity physical activity, *e.g.* running or walking. This is achieved through a high number of repetitions of muscle contractions at a low resistance. Each muscle contraction in strength training might be 75–90%, while in an aerobic training session it might be 15–20%, of the maximal voluntary force production of that muscle. The biochemical changes in response to these types of exercise are distinct.

Strength training has minimal effects on muscle biochemistry. The increase in force production capacity that occurs with strength training is due to increased cell size, *i.e.* hypertrophy. The hypertrophy of individual muscle cells occurs as a result of an increase in structural and sarcomeric proteins. With more myofibrils and sarcomeres (the contractile units of muscle) comes an increase in force production capability. When glycolytic enzymes are examined and normalized to the increased cell size, there is no change with strength training. When mitochondrial enzyme activity is normalized to the increased cell size of strength training, there is usually a slight decrease, suggesting that while force production

capacity increases, ATP production capacity (at least based on the size of the cell) has slightly decreased. In terms of contraction speed and sarcomere cross-bridge cycling, this is primarily determined by myosin-ATPase activity, which remains relatively unchanged in response to resistance training.

Endurance or aerobic training increases the oxidative metabolic capacity of muscle

In response to aerobic training, the primary biochemical change is an increase in capacity to metabolize fat, supported by increases in mitochondrial number, size and enzymes. All muscle fiber types (fast and slow) will increase their concentration and activity of citrate synthase and cytochrome-*c* by 2–3-fold, resulting in increased ATP production at a given workload (i.e. exercise intensity), so that muscle can then rely more on fat oxidation and less on anaerobic metabolism. The shift toward aerobic metabolism delays muscle fatigue; there are only minor effects on glycolytic enzymes in response to aerobic training, and the effects on cell size due to aerobic training are also minimal. Small shifts in myosin-ATPase composition may also occur, leading to a slower muscle phenotype (slower cross-bridge formation during contraction) due to aerobic training. Increases in glucose utilization as a result of increased expression of GLUT-4 and hexokinase also develop more in response to aerobic training, as opposed to strength training, but it is easy to see, considering the amount of skeletal muscle in the body, how blood glucose is decreased in a person with diabetes by an exercise program. It should also be noted that nearly all of these adaptations will occur in reverse in response to any form of de-training, whether that be due to cessation of an exercise program or bed rest due to injury or disease. Decreased use of muscle causes it to become much less metabolically efficient; unfortunately this de-adaptation becomes apparent within a few days following cessation of exercise. Other factors induced by endurance training include changes in cardiac output, increases in capillary density, and increases in glycogen stores. Of critical importance to health and medicine is the continuum within which these adaptations occur and the fact that small changes may impact many chronic diseases, including diabetes, atherosclerosis and cancer cachexia. Further, as changes occur relative to the original status of the muscle, older sedentary persons will see responses in muscle biochemistry comparable to those observed in younger persons. Thus, regardless of age, sedentary individuals who start even a moderate exercise program are likely to see substantial biochemical adaptations and health benefits.

Much research is still ongoing in these areas in an attempt to understand the molecular genetic and signaling pathways that bring about these responses and understanding how they might be modified after injury or disease.

Summary

■ Muscle is the major consumer of fuels and ATP in the body. Glycogenolysis, blood glucose, glycolysis and lipid metabolism are essential for optimal muscle activity. Reliance on these energy-producing pathways varies with muscle type and its prior contractile activity.
■ Skeletal, cardiac, and smooth muscle have a common actomyosin contractile complex, but differ in innervation, contractile protein arrangement, calcium regulation of contraction, and propagation of depolarization from cell to cell.
■ The sarcomere is the fundamental contraction unit of striated muscle and is defined by Z-lines and thick and thin filament overlap.
■ Contraction is described by a 'sliding filament' model in which hydrolysis of ATP is catalyzed by an influx of Ca^{2+} into the sarcoplasm and is coupled to changes in the conformation of myosin. Relaxation of the high-energy conformation of myosin during interaction with actin produces a 'power stroke', resulting in increased overlap of the actin–myosin filaments and shortening of the sarcomere.
■ The ATP produced in muscle drives the maintenance of ion gradients, restoration of intracellular calcium levels, and the contractile process.
■ Fast-glycolytic muscle relies largely on glycogen and anaerobic glycolysis for short, high-intensity bursts of muscle activity.
■ Slow-oxidative muscle is an aerobic tissue; at rest, it uses fats as its primary source of energy. During the initial phases of exercise, it relies on glycogenolysis and glycolysis, but then gradually shifts to fat metabolism for long-term energy production. Enzymes and proteins are released from muscle in response to damage.
■ Measurements of plasma CK-MB activity and troponin concentration are used as biomarkers of damage to cardiac muscle and are commonly used in the diagnosis and treatment of myocardial infarction.
■ Exercise is good medicine; it increases insulin sensitivity and glucose disposal, and assists in maintenance of muscle mass and function during aging.

Active learning

1. When chickens are frightened, they squawk a lot, may jump high and fly for short distances, but are unable to take flight and fly for great distances, either normally or to escape danger. In contrast,

geese have the ability to fly for great distances during semi-annual migrations. Compare the types of muscle fibers and energy sources in the breast of chicken and geese and explain how the differences in fiber type are compatible with the flying capacity of these birds.
2. Discuss the impact of muscle glycogen phosphorylase deficiency (McArdle's disease) and carnitine or carnitine palmitoyl transferase I deficiency on muscle performance during short-and long-duration exercise.
3. Review the merits of blood doping, carbohydrate loading and creatine supplementation to enhance performance during marathon events.

Further reading

Cooke, R. The sliding filament model: 1972–2004. *J Gen Physiol*. 2004; 123:643–656.

Mayer, F, Scharhag-Rosenberger, F, Carlsohn, A, *et al.* The intensity and effects of strength training in the elderly. *Dtsch Arztebl Int*. 2011; 108:359–364.

Mitchell, WK, Williams, J, Atherton, P, *et al.* Sarcopenia, dynapenia, and the impact of advancing age on human skeletal muscle size and strength; a quantitative review. *Front Physiol*. 2012; 3:260.

Muscaritoli, M, Lucia, S, Molfino, A, *et al.* Muscle atrophy in aging and chronic diseases: is it sarcopenia or cachexia? *Intern Emerg Med*. 2013; 8:553–560.

Shave, R, Baggish, A, George, K, *et al.* Exercise-induced cardiac troponin elevation: evidence, mechanisms, and implications. *J Am Coll Cardiol*. 2010; 56:169–176.

Tiwari, RP, Jain, A, Khan, Z, *et al.* Cardiac troponins I and T: molecular markers for early diagnosis, prognosis, and accurate triaging of patients with acute myocardial infarction. *Mol Diagn Ther*. 2012; 16:371–381.

Websites

Muscular dystrophies. www.muscular-dystrophy.org/conditions.

Animations: numerous excellent videos of muscle structure and contraction a. www.youtube.com

CHAPTER 21

Glucose Homeostasis and Fuel Metabolism

Diabetes Mellitus

Marek H. Dominiczak

Learning objectives

After reading this chapter you should be able to:

- Characterize the main energy substrates (metabolic fuels).
- Outline the actions of insulin and glucagon.
- Compare and contrast metabolism in the fasting and postprandial states.
- Describe the metabolic response to injury and compare it with metabolism in diabetes.
- Characterize type 1 and type 2 diabetes mellitus.
- Explain the basis of laboratory tests relevant to fuel metabolism and monitoring of diabetes.

Introduction

Continuous provision of energy is essential to maintain life. This chapter describes metabolism of compounds known as energy substrates (or metabolic fuels). It also discusses the most common metabolic disease, diabetes mellitus.

Glucose and fatty acids are the most important energy substrates: both can be stored in the body

The most important energy substrates in mammals are glucose and fatty acids. After ingestion of food, their excess is stored to be released again in case of need. This storage-release paradigm safeguards the energy supply between meals, and in extreme circumstances can ensure an organism's survival for weeks and months. The main pathways of fuel metabolism and the key metabolites are listed in Table 21.1.

Table 21.1

Principal anabolic and catabolic pathways, and their main substrates and products

Pathway	Main substrates	End products
Anabolic		
Gluconeogenesis	Lactate, alanine, glycerol	Glucose
Glycogen synthesis	Glucose-1-phosphate	Glycogen
Protein synthesis	Amino acids	Proteins
Fatty acid synthesis	Acetyl-CoA	Fatty acids
Lipogenesis	Glycerol, fatty acids	Triacylglycerols (triglycerides)
Catabolic		
Glycolysis	Glucose	Pyruvate, ATP
Tricarboxylic acid cycle	Pyruvate	NADH + H^+, $FADH_2$
		CO_2, H_2O, ATP
Glycogenolysis	Glycogen	Glucose-1-phosphate, glucose
Pentose phosphate	Glucose-6-phosphate	NADPH + H^+,
pathway		pentoses, CO_2
Fatty acid oxidation	Fatty acids	Acetyl-CoA CO_2, H_2O, ATP (ketone bodies)
Lipolysis	Triglycerides	Glycerol, fatty acids
Proteolysis	Proteins	Amino acids, glucose

Note that metabolites such as pyruvate and acetyl-CoA are common to several pathways. Note also which pathways generate reducing equivalents (NADH, NADPH and $FADH_2$), which in turn are substrates for the mitochondrial respiratory chain.

Metabolism is geared towards safeguarding continuous

glucose supply. Glucose is stored as glycogen and can also be synthesized from noncarbohydrate compounds

Continuous supply of glucose is essential for survival because in normal circumstances glucose is the only fuel which the brain can use. Glucose is also the preferred fuel for muscle during initial stages of exercise. In spite of the fact that from time to time there is large demand for glucose, the amount of glucose present in the extracellular fluid is only about 20 g (1 oz), the equivalent of 80 kcal (335 kJ). To safeguard supply, glucose can be released into the circulation from the 'emergency store' of its polymer, **glycogen**. Glycogen is stored in the liver (approximately 75 g; 2.5 oz) and in muscle (400 g; 1 lb). This is equivalent to about 1900 kcal (7955 kJ) and can supply glucose for approximately 16 h of fasting (compare Table 19.1 and 20.3). When the period of fast is longer, another mechanism of glucose supply comes into play: its synthesis de novo from noncarbohydrate compounds known as **gluconeogenesis**.

Fatty acids are stored as esters of glycerol (triacylglycerols or triglycerides)

The body has virtually unlimited capacity for the accumulation of fat in the adipose tissue, and it is stored as esters of glycerol (triacylglycerols). The caloric value of fat (9 kcal/g, 37 kJ/g) is higher than that of either carbohydrates (4 kcal/g, 17 kJ/g) or proteins (4 kcal/g) (Table 21.2). A 70 kg (154 lb) man will store approximately 15 kg (33 lb) of fat. This is equivalent to over 130,000 kcal (544,300 kJ), a vast amount compared to the caloric value of the stored glycogen. Fatty acids support body energy needs during prolonged periods of fasting, and during prolonged exercise. In extreme circumstances, people can fast for as long as 60–90 days: obese persons may survive for over a year without food.

Table 21.2

Reciprocal effects of insulin and glucagon on the key enzymes of gluconeogenesis

Enzyme	Effect of glucagon	Effect of insulin
Glucose-6-phosphatase (Glc-6-Pase)	Induction	Repression
Fructose-1,6-biphosphatase (Fru-1,6-BPase)	Induction	Repression
Phosphoenolpyruvate carboxykinase (PEPCK)	Induction	Repression

On a high-carbohydrate diet, insulin induces gene transcription of the glycolytic enzymes

glucokinase, phosphofructokinase (PFK), pyruvate kinase (PK), and glycogen synthase. At the same time, it represses the key enzymes of gluconeogenesis, pyruvate carboxylase (PC), PEPCK, Fru-1,6-BPase, and Glc-6-Pase. Glucagon effects oppose those of insulin. **On a high-fat diet,** glucagon represses the synthesis of glucokinase, PFK-1, and PK, and induces the transcription of PEPCK, Fru-6-Pase, and Glc-6-Pase.

Amino acids can be used as a fuel after conversion to glucose

Amino acids normally serve as substrates for synthesis of body's proteins. However, in certain situations they become energy substrates. During a prolonged fast or periods of metabolic stress induced by illness or injury, body proteins are degraded, and the released amino acids are converted into glucose in the course of gluconeogenesis. When an excessive amount of amino acids is taken as food, they are converted to carbohydrates and are either stored or metabolized (Chapter 19).

Different organs and tissues handle fuels differently

At rest, **the brain** uses approximately 20% of all oxygen consumed by the body. Glucose is normally its only fuel: during starvation, however, the brain adapts to the use of ketone bodies as an alternative energy source.

The two pathways that provide glucose are glycogenolysis and gluconeogenesis. When glucose concentration in the extracellular fluid decreases, it is first replenished by degrading liver glycogen. However, when the fasting period extends, gluconeogenesis is initiated. Gluconeogenesis takes place mostly in the **liver**, and the **kidneys** also contribute during prolonged fast. Its main substrates are lactate (from anaerobic glycolysis), alanine (from the amino acids released during breakdown of muscle protein) and glycerol (from the breakdown of triacylglycerols in the adipose tissue (Chapter 13)).

Muscle uses both glucose and fatty acids as energy sources. During short-term exercise, glucose is the preferred substrate, however, at rest and during prolonged exercise, fatty acids are the main energy source (Chapter 20). The myocyte cannot release glucose directly into the circulation because it does not contain the enzyme glucose-6-phosphatase. Thus, it uses glycogen only for its own energy needs. However, it does contribute to gluconeogenesis by releasing lactate, which is transported to the liver. No gluconeogenesis takes place in muscle.

Glucose concentration in **plasma** reflects the balance between, on the one

hand, its intake (absorption from the gut) or its endogenous production (glycogenolysis and gluconeogenesis), and on the other, its tissue utilization in glycolysis, pentose phosphate pathway, tricarboxylic acid (TCA) cycle and in glycogen synthesis (see Fig. 13.2). In the fasting state, a person weighing 70 kg (154 lb) metabolizes glucose at a rate of approximately 2 mg/kg/min (200 g/24 h).

Insulin and glucagon are the main hormones that control glucose disposal, its plasma concentration, and, generally, the state of fuel metabolism

Glucose homeostasis is controlled, on the one hand, by the anabolic hormone insulin and on the other, by the catabolic hormones (glucagon, catecholamines, cortisol, and growth hormone), also known as counterregulatory hormones (Fig. 21.1). Insulin and glucagon are both secreted from the pancreatic islets of Langerhans. Insulin is secreted by β-cells (approximately 70% of all islet cells) and glucagon by the α-cells. The molar ratio of insulin to glucagon at any given time is the key determinant of fuel metabolism.

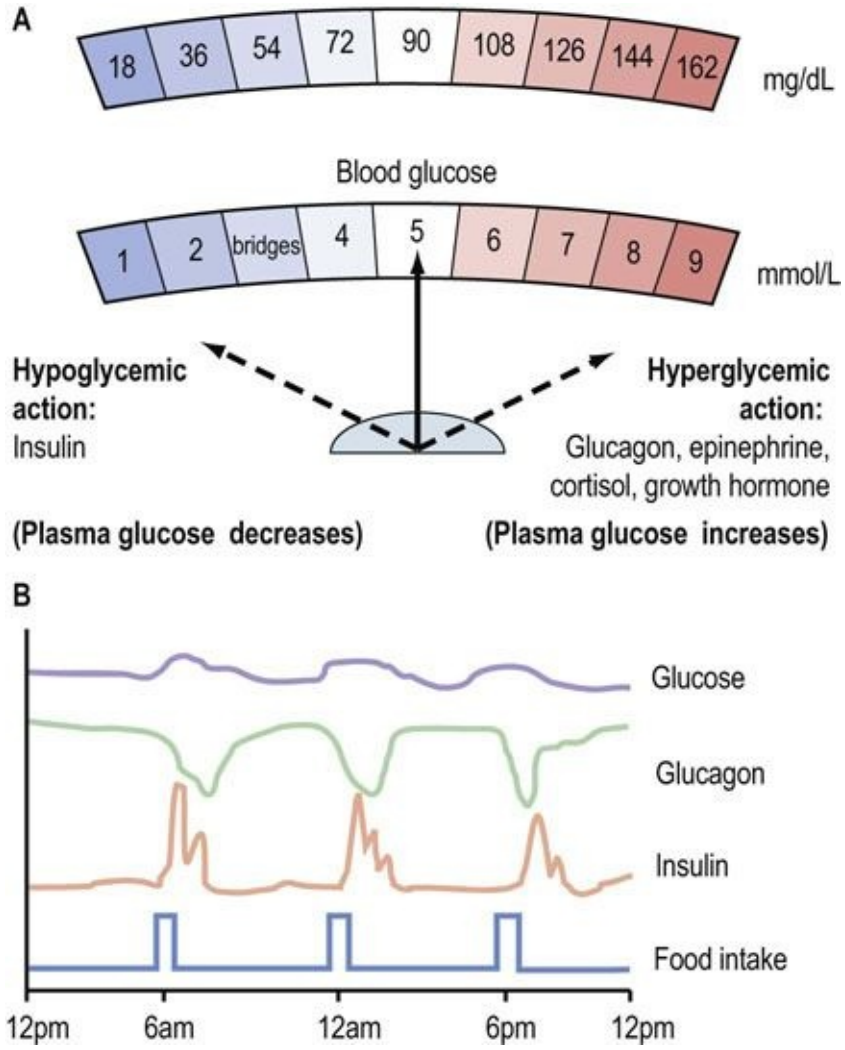


FIG. 21.1 Hormonal control of glucose homeostasis.
(A) Plasma glucose concentration reflects the balance between the hypoglycemic (glucose-lowering) action of insulin and the hyperglycemic (glucose-increasing) action of the anti-insulin hormones. **(B)** Daily patterns of insulin and glucagon secretion, and corresponding plasma glucose concentrations. Plasma glucose concentration is maintained within a narrow range throughout the day. To obtain glucose concentrations in mg/dL, multiply the value in mmol/L by 18.

Insulin

Insulin was discovered in 1921–22 by Frederick Banting, Charles Best and John Macleod, all working in Toronto (see Further Reading). Much later, in 1979, insulin became the first recombinant human protein to be produced commercially. The insulin molecule consists of two peptide chains linked by two disulfide bonds. Its molecular weight is 5500 Da. The α-chain contains 21 amino acids, and the β-chain 30 amino acids. Insulin is synthesized in the rough endoplasmic reticulum of the pancreatic β-cells and is packaged into the secretory vesicles in the Golgi apparatus. The precursor of insulin is the single chain molecule, preproinsulin. First, a 24-amino acid signal sequence is cleaved from preproinsulin by a peptidase, yielding proinsulin. Proinsulin is then split by endopeptidases into insulin and C-peptide (Fig. 21.2), both of which are released from the cell in equimolar amounts. This is exploited in the clinical laboratories for the assessment of the β-cell function in patients treated with insulin. In such persons, endogenous insulin cannot be measured directly, because the administered insulin would interfere in the assay. However, since C-peptide is present in the same molar concentration as native insulin, its serves as a marker of β-cell function.

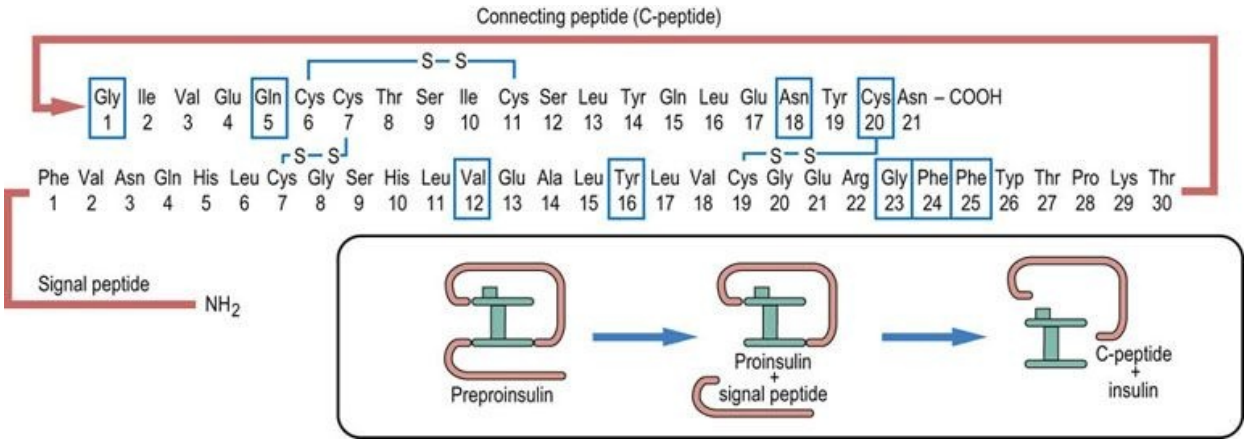


FIG. 21.2 Insulin.
The insulin molecule consists of two polypeptide chains joined by two disulfide bridges. The third bridge is internal to the β-chain. Insulin is synthesized as a longer peptide, preproinsulin, which is cleaved into the signal peptide and proinsulin. Before being secreted from the β-cell, proinsulin is split further into the C-peptide and insulin. Boxes drawn around amino acid residues indicate the amino acid involved in binding to the

insulin receptor.

Glucose metabolism in the β-cell determines insulin secretion

When glucose is being absorbed during a meal, the β-cell takes it up using the membrane transporter GLUT-2 (Chapter 8 and Table 8.2). On entering the cell, glucose is phosphorylated by glucokinase. As glucose metabolism is stimulated, the ATP/ADP ratio in the cell increases. This closes the ATP-sensitive potassium channels in the cell membrane, decreasing potassium efflux. The cell is depolarized, and the L-type calcium channels open. Calcium ions enter the cell (Chapter 40) and activate the Ca^{2+}-dependent proteins that cause the release of secretory granules containing preformed insulin. This is known as the first phase of insulin secretion (Fig. 21.3) (compare this with the neurosecretory granules, Chapter 41.1). Loss of this phase is the early functional sign of islet cell damage. The second phase of insulin secretion involves synthesis of new insulin and responds to signals such as an increase in the concentration of the cytosolic long-chain acetyl-CoA.

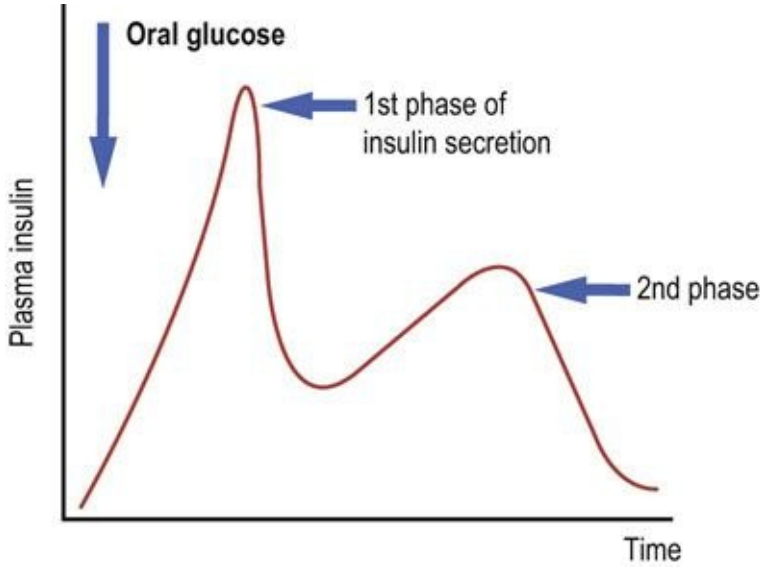


FIG. 21.3 Insulin secretion.
Note the two phases of insulin secretion. Glucose is the most important stimulator of insulin secretion. Other stimulators are some of the amino acids (branched-chain amino acids), stimulation of the vagus nerve and the hormones secreted by the gut (incretins).

The process of ‘matching’ insulin secretion to the sensed plasma glucose concentration by the β-cell is known as the **metabolism-secretion coupling**.

Insulin action

Insulin signaling within cells involves a membrane receptor and the phosphorylation cascades

Cellular propagation of the insulin signal includes multiple protein phosphorylations (Fig. 21.4; compare Fig. 13.7). They occur at different sites within the target protein molecules: tyrosine phosphorylation sends signals differently from serine-threonine phosphorylation. The phosphorylations help create conformational changes that enable binding (docking) of other proteins and thus further propagation of the signal.

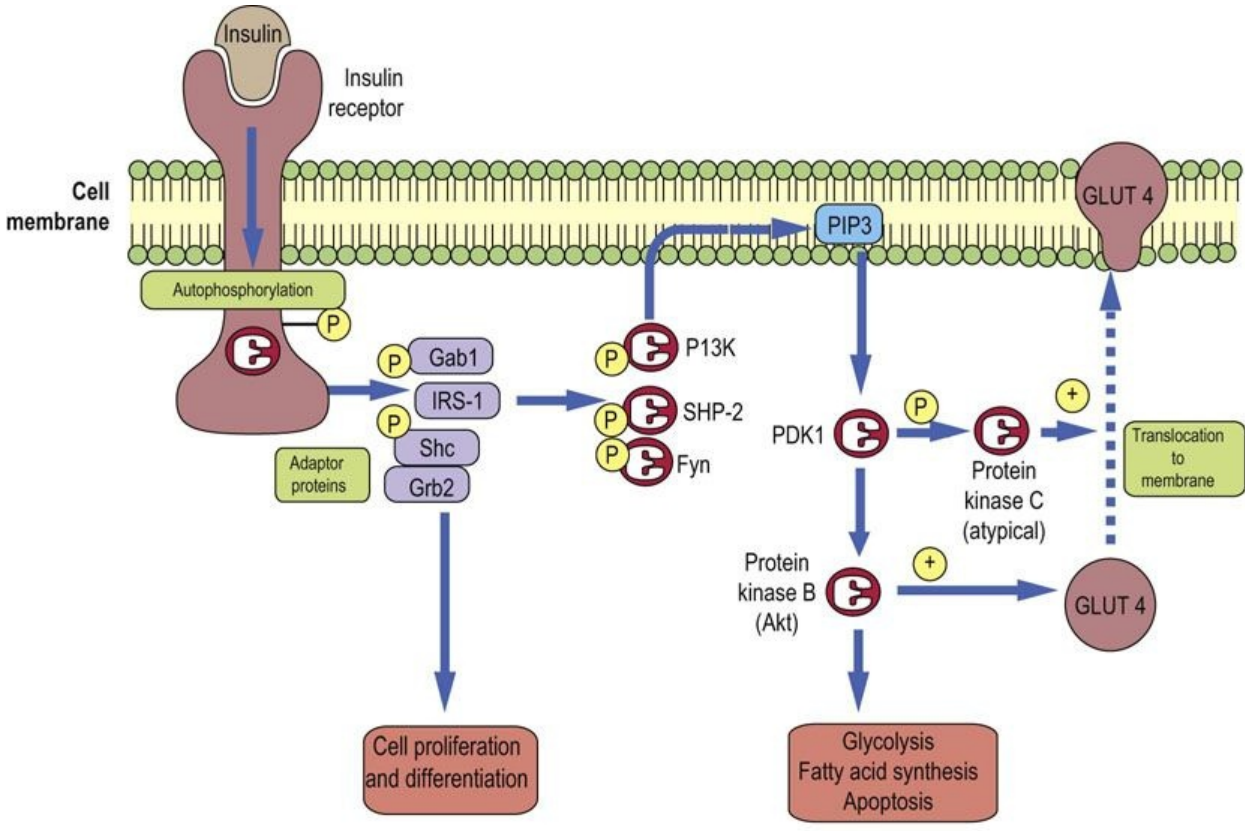


FIG. 21.4 Insulin signaling.
Insulin signaling cascades transfer the signal from the insulin molecule to its target

molecules such as regulatory enzymes or the membrane glucose transporter GLUT-4. One signaling pathway involves insulin receptor phosphorylation, phosphorylation of the IRS-1, the phosphoinositol-3-kinase (PI3K), then PDKs and two downstream kinases PKB/Akt and PKC. GLUT4 translocation can also be affected by the PI3K-independent pathway (not shown here). Compare Fig. 13.7. IRS-1, insulin receptor substrate 1; PDK, phosphoinositide-dependent kinase.

First, insulin binds to its receptor, which is a four-subunit protein spanning the cell membrane of a target cell. One subunit of the receptor contains an ATP-binding site and has tyrosine kinase activity. Binding of insulin causes the receptor to phosphorylate itself. The receptor also recruits and phosphorylates several other proteins such as the insulin receptor substrates (designated IRS-1–4), and adaptor proteins Gab1, Shc, Grb2, and others. These in turn dock another set of proteins such as the phosphatidylinositol 3-kinase (PI3K), protein tyrosine phosphatase (SHP2) and Fyn kinase.

As the signal spreads, PI3K generates lipid-based messengers, the phosphoinositol phosphates, the most important being phosphatidylinositol -3,4,5-trisphosphate (PIP_3; Chapter 40). They in turn recruit the serine-threonine kinase (**protein kinase B, PKB, also called Akt**) and the 3′-phosphoinositide-dependent kinases 1 and 2 (PDK1 and 2). PDK1 phosphorylates PKB and protein kinase C (PKC). A form of PKC known as PKC λε is involved in glucose transport and translocation in the adipocytes. Serine-threonine kinases can also phosphorylate IRS-1; this decreases its susceptibility to phosphorylation by the insulin receptor, introducing another regulatory loop. Phosphatases such as the phosphotyrosine phosphatase 1B contribute to the termination of insulin signal.

There is also a **PI3K-independent pathway** of insulin signaling that also activates glucose transport; the insulin receptor phosphorylates the protein called Cbl, using associated protein substrate (APS) as an adaptor. Phosphorylated Cbl binds to Cbl-associated protein (CAP). CAP in turn binds to flotillin, a protein associated with lipid rafts present in the membrane. Flotillin docks guanyl nucleotide exchange factor (C3G) with the mediation of another adaptor molecule, CrkII. This activates the G-protein called TC-10, which participates in the translocation of the **GLUT-4 transporter** in adipocytes to the cell membrane, thus affecting the transport of glucose.

Metabolic effects of insulin

Insulin promotes anabolism: the storage of carbohydrates and lipids, and protein

synthesis. In parallel, it suppresses the catabolic pathways. Insulin acts on three main target tissues: the liver, adipose tissue and skeletal muscle (Fig. 21.5). The liver is the main target of insulin action in the fasting state. After a meal, the main targets are muscle and adipose tissue.

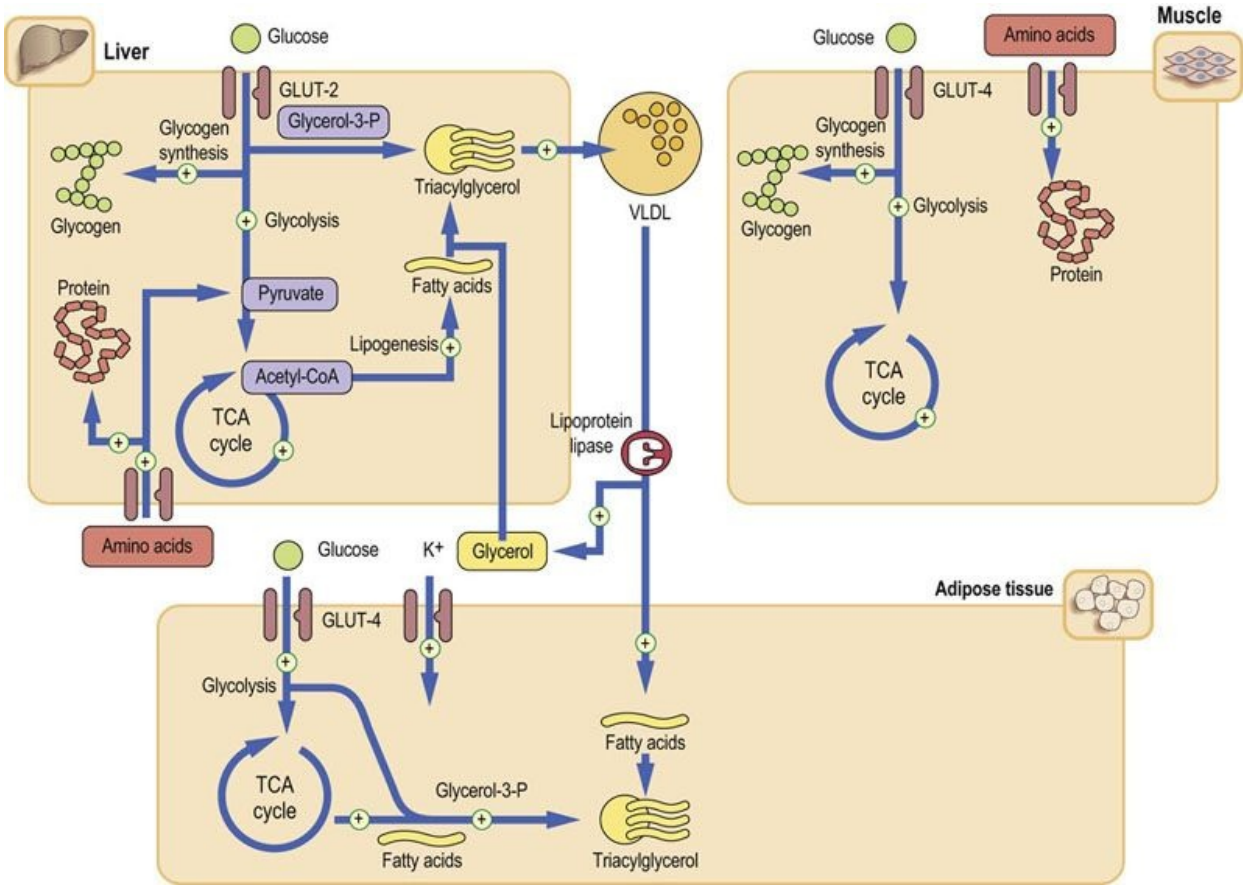


FIG. 21.5 Metabolic effects of insulin.
The principal insulin target tissues are liver, muscle, and adipose tissue. Insulin affects carbohydrate, lipid, and protein metabolism. Insulin, also promotes cellular potassium uptake. The plus sign indicates pathways stimulated by insulin. In most instances insulin also inhibits the opposite processes. Glucose transport in muscle and adipose tissue is mediated by the GLUT-4 transporter and is insulin dependent. On the other hand, the glucose transporter in the liver (GLUT-2) is insulin-independent (compare Fig. 13.7).

In the liver, insulin stimulates glycolysis and glycogen synthesis. At the same time, it suppresses lipolysis and gluconeogenesis. It stimulates synthesis of long-chain fatty acids and lipogenesis – the synthesis of triacylglycerols. It promotes the assembly of the very low-density lipoproteins (VLDL), which transport lipid

to the peripheral cells.

In the adipose tissue, insulin stimulates triglyceride synthesis from glycerol-3-phosphate and fatty acids.

In muscle, it stimulates glucose transport, glucose metabolism, and glycogen synthesis. It increases cellular uptake of amino acids and stimulates protein synthesis.

Insulin also induces the endothelial lipoprotein lipase, an enzyme that liberates triacylglycerols from the chylomicrons and VLDL (Chapter 18).

Insulin stimulates glucose transport across cell membrane

Insulin-dependent glucose entry into cells is mediated by proteins known as glucose transporters (Table 8.2). The GLUT-4 transporter controls glucose uptake in skeletal muscle and adipocytes. GLUT-4 cycles between the endosomes and membrane. In an unstimulated cell, most of the GLUT-4 molecules reside intracellularly and no more than 10% are present in the plasma membrane. In humans, insulin doubles the GLUT-4 recruitment to cell membrane. Importantly, muscular contraction increases expression of the GLUT-4 independently of insulin. On the other hand, fatty acids decrease the expression of the GLUT-4 in muscle.

Insulin resistance is a condition in which a given dose of insulin produces a less than expected response in the cell

Insulin resistance affects all the main insulin-responsive tissues. In the liver this results in an increased VLDL production and a procoagulant status (increased fibrinogen synthesis and increased plasminogen activator inhibitor 1, PAI-1), in muscle it causes a decreased glucose uptake, and in adipose tissue the overproduction of free fatty acids and changes in the secretion of adipokines (a decrease in adiponectin and an increase in resistin production).

Advanced concept box There are several methods of assessment of insulin resistance

Insulin resistance can be assessed by function tests, which are used mostly for research purposes. One such test is the **hyperinsulinemic euglycemic clamp:** the insulin is infused at a

constant rate together with variable amounts of glucose. The rate of glucose infusion is adjusted to keep plasma glucose concentration at 5–5.5 mmol/L (90–99 mg/dL). When a steady state is attained, the rate of glucose infusion is equal to the peripheral glucose uptake and is a measure of insulin sensitivity/resistance. Other tests of insulin resistance are based on the oral glucose tolerance test (see below) and involve frequent measurements of plasma glucose and insulin concentrations after glucose load.

The most important cause of insulin resistance is defective insulin signaling

Within a cell, insulin resistance may be caused by defects at several levels (Table 21.3). It could be a compromised insulin binding to its receptor, for instance due to a very rare mutation in the insulin receptor gene. Resistance can also be caused by antireceptor autoantibodies. The most important causes, however, are defects in the insulin signaling pathways. In type 2 diabetes and in persons with a strong family history of diabetes, the IRS-phosphoinositol kinase pathway may not operate normally. This impairs cellular translocation of GLUT-4, and consequently the glucose uptake. Reduced expression and impaired translocation of GLUT-4 in adipocytes (but not in the skeletal muscle) have been observed both in obesity and in diabetes. Interestingly, adipocytes of GLUT-4-knockout mice show overexpression of a gene for another protein, the retinol-binding protein (RBP). Although RBP is synthesized in adipocytes, it was found to impair insulin signaling in muscle; thus, it serves as an inter-tissue messenger contributing to insulin resistance.

Table 21.3

Sites of insulin resistance

Site of resistance	Possible defect	Comment
Pre-receptor	Insulin receptor antibodies, abnormal molecule	Rare
Receptor	Decreased number or affinity of insulin receptors	Not significant in diabetes
Post-receptor	Defects in signal transduction: defective tyrosine phosphylation, mutations in genes coding for IRS-1, phosphatidylinositol-3'-kinase, defective translocation of GLUT-4 to cell membrane, elevated concentration of fatty acids	Postreceptor resistance is the most common type of insulin resistance

Other causes of insulin resistance include variants of genes coding for IRS-1 and phosphatidylinositol-3-kinase. Peripheral insulin resistance, particularly in muscle, is induced by the presence of excess fatty acids. Fatty acids can also exert a direct harmful effect on the β-cell. The elevated fatty acids concentrations to insulin resistance by inhibiting peripheral glucose disposal, enhancing hepatic glucose output and damaging β-cell function. Steatosis (accumulation of triacylglycerols in liver and muscle), seen in patients with high plasma triglyceride concentrations, also contributes to insulin resistance. These untoward phenomena stimulated by fatty acid excess are known as **lipotoxicity**.

Metabolic effects of glucagon and other anti-insulin hormones

Anti-insulin hormones antagonize the effects of insulin, promoting hyperglycemia

The anti-insulin hormones mobilize glucose by stimulating glycogenolysis and

gluconeogenesis. This leads to an increase in plasma glucose concentration (hyperglycemia).

Glucagon acts on the liver

Glucagon is a single-chain, 29-amino acid peptide, with a molecular weight of 3485 Da. It acts on the liver. It mobilizes fuel reserves to maintain blood glucose concentration between meals. It stimulates glycogenolysis, gluconeogenesis (Table 21.2), the oxidation of the fatty acids and ketogenesis, and inhibits glycolysis, glycogen synthesis and the synthesis of triacylglycerols (Fig. 21.6).

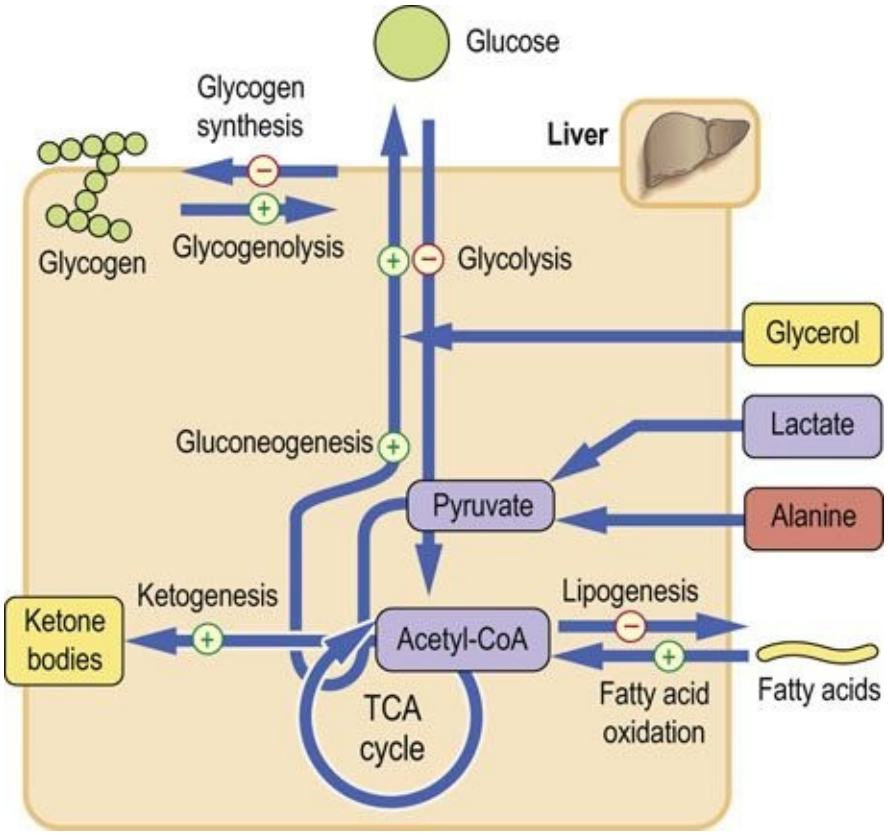


FIG. 21.6 Metabolic effects of glucagon.
Glucagon mobilizes glucose from every available source. It also increases lipolysis, and ketogenesis from acetyl-CoA. Glucagon actions are confined to the liver (compare Fig. 13.4).

Glucagon binds to its own membrane receptor (Chapter 13, Fig. 13.4), which signals through the membrane-associated G-proteins and the cyclic AMP cascade. First, the glucagon–receptor complex causes binding of guanosine 5′-

triphosphate (GTP) to a G-protein complex (Chapter 40). This leads to the dissociation of G-protein subunits. One of these subunits (Gα) activates the adenylate cyclase, which in turn converts ATP into cyclic AMP (cAMP). cAMP in turn activates cAMP-dependent protein kinase which, through phosphorylation of several regulatory enzymes, controls the key steps in carbohydrate and lipid metabolism (Figs 21.6 and 21.7).

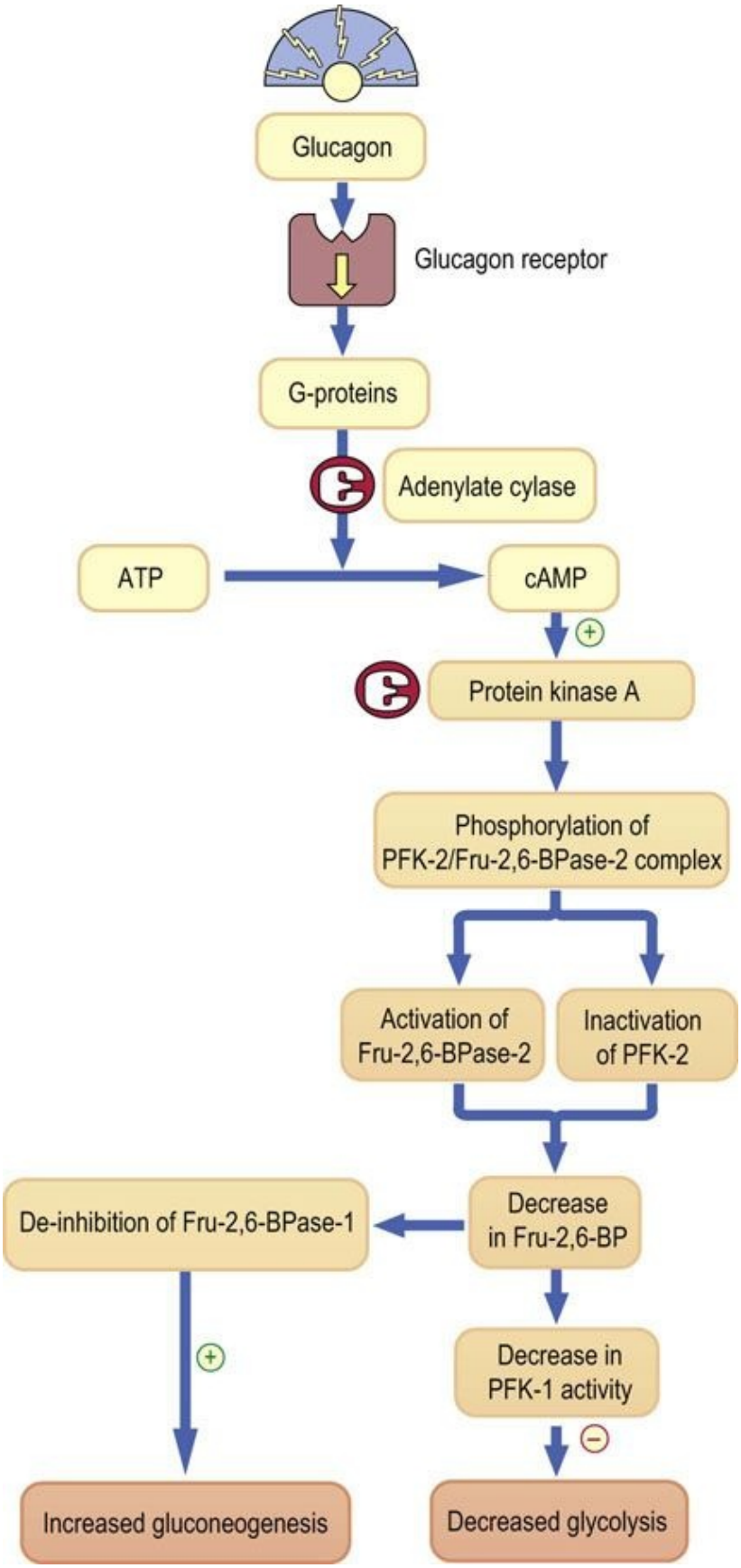

Glucagon
Glucagon receptor
G-proteins
Adenylate cylase
ATP
cAMP
Protein kinase A
Phosphorylation of
PFK-2/Fru-2,6-BPase-2 complex
Activation of
Fru-2,6-BPase-2
Inactivation
of PFK-2
De-inhibition of Fru-2,6-BPase-1
Decrease
in Fru-2,6-BP
Decrease in
PFK-1 activity
Increased gluconeogenesis
Decreased glycolysis

FIG. 21.7 Regulation of glycolysis and gluconeogenesis by phosphofructokinase. Glucagon regulates gluconeogenesis by controlling the bifunctional enzyme with activities of both **phosphofructokinase-2** (PFK-2) and **fructose 2,6-biphosphatase-2** (Fru-2,6-BPase-2). Glucagon binds to its membrane receptor and signals through G-proteins, and adenylate cyclase, generating cyclic AMP. cAMP in turn activates protein kinase A. Subsequently, this kinase phosphorylates the PFK-2:Fru-2,6-BPase complex. Phosphorylation activates the bisphosphatase, which degrades Fru-2,6-BP, and lowering of Fru-2,6-BP disinhibits another enzyme, Fru-2,6-BPase-1, in the main pathway of gluconeogenesis. Thus, **gluconeogenesis is stimulated**. Ingeniously, the decrease in Fru-2,6-BP has a reciprocal inhibitory effect on the key glycolytic enzyme, phosphofructokinase (PFK-1). Thus, **glycolysis is inhibited**.

There are no glucagon receptors on muscle cells; muscle glycogenolysis is stimulated by another anti-insulin hormone, epinephrine.

Epinephrine acts on liver and muscle

Epinephrine (adrenaline) is a key hormone responsible for hyperglycemia in response to stress. Its metabolic effects are similar to these of glucagon. It inhibits glycolysis and lipogenesis, and stimulates gluconeogenesis. Its receptors are the α-and β-adrenergic receptors (mainly the β_2-receptor; Fig. 13.5) and they link to the cAMP signaling cascade.

Hormones regulate key enzymes in the metabolic pathways by several different mechanisms

The rate and direction of metabolic pathways depend on the activity of key (regulatory) enzymes. Hormones target these regulatory enzymes. Regulation can be exerted through several mechanisms: induction or repression of genes that code for these enzymes, phosphorylation and dephosphorylation of enzyme proteins, and other mechanisms such as substrate interactions and allosteric changes. Cell energy level and the redox state of the cell also control their activity (Chapter 6).

Transcription factors directly control gene induction

Induction or repression of genes (Table 21.2) takes days or weeks to take effect, and may occur in response to diet, chronic stress or illness. For example, activities of several hepatic enzymes are different in people on a high-fat diet compared to those on a high-carbohydrate diet.

Gene induction is controlled by transcription factors, proteins which bind to

the response elements in gene promoters (Fig. 35.2). Transcription factors are often targets of the hormonal signaling cascades.

Some nuclear receptors such as the peroxisome proliferator-activated receptors (PPARs) act as transcription factors in cooperation with other nuclear receptors. PPARs are important targets of antidiabetic drugs (Fig. 21.9).

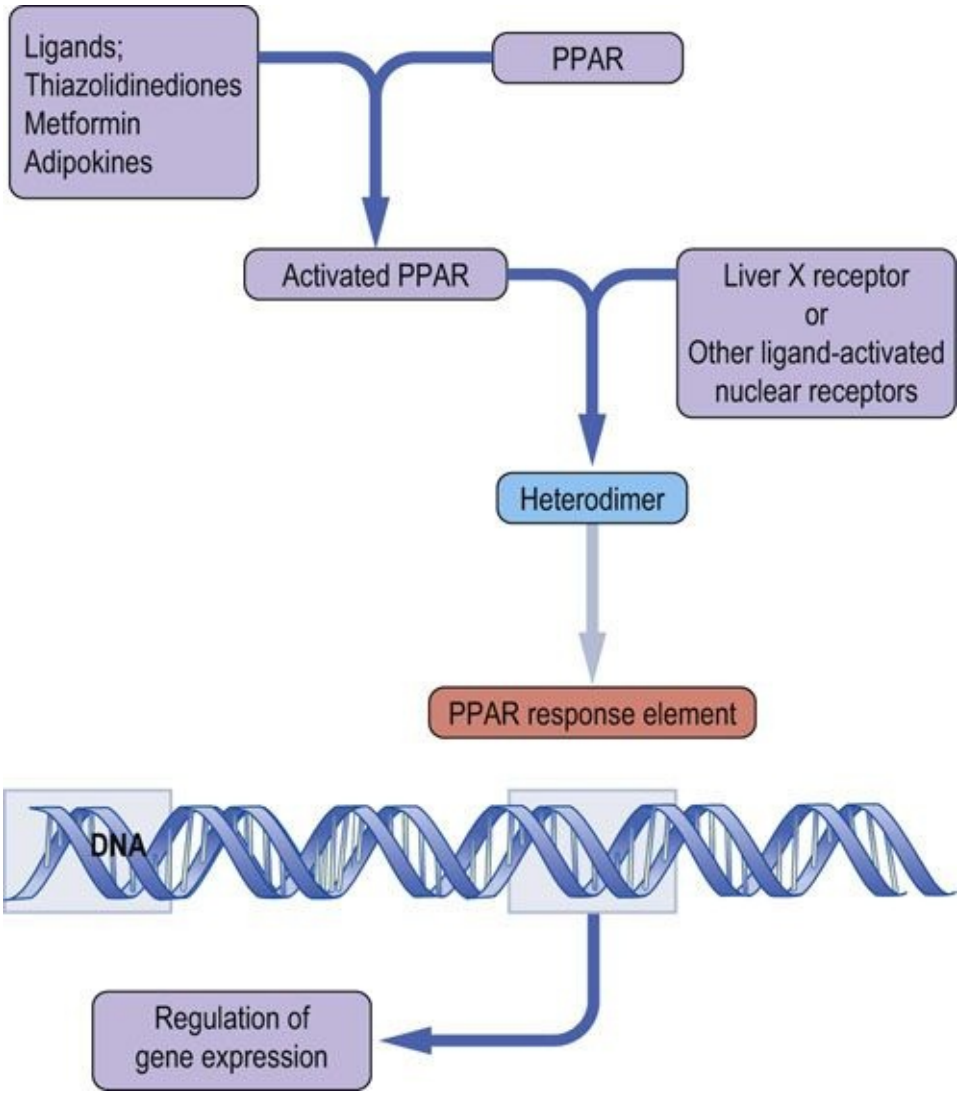


FIG. 21.9 Transcriptional regulation by peroxisome proliferator-activated receptors (PPARs).
PPARs are activated by ligands that may be metabolites or drugs. They form heterodimer complexes with a range of other nuclear receptors. Resulting complexes bind to the PPAR response elements located in gene promoters, regulating gene expression.

Phosphorylation is a key regulatory mechanism in metabolic pathways and signaling cascades

Phosphorylation and dephosphorylation is a fast-acting regulatory mechanism. It commonly controls the key enzymes in the metabolic pathways. It is further fine-tuned by involving several levels of kinases and phosphatases. Phosphorylation/dephosphorylation can also regulate transcription factors. Figure 21.8 illustrates how this mechanism operates in glycogenolysis, gluconeogenesis and lipolysis.

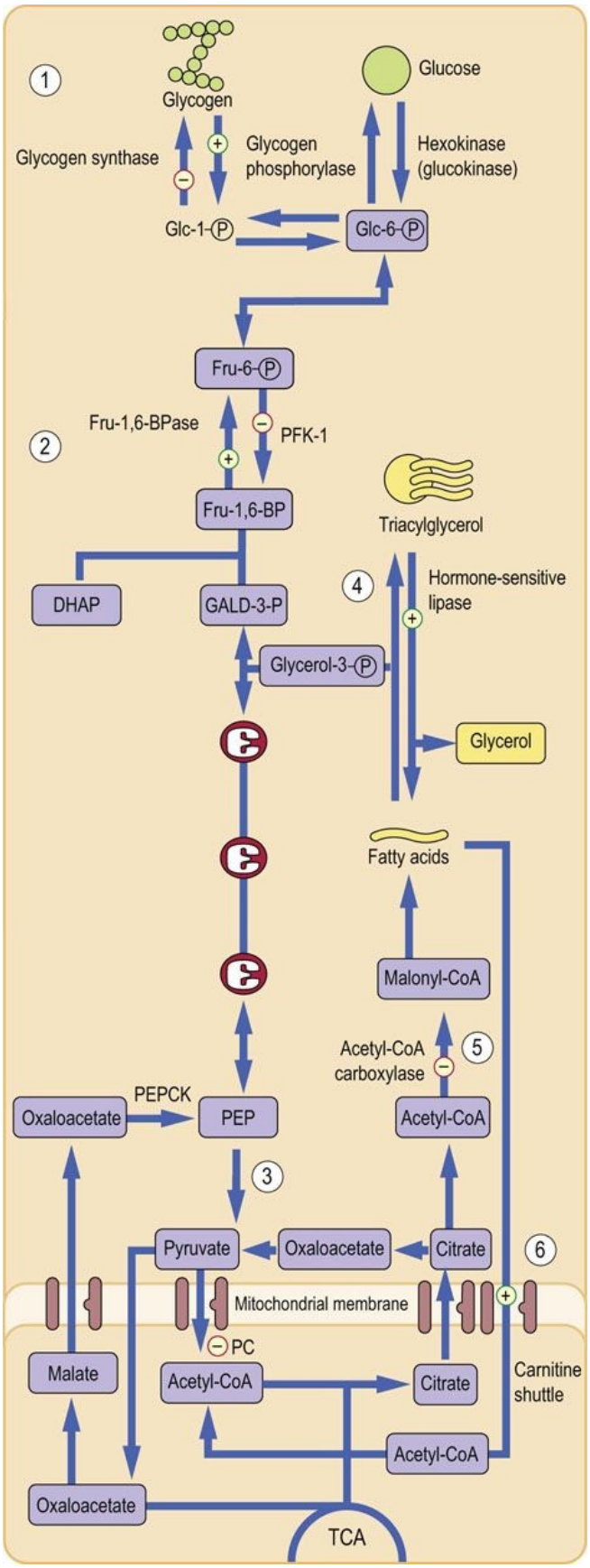

1
Glycogen
Glucose
Glycogen synthase
Glycogen phosphorylase
Hexokinase (glucokinase)
Glc-1-Ⓟ
Glc-6-Ⓟ
Fru-6-Ⓟ
Fru-1,6-BPase
PFK-1
2
Fru-1,6-BP
Triacylglycerol
4
Hormone-sensitive lipase
DHAP
GALD-3-P
Glycerol-3-Ⓟ
Glycerol
Fatty acids
Malonyl-CoA
Acetyl-CoA carboxylase
5
PEPCK
Oxaloacetate
PEP
Acetyl-CoA
3
Pyruvate
Oxaloacetate
Citrate
6
Mitochondrial membrane
PC
Malate
Acetyl-CoA
Citrate
Carnitine shuttle
Acetyl-CoA
Oxaloacetate
TCA

FIG. 21.8 Regulatory role of enzyme phosphorylation.
Phosphorylation of the key enzymes, controlled by glucagon and epinephrine, is the major mechanism of regulation of carbohydrate and lipid metabolism. Phosphorylation usually stimulates enzymes in the catabolic pathways and inhibits enzymes in the anabolic ones. **Regulation of glycogen metabolism:** glycogen phosphorylase is activated and glycogen synthase is inactivated. This promotes glycogen breakdown (1). **Regulation of gluconeogenesis:** Fru-2,6-BPase-2 is activated and PFK-2 is inhibited. This decreases Fru-2,6-BP formation and in turn inhibits PFK-1 (glycolysis) and stimulates FBPase (gluconeogenesis) (2). **Regulation of glycolysis:** normally, Fru-1,6-BP also allosterically activates pyruvate kinase downstream in the glycolytic pathway. Because its formation is decreased, glycolysis slows down (3). **Regulation of lipolysis:** phosphorylation stimulates hormone-sensitive lipase-stimulating lipolysis (the release of fatty acids from triglycerides) (4). **Regulation of fatty acid oxidation:** phosphorylation inhibits acetyl-CoA carboxylase, inhibiting the generation of malonyl-CoA (5). Malonyl-CoA normally inhibits carnitine-palmitoyl transferase-1. Lack of malonyl-CoA disinhibits it (6), facilitating the entry of fatty acids into mitochondria. This stimulates lipid oxidation. DHAP, dihydroxyacetone phosphate; GALD-3-P, glyceraldehyde-3-phosphate; PEP, phosphoenolpyruvate.

Insulin and glucagon switch genes on and off during feed–fast cycle

Insulin regulates synthesis of key enzymes by controlling the activity of so-called forkhead transcription factors (they have a helix-turn-helix structure with two additional loops or wings). Two such transcription factors, forkhead box protein 1 (FOXO1) and HNF-3B (FOXA2), are essential for switching metabolism from anabolism to catabolism. FOXO1 promotes gluconeogenesis in the liver in the fasted state, and FOXA2 regulates fatty acid oxidation. They are both inactivated by phosphorylation by kinases in the IRS-1/PKB/Akt pathway.FOXO1 and its coactivators PGC-1 and TORC2 (the latter being cAMP responsive) stimulate gluconeogenesis by activating genes that code for the rate-limiting enzymes phosphoenolpyruvate carboxykinase (PEPCK) and Glc-6-Pase.

During feeding, insulin, again through the IRS-PKB/Akt pathway, phosphorylates and inactivates FOXO1. This inhibits hepatic gluconeogenesis. Insulin also deactivates TORC2. Glucagon, on the other hand, dephosphorylates TORC2, activating it. This leads to induction of the gluconeogenic enzymes.

FOXA2 regulates fat breakdown in the fasted state by inducing genes encoding enzymes of glycolysis, fatty acid oxidation, and ketogenesis. This leads to an increase in the concentrations of free fatty acids, ketone bodies and triacylglycerols, and to a decrease in liver triacylglycerol content. Insulin

phosphorylates and inhibits FOXA2 through the IRS-PKB/Akt pathway (Fig. 21.4).

The incretin hormones

Hormones secreted in the gut potentiate insulin secretion

Gastrointestinal hormones, such as **glucagon-like peptide-1** (GLP-1) and the **glucose-dependent insulinotropic peptide** (also known as gastric inhibitory peptide; GIP) and also **cholecystokinin**, and **vasoactive intestinal peptide** (VIP), potentiate insulin secretion. They are secreted following ingestion of foods (being stimulated by the increase of glucose concentration in the intestinal lumen). This is known as the **incretin effect** and explains why insulin response to orally administered glucose is greater than to its intravenous infusion. Insulin secretion is also stimulated by amino acids, such as leucine, arginine, and lysine.

GLP-1 is secreted in the intestinal mucosa by the L-cells, mostly in the distal ileum and colon. It occurs in two forms, GLP-1 (7–37) and GLP-1 (7–36) amide, the latter present in a higher concentration. GLP-1 secretion increases rapidly after a meal. In the presence of elevated glucose concentration, GLP-1 increases insulin secretion by the pancreatic β-cells. It also decreases glucagon secretion and thus decreases endogenous glucose production. This effect diminishes with decreasing glucose concentration. GLP-1 also decreases gastric emptying and increases the feeling of satiety. It also stimulates proliferation of the β-cells. GIP is a 42 amino acid molecule synthesized in the K-cells in the duodenum and jejunum.

GLP-1 and GIP act through G-protein-coupled receptors. The GLP-1 receptor is present in the β-and α-cells in the pancreatic islets, and also in the peripheral tissues. In addition to increasing insulin secretion they also stimulate β-cell proliferation and decrease apoptosis, acting through the PKA-mediated pathway. GLP-1 and GIP are inactivated by the **dipeptidyl peptidase-4** (DPP-4).

The feed–fast cycle

Human metabolism oscillates between the fed state and the fasting state. The ‘switch’ which determines metabolic changes is the **molar ratio of insulin to glucagon in plasma**. The **fed state** (also called the absorptive or postprandial state) occurs during a meal and for several hours after and is characterized by high insulin and low glucagon concentration (a high insulin/glucagon ratio). **Fasting,** on the other hand, is a low-insulin/high-glucagon state. Fasting for 6–12 h is called the **postabsorptive state**. Fasting that lasts longer than 12 h is ‘prolonged fasting’ or starvation.

Metabolism in the fed state

A meal stimulates insulin release and inhibits glucagon secretion

Nutrients present in a meal stimulate insulin release and suppress the secretion of glucagon. This affects metabolism in the liver, adipose tissue, and muscle (Fig. 21.10). Glucose utilization by the brain remains unchanged. There is an increase in glucose uptake in the insulin-dependent tissues, principally in the skeletal muscle. Glucose oxidation and glycogen synthesis are stimulated, and lipid oxidation is inhibited. Glucose taken up by the liver is phosphorylated by glucokinase, yielding glucose-6-phosphate (Glc-6-P). Excess glucose is directed into the pentose phosphate pathway, generating NADPH + H^+ that is used in biosyntheses that require reductions, such as synthesis of fatty acids and cholesterol.

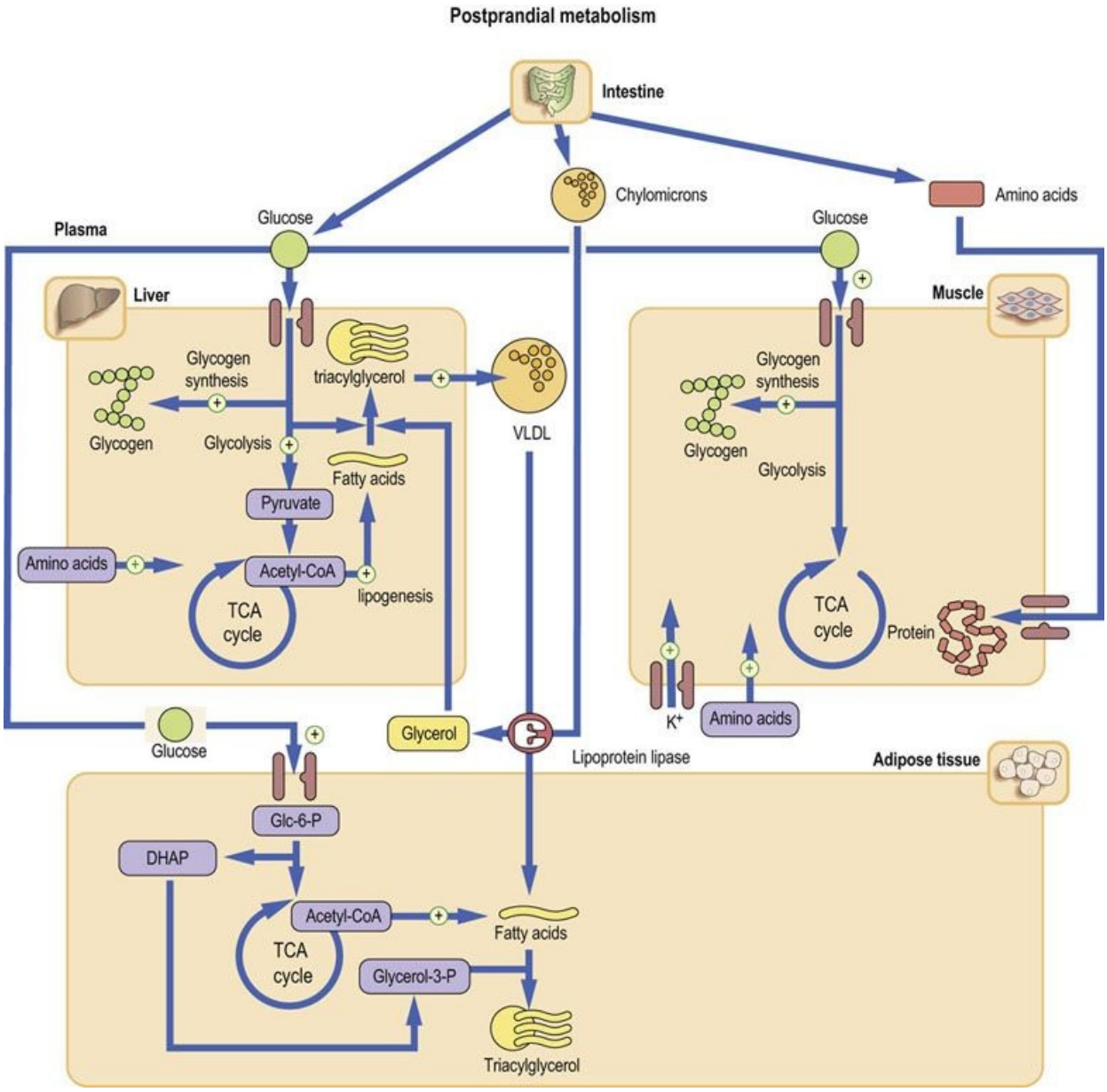


FIG. 21.10 Metabolism in the fed (postprandial) state.
Carbohydrates, amino acids and fats are absorbed in the intestine, and insulin secretion is stimulated. Insulin directs metabolism towards storage and synthesis (anabolism). In the liver, glucose is taken up by the GLUT-2 transporter and is channeled into glycolysis and glycogen synthesis. Aerobic glycolysis supplies acetyl-CoA, which is a key substrate for fatty acid synthesis. The fatty acids are subsequently esterified by glycolysis-derived glycerol, forming triacylglycerols in a process known as lipogenesis. Triacylglycerols are packaged into VLDL for transport to peripheral tissues. In muscle, glycogen synthesis, amino acid uptake and protein synthesis are stimulated. In the adipose tissue, VLDL triacylglycerols are hydrolyzed and fatty acids are taken up by cells. Triacylglycerols are resynthesized intracellularly, becoming adipocyte storage material. DHAP, dihydroxyacetone phosphate; Glc-6-P, glucose-6-phosphate.

Fat (triacylglycerols) absorbed in the intestine is transported in chylomicrons to peripheral tissues, where it is hydrolyzed to glycerol and free fatty acids by lipoprotein lipase (Chapter 18). The released fatty acids are taken up by cells. In muscle they are used as fuel. In the adipose tissue they are reassembled into triacylglycerols and stored. The reassembly of triacylglycerols requires glycerol, which is provided from glycolysis (triose phosphate being reduced to glycerol-3-phosphate). Insulin stimulates synthesis of fatty acids in the liver and adipose tissue. It also stimulates amino acid uptake and protein synthesis, and decreases protein degradation in liver, muscle, and adipose tissue.

Metabolism in the fasting state

During fasting the liver switches from a glucose-utilizing to a glucose-producing organ

During fasting, insulin secretion decreases and glucagon secretion increases (Fig. 21.11). This leads to a decrease in glycogen synthesis and an increase in glycogenolysis. The liver is transformed from a glucose-utilizing to a glucose-producing organ. After an overnight fast, a steady state is attained, with hepatic glucose production becoming equal to its peripheral uptake.

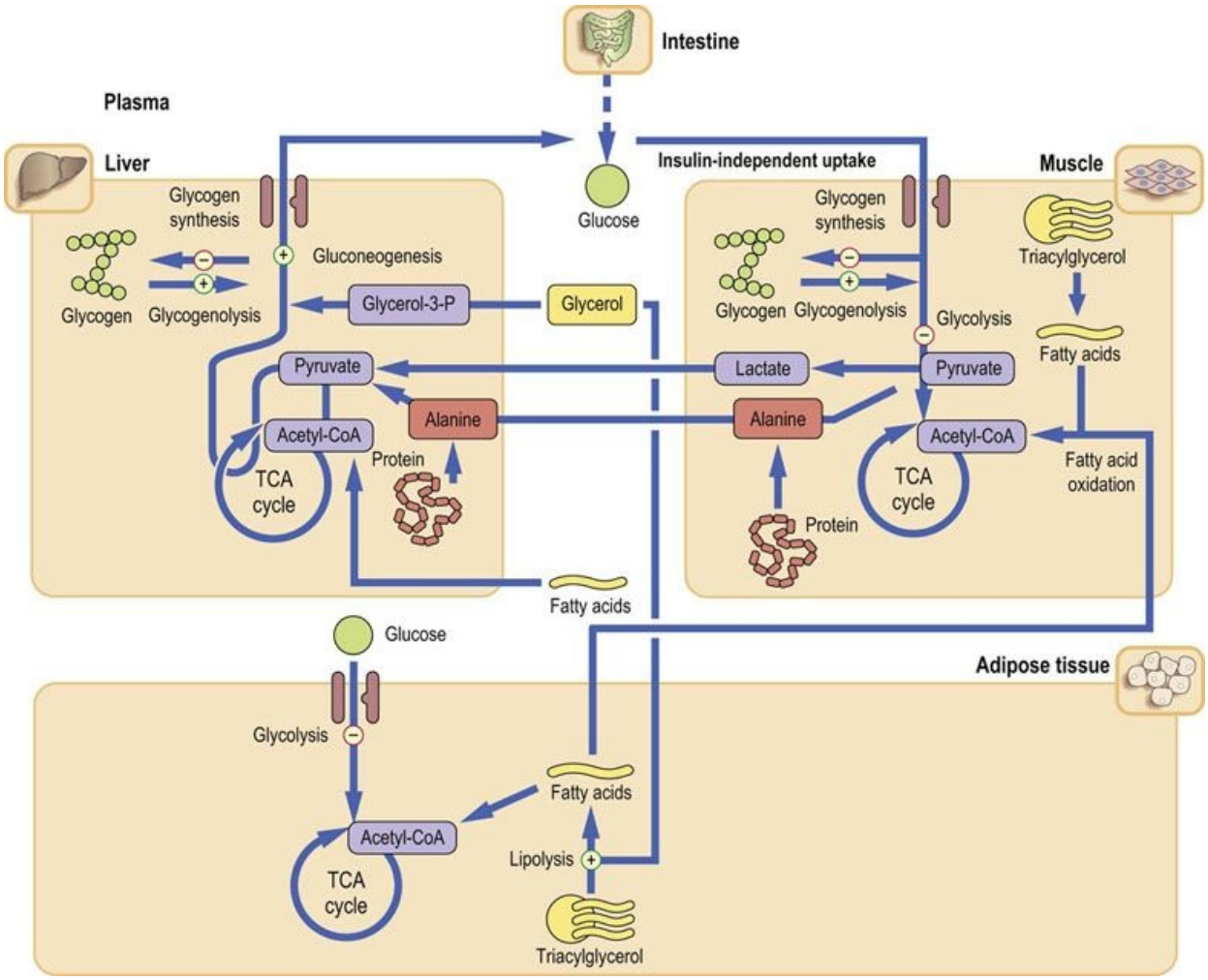


FIG. 21.11 Metabolism after an overnight fast (postabsorptive state).
In the postabsorptive state, liver metabolism changes from glucose utilization to glucose production (through gluconeogenesis stimulated by glucagon). Glucagon also stimulates glycogenolysis and inhibits glycolysis. The substrates for gluconeogenesis are **alanine, lactate and glycerol**. Alanine and lactate are transported to the liver from muscle (see Fig. 21.13). Glucose uptake by the muscle and adipose tissue decreases. Degradation of triacylglycerols (lipolysis) and subsequent fatty acid oxidation are stimulated, providing energy.

The three key substrates for gluconeogenesis are lactate, alanine and glycerol

In the fasting state the insulin-dependent tissues use relatively little glucose – the muscle and adipose tissue together use only 20% of all available glucose. As much as 80% of all glucose is taken up by insulin-independent tissues. Of this, 50% goes to the brain and 20% to the erythrocytes.

After a 12 h fast, 65–75% of synthesized glucose is still derived from glycogen; the rest comes from gluconeogenesis. During longer fast, the contribution of gluconeogenesis steadily increases. Muscle facilitates gluconeogenesis by releasing lactate, which is taken up by the liver. It is oxidized to pyruvate, which enters gluconeogenesis. The newly synthesized glucose is released from the liver and returns to the skeletal muscle. This closes the loop known as the **Cori cycle** (Fig. 21.12).

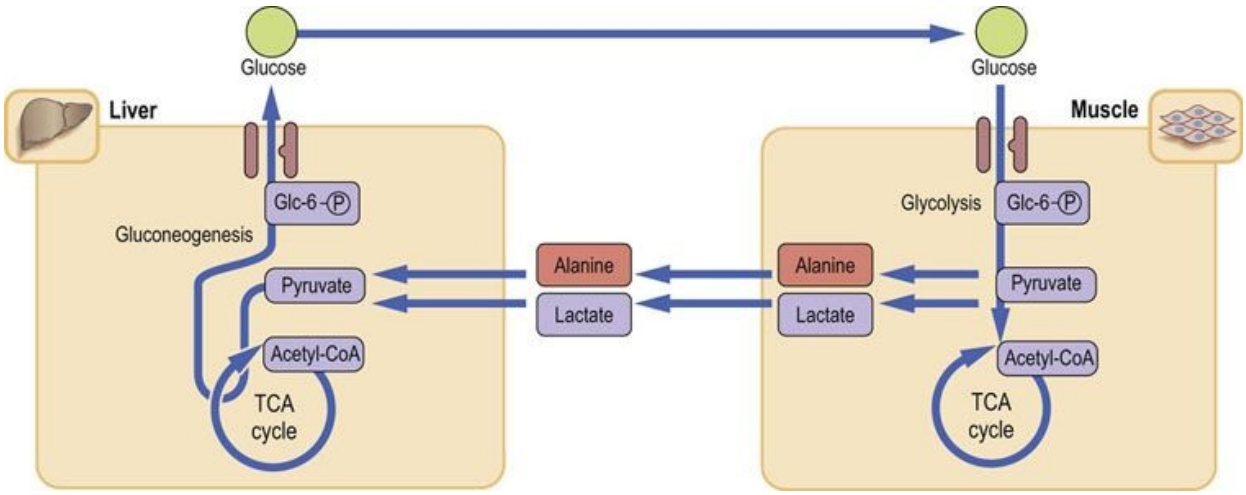


FIG. 21.12 Cori cycle and glucose–alanine cycle.
The Cori (glucose–lactate) cycle allows recycling of lactate back to glucose. Alanine is derived mostly from muscle proteolysis.

Low insulin concentration contributes to the release of amino acids (primarily alanine and glutamine) from muscle by stimulating muscle proteolysis. Alanine, analogously to lactate, is taken up by the liver, and converted to pyruvate. This **glucose–alanine cycle** parallels the Cori cycle.

The third gluconeogenic substrate, glycerol, is released during the hydrolysis of triacylglycerols by the hormone-sensitive lipase (lipolysis), a process stimulated by glucagon. The fatty acids liberated from triacylglycerols stimulate ketogenesis from the acetyl-CoA. Ketogenesis in turn yields acetoacetate, hydroxybutyrate, and the product of spontaneous decarboxylation of acetoacetate, acetone (the products of ketogenesis are collectively known as ketone bodies). During prolonged fasting, ketone bodies can be used as energy substrates in the heart and skeletal muscle, and also the brain.

Prolonged fasting

Prolonged fasting (starvation) is a chronic low-insulin, high-glucagon state. It is also accompanied by a decrease in the concentration of thyroid hormones: this decreases the metabolic rate. Free fatty acids now become the major energy substrate – the β-oxidation of the fatty acids generates acetyl-CoA to be used in the TCA cycle.

However, because the ongoing gluconeogenesis depletes oxaloacetate (the TCA cycle metabolite), the activity of the TCA cycle (Chapter 14) decreases. This causes accumulation of the acetyl-CoA and its channeling into ketogenesis. Thus, during prolonged fasting, the concentration of ketone bodies in plasma increases.

In addition, there are mechanisms that protect body proteins during starvation, by decreasing the need for endogenous glucose production.

The use of proteins as substrates for gluconeogenesis is minimized by almost total dependence on fat as energy source (Fig. 21.13). The Cori cycle also decreases the requirement for endogenous glucose. Further, the amount of the GLUT-4 transporter in the adipose tissue and muscle decreases, decreasing cellular glucose uptake and thus ‘saving’ even more glucose. Eventually, the brain also adapts to the use of ketone bodies as fuel. Some brain tissue actually switches from complete oxidation of glucose to glycolysis.

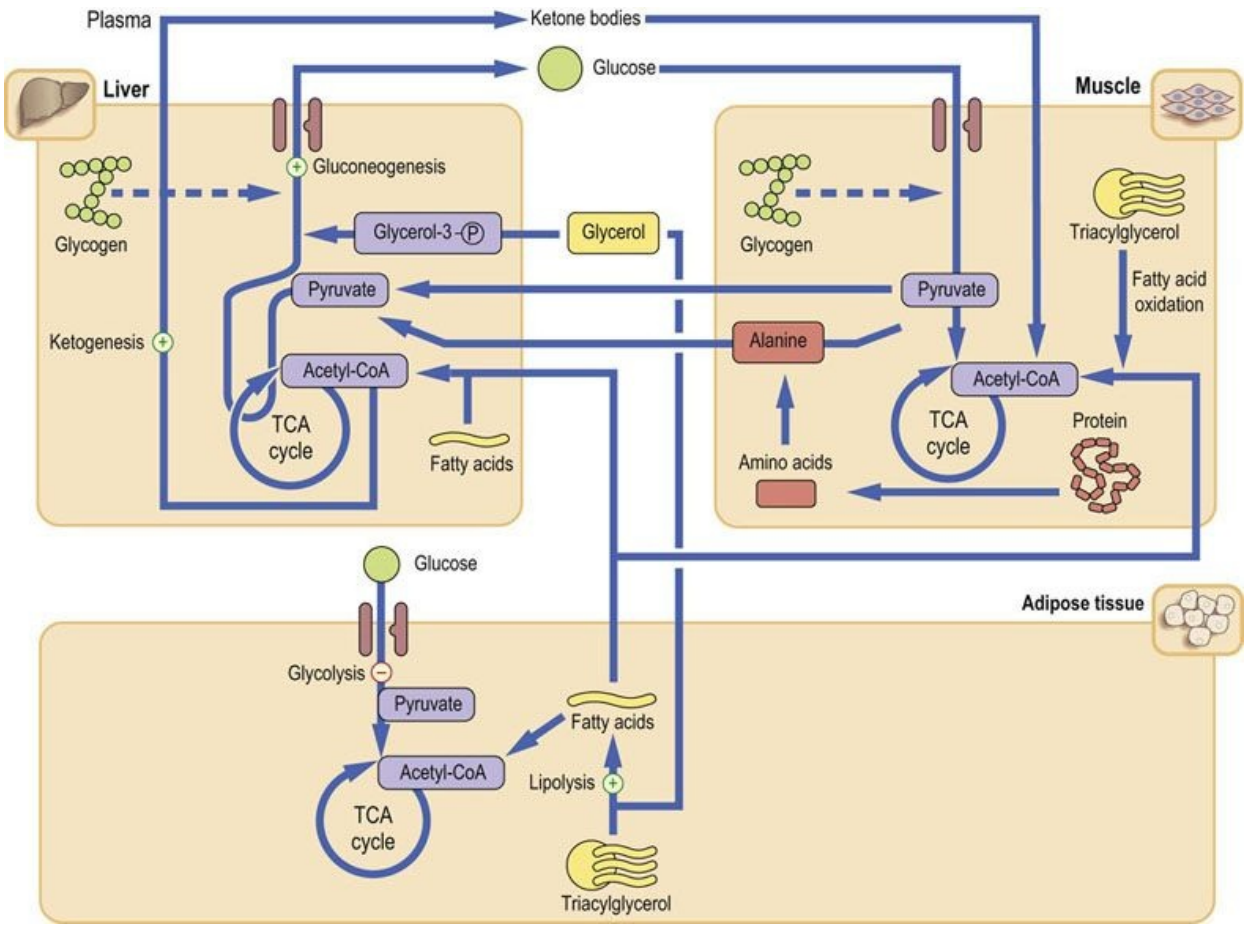


FIG. 21.13 Metabolism during prolonged fast (starvation).
The catabolic state is similar to that during short-term fasting, but adaptive responses are in operation. At this stage glycogen stores are depleted, and the supply of metabolic fuels depends on gluconeogenesis and lipolysis. Ketone bodies produced from large amounts of acetyl-CoA generated by fatty acid oxidation become an important energy source for the muscle. The brain also adapts to their use. Importantly, decreased demand for glucose (and thus gluconeogenesis) in turn decreases the demand for alanine, ‘sparing’ muscle proteins.

Metabolism during stress

Response to stress is driven by the anti-insulin hormones

The stress response occurs not only during ‘fight and flight’ situations but is also triggered by **trauma, burns, surgery and infection**: therefore it is fundamentally important in medicine. Generally, it is associated with hypermetabolism and the increased activity of the sympathetic nervous system. The metabolic response is driven by the anti-insulin hormones: catecholamines, primarily epinephrine (Fig. 13.5), glucagon and cortisol. The anabolic pathways (glycogen synthesis, lipogenesis) are suppressed, and the catabolic pathways (glycogenolysis, lipolysis, and proteolysis) are stimulated. The insulin-independent peripheral glucose uptake increases (Fig. 21.14). Clinically, there is early vasoconstriction, which aims to limit possible blood loss. There is also fever, tachycardia, tachypnea (increased respiratory rate), and leukocytosis.

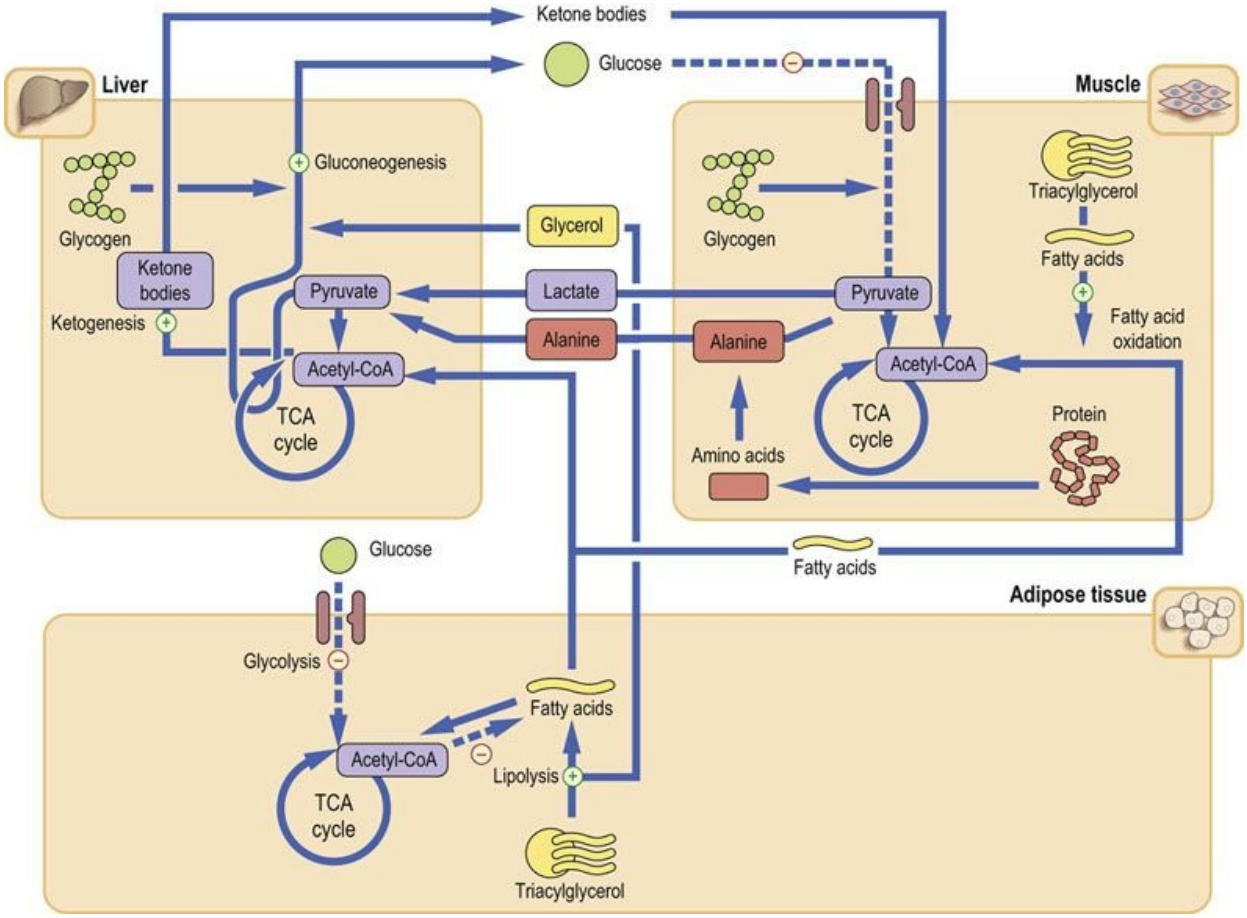


FIG. 21.14 Metabolism during stress and injury.
The metabolic response is catabolic, and is broadly analogous to fasting. Glucose is mobilized from all available sources. However, here epinephrine plays a key role and, together with glucagon, inhibits insulin secretion. Stress also induces peripheral insulin resistance, further sparing glucose. Energy is provided from glucose, fatty acids and from protein catabolism.

Stress response mobilizes energy substrates from all available sources

The priority is to provide glucose for the brain: epinephrine and glucagon stimulate glycogenolysis and gluconeogenesis. In addition, decreased peripheral uptake of glucose makes more of it available to the brain. Later, the metabolic rate increases and fatty acids become the major source of energy. The amino acids for gluconeogenesis are supplied from the muscle; this leads to the negative nitrogen balance within 2–3 days of injury.

Stress response includes insulin resistance

Insulin-dependent transport of glucose into cells, mediated by GLUT-4, decreases under the influence of the glucocorticoids. Also, glucocorticoids, by inducing genes coding for Glc-6-Pase and PEPCK, facilitate stimulation of gluconeogenesis by glucagon and catecholamines (Table 21.3). Insulin-independent glucose uptake increases, particularly in muscle. This is mediated by tumor necrosis factor α (TNF-α) and cytokines such as interleukin-1 (IL-1) (Chapter 42). TNF-α also stimulates muscle glycogen breakdown. IL-6 helps to induce PEPCK, stimulates lipolysis in the adipose tissue, and contributes to muscle proteolysis. There is an increase in lactate production.

Stress response affects results of laboratory tests

Metabolic response to stress affects the results of common laboratory measurements. The key finding is hyperglycemia; importantly, during stress a degree of hyperglycemia should not be confused with diabetes mellitus. Also, injury, infection, and trauma are associated with the acute phase response, which stimulates synthesis of a range of proteins, such as α_1-antitrypsin, C-reactive protein (CRP), haptoglobin, α_1-acid glycoprotein, complement and others. On the other hand, the albumin synthesis is suppressed. The measurements of CRP are essential in monitoring the progress of treatment in patients with severe infections (Chapter 4).

Clinical box A woman with chest pain and elevated plasma glucose concentration

a stress-induced hyperglycemia

A 66-year-old woman was admitted to the cardiology ward with chest pain. A myocardial infarction was diagnosed on the basis of EKG and increased plasma troponin concentration. She was successfully treated with thrombolysis. At that time, her random plasma glucose concentration was 10.5 mmol/L (189 mg/dL). Next day the fasting blood glucose was only slightly raised at 7.5 mmol/L (117 mg/dL). Normal fasting plasma glucose is 4.0–6.0 mmol/L (72–109 mg/dL).

Comment.

Major stress associated with myocardial infarction is associated with the counterregulatory hormone response and this leads to the elevation of blood glucose concentration. Care is needed in the interpretation of raised fasting plasma glucose levels in the context of acute illness. A glucose tolerance test should not be performed during acute illness.

Diabetes mellitus

According to the International Diabetes Federation (2012 data) there are 371 million people living with diabetes worldwide (these numbers include the predicted number of undiagnosed cases). The prevalence of diabetes is on the increase. There were 80 million diabetic individuals in 1985 and it is predicted that there will be up to 430 million in 2030 (see Tuchman in the Further reading section). This increase is linked to lifestyles that include excess of high-energy foods combined with little physical exercise – and to the rising prevalence of obesity. Diabetes results from a combined effect of genetic and environmental influences. In other words, the environmental effects and lifestyle are likely to cause diabetes in persons who are genetically (and epigenetically, see Chapter 36) susceptible. More than 40 loci have been currently associated with type 2 diabetes alone: however, only around 10% of heritability of diabetes can be explained by these loci.

Diabetes is a disorder of fuel metabolism characterized by hyperglycemia and (later) by vascular damage

There are two major components to the syndrome of diabetes mellitus: the hyperglycemia and the vascular complications. There are **four main forms of diabetes mellitus**: type 1 diabetes (T1D), type 2 diabetes (T2D), secondary diabetes, and gestational diabetes (GDM, diabetes of pregnancy).

The most common is T2D: 90% of all diabetic patients have T2D, and 5–10% have T1D. The other two forms are, in comparison, rare. Diabetes has a strong genetic component but is often unmasked by lifestyle factors. Importantly, some diabetic patients have no clinical symptoms at all. In these individuals the diagnosis is made solely on the basis of laboratory results – hence the importance of laboratory tests (Table 21.4).

Table 21.4

Classification of diabetes mellitus

Syndrome	Comments
Type 1	Autoimmune destruction of β-cells
Type 2	Impairment of β-cells: B-cell inability to compensate for insulin resistance
Other types	Genetic defects of β-cells (e.g. mutations of glucokinase gene). Rare insulin resistance syndromes
	Diseases of exocrine pancreas. Endocrine diseases (acromegaly, Cushing's syndrome). Drugs and chemical-induced diabetes, infections (e.g. mumps)

	Rare syndromes characterized by the presence of antireceptor antibodies Diabetes accompanying other genetic diseases (e.g. Down syndrome)
Gestational diabetes	Any degree of glucose intolerance diagnosed in pregnancy

Type 1 diabetes was in older literature described as insulin-dependent diabetes (IDDM) and type 2 as noninsulin-dependent diabetes (NIDDM) or maturity-onset diabetes.

In the long term, diabetes leads to changes in the walls of small (**microangiopathy**) and large (**macroangiopathy**) arteries. Microangiopathy occurring in the kidney (diabetic nephropathy) may lead to kidney failure (see Box on p. 329). Microangiopathy developing in the retina (diabetic retinopathy) may cause blindness, and that occurring in the nervous system (diabetic neuropathy) leads to impairment of autonomic nerve function. Diabetic patients also develop lens opacities (cataracts). Diabetes is the main cause of blindness in the Western world and one of the main causes of kidney failure.

Diabetic macroangiopathy is associated with a 2–3 times greater risk of myocardial infarction than in nondiabetic persons of similar age. When macroangiopathy affects peripheral arteries, it leads to diabetic peripheral vascular disease, and to foot ulceration: diabetes remains a major cause of lower limb amputations. However, the most prevalent complication of diabetes is cardiovascular disease, which is the cause of death in more than 80% of type 2 diabetic patients.

Type 1 diabetes

Type 1 diabetes is an autoimmune disease

T1D usually develops in people below 35 years of age, with the peak incidence at approximately 12 years. However, this seems to change: more recent data suggest that only 50–60% of patients are now younger than 16-18 years at presentation. T1D is caused by autoimmune destruction of the pancreatic β-cells. Its precipitating cause remains unclear; it could be a viral infection (e.g. congenital rubella), environmental toxins or foods. The autoimmune reaction could be initiated by the cytokine response to infection.

Susceptibility to type 1 diabetes is inherited

The concordance rate for T1D in monozygotic twins is 30–40%. The risk of diabetes is greater if there is paternal (compared to maternal) history. Susceptibility genes are located on chromosome 6 in the major histocompatibility complex (MHC; Chapter 38). Approximately 50% of the

genetic susceptibility to diabetes resides in the HLA genes: HLA genotypes DR and DQ, and to a lesser extent in other loci known as IDDM2 (insulin-VNTR) and IDDM12 (CTLA-4). Both genes imparting risk (DR3/4, DQA1*0301-DQB1*0302 and DQA1*0501-DQB1*0201) and genes imparting protection (DQA1*0102-DQB1*0602) have been identified within the HLA complex. Interestingly, this region also contains susceptibility genes associated with other autoimmune diseases; this means that patients with type 1 diabetes are more susceptible to other autoimmune disorders such as Graves' disease, Addison's disease and celiac disease.

In addition to the inflammatory infiltration of the islets resulting from the abnormal T cell response (Chapter 38), a proportion of patients demonstrate an abnormal B cell response with circulating antibodies against various β-cell proteins. The autoantibodies may appear years before diagnosis and might be directed against insulin, glutamic acid decarboxylase (GAD) and to protein tyrosine phosphatase, or against islet antigens.

Persons with T1D are prone to the development of ketoacidosis and are dependent on insulin treatment. The development of symptomatic disease may happen quickly, and diabetic ketoacidosis may be present at diagnosis.

Type 2 diabetes

Impairment of β-cell function is the main factor in the development of type 2 diabetes

T2D usually develops in obese patients who are over 40 years old. However, in recent years, it has been increasingly observed in younger people. There has been a long-standing discussion of which of the two main contributing factors is dominant: the impairment of the pancreatic β-cell function (and consequently insulin secretion) or the peripheral insulin resistance. Recent interpretations favor the former. The majority of the discovered diabetes susceptibility genes are linked to insulin secretion and to the β-cell function. Thus, the inability of the β-cell to deal with the excessive nutrient supply, and to compensate for the existing insulin resistance, is the main factor leading to the development of diabetes. There is also a wider derangement of the glucose homeostatic system. This includes increased glucagon secretion, reduced incretin response, and also decreased secretion of adiponectin and inflammation in the adipose tissue.

Obesity is a key risk factor for type 2 diabetes

The two most important risk factors for T2D are family history and obesity. Obesity, in turn, is closely linked to insulin resistance.

Insulin resistance first presents as hyperinsulinemia with normal glucose concentration – a sign of the β-cells effectively compensating for the peripheral insulin resistance. When such compensation fails, plasma glucose concentration may increase slightly on fasting (this condition is known as impaired fasting glucose, IFG) or in response to glucose load (impaired glucose tolerance, IGT). Both IFG and IGT are predictors of diabetes. IGT is also associated with an increased risk of macrovascular complications. Further increase in insulin resistance and the consequent impairment of insulin secretion lead to overt T2D. Importantly, the deterioration of glucose tolerance can be slowed down or, occasionally, reversed by weight reduction and exercise. Thus, IFG and IGT are strong signals for individuals to change their lifestyles and minimize the chance of developing overt diabetes later.

Heritability of type 2 diabetes is greater than 50%

Monozygotic twins are approximately 70% concordant with respect to T2D, and the concordance rate in dizygotic twins is 20–30%. First-degree relatives of diabetic persons have a 40% chance of developing the disease.

Population testing identified six genes associated with T2D. They include the one coding for peroxisome proliferator-activated receptor gamma (PPAR-γ). Another gene (IRS1) codes for the IRS-1, and is associated with an impaired peripheral response to insulin. The gene coding for the potassium inwardly-rectifying channel, subfamily J, member 11 (KCNJ11) affects the insulin secretion: it encodes two proteins, which constitute an ATP-sensitive potassium channel that regulates glucose-dependent insulin secretion in the β-cells. Its mutation causes a rare form of neonatal diabetes. The WFS1 gene codes for wolframin, a protein detected in patients with a syndrome including both diabetes insipidus and juvenile diabetes, as well as deafness and optic atrophy. Finally, the genes coding for the HNF1 homeobox A (HNF1A) and HNF1 homeobox B (HNF1B) are associated with forms of monogenic diabetes (MODY, see below).

In 2007 the genome-wide association (GWAS) study identified further two genes associated with T2D, the most important being the solute carrier family 30, member 8 (SLC30A8) – a gene that, somewhat surprisingly, codes for a zinc

transporter. Another identified gene was the hematopoietically expressed homeobox HHEX. Still more susceptibility genes were identified by subsequent large meta-analyses of the GWAS studies. These were further genes associated with insulin secretion such as MTNR1B (melatonin receptor), which is associated with impairment of the first phase of insulin secretion. The other ones, such as GCKR and IGF1, are associated with insulin resistance. The obesity-associated genes such as FTO (Chapter 22) are also linked to T2D.

Maturity-onset diabetes of the young (MODY) is a rare form of type 2 diabetes

MODY develops before the age of 25 years, is characterized by persistent C-peptide secretion and a clear pattern of inheritance. MODY results from mutations of at least six different genes; among them are the one coding for the glucokinase (affecting β-cell sensing of glucose and causing MODY2), and for transcription factor HNF1A, which causes MODY3, and HNF1B, which causes MODY5. There is also a mitochondrial DNA mutation that leads to impaired oxidative phosphorylation and causes the so-called mitochondrial diabetes.

The MODY-associated genes also give a fascinating insight into the genetic determination of the response to treatment. In contrast to MODY types caused by the transcription factor mutations, MODY2 responds to diet, and these patients do not require insulin. Patients with HNF1A mutations respond to sulfonylurea drugs.

A continuum between type 1 and type 2 diabetes?

Approximately 10% of patients diagnosed with T2D also have autoantibodies to glutamic acid decarboxylase; these are actually individuals who are developing a form of T1D, sometimes called latent autoimmune diabetes in adults (LADA).

The so-called accelerator hypothesis (developed by Wilkin) proposes that there are the three main ‘accelerators’ of diabetes: the potential for β-cell apoptosis, the insulin resistance and the predisposition to β-cell immunity. If all three are present, T1D develops. If the predisposition to autoimmunity is absent, T2D ensues.

In type 2 diabetes, ketoacidosis is rare

Patients with T2D may develop microvascular complications as in T1D but the main problem is the macrovascular complications, which eventually lead to

coronary heart disease, peripheral vascular disease and stroke. T1D and T2D are compared in Table 21.5.

Table 21.5

Comparison of type 1 and type 2 diabetes

	Type 1	Type 2
Onset	Usually under 20 years of age	Usually over 40 years of age
Insulin synthesis	Absent: immune destruction of β-cells	Preserved: combination of insulin resistance and impaired β-cell function
Plasma insulin concentration	Low or zero	Low, normal or high
Genetic susceptibility	Yes	Yes
Islet cell antibodies at diagnosis	Yes	No
Obesity	Uncommon	Common
Ketoacidosis	Yes	Rare – can be precipitated by major metabolic stress
Treatment	Insulin	Hypoglycemic drugs and insulin

Metabolism in diabetes

In type 1 diabetes, glucose cannot enter insulin-dependent cells such as adipocytes and myocytes because of lack of insulin

Lack of insulin means that metabolism, by default, enters the glucagon-controlled mode. Glycolysis and lipogenesis are inhibited, whereas glycogenolysis, lipolysis, ketogenesis, and gluconeogenesis are stimulated by glucagon (Fig. 21.15). The liver turns into a glucose-producing organ. This, combined with impaired glucose transport into cells, leads to fasting hyperglycemia.

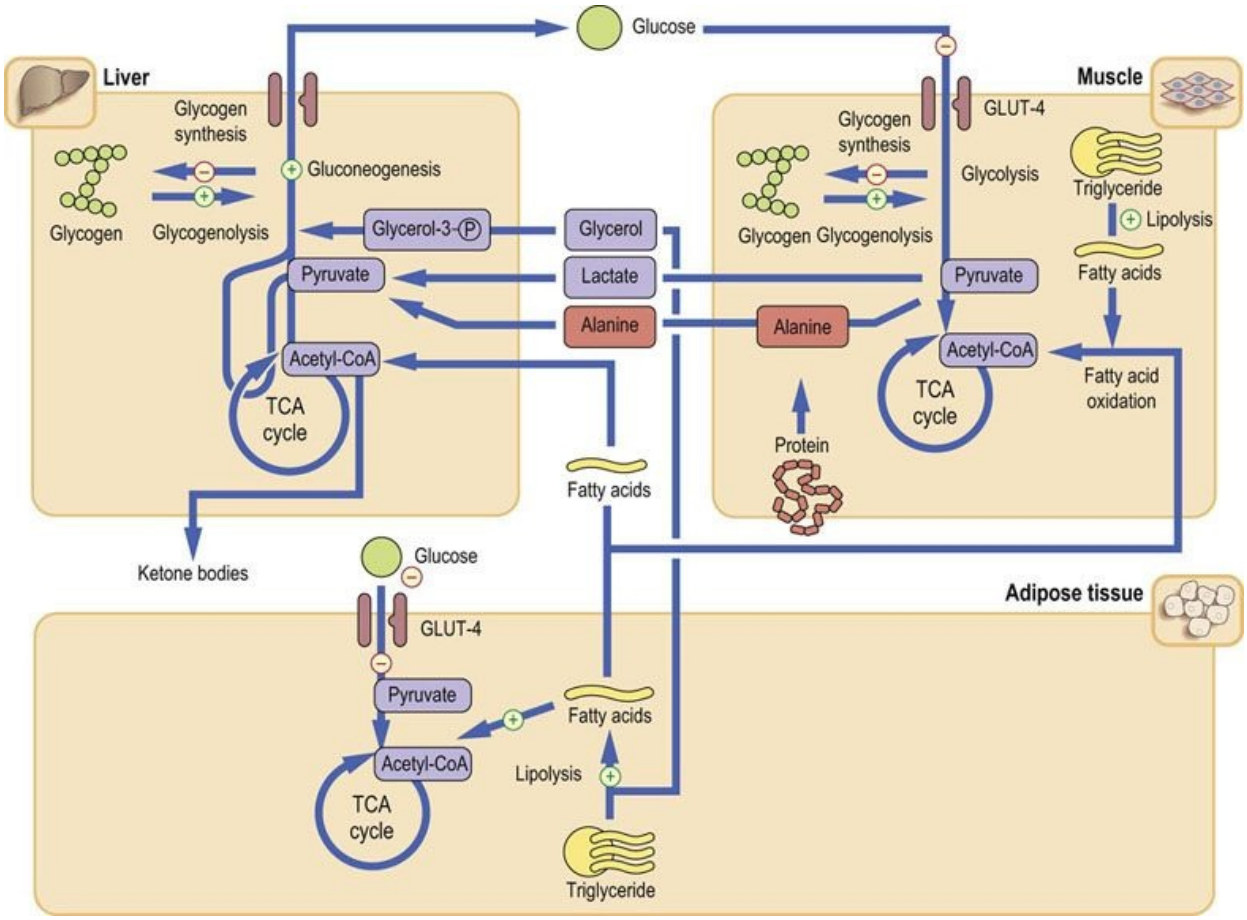


FIG. 21.15 Metabolism in diabetes mellitus.
In diabetes there is a decreased ability of tissues to use glucose because of insulin lack, defective insulin action, or both. Hyperglycemia results from the combined effect of impaired peripheral glucose uptake and increased liver gluconeogenesis. The excess of fatty acids available to the liver, together with the less efficient TCA cycle due to oxaloacetate being used for gluconeogenesis, results in the accumulation of acetyl-CoA and its conversion into ketone bodies.

When plasma glucose concentration exceeds renal capacity for reabsorption, glucose appears in the urine. Because glucose is osmotically active, its excretion is accompanied by increased water loss (the osmotic diuresis). Poorly controlled diabetic patients pass large volumes of urine (polyuria) and drink excessive amount of fluids (polydipsia). Fluid loss eventually leads to dehydration (Chapter 23). In parallel to the disturbed water balance, lipolysis generates an excess of acetyl-CoA, which enters ketogenesis. The concentration of ketone bodies in plasma increases (ketonemia) and they are excreted in the urine (ketonuria). Overproduction of acetoacetic and β-hydroxybutyric acids increases the blood hydrogen ion concentration (the blood pH decreases). This form of

metabolic acidosis (Chapter 24) is known as the diabetic ketoacidosis (Fig. 21.16). In some patients, acetone can be smelled on the breath.

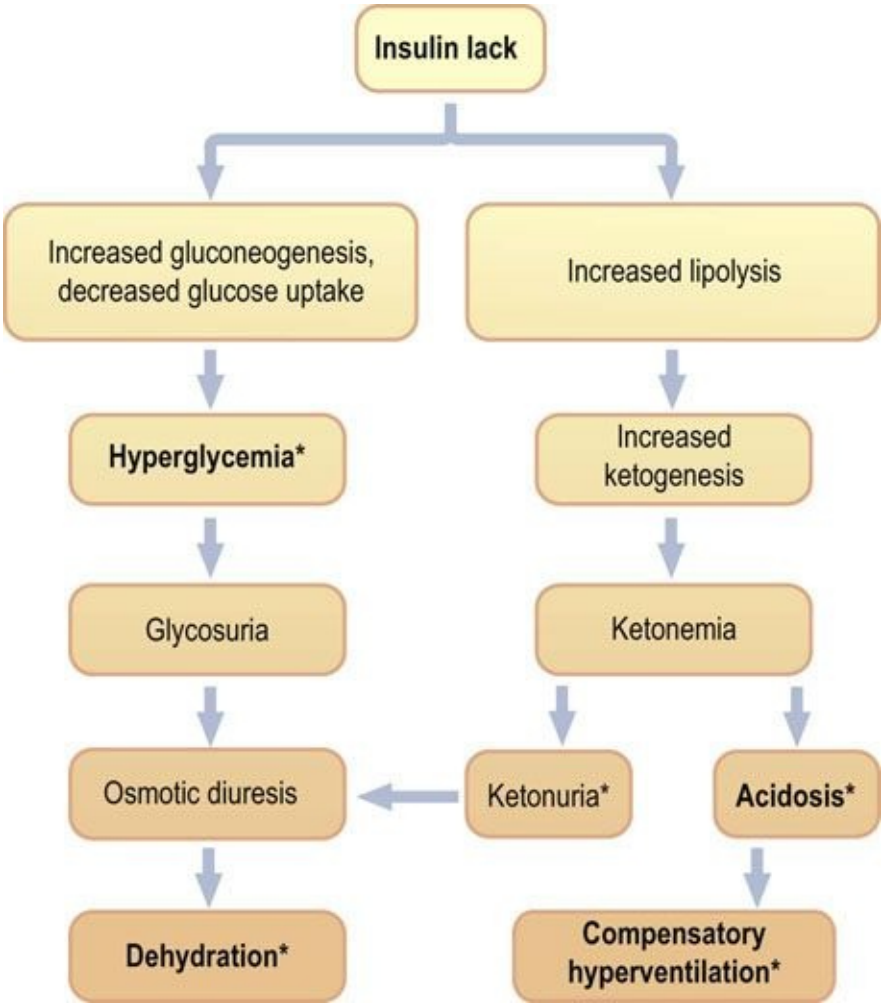


FIG. 21.16 Diabetic ketoacidosis.
Clinical picture of ketoacidosis is a consequence of insulin lack and resulting hyperglycemia and its complications (osmotic diuresis, dehydration), as well as the increased lipolysis and ketogenesis (ketonemia and acidosis). Treatment of ketoacidosis is devised to combat these problems and includes insulin infusion, rehydration and potassium supplementation. *Indicates the most important clinical and laboratory findings.

Ketoacidosis is a major acute complication of poorly controlled diabetes

The key features of diabetic ketoacidosis are hyperglycemia, ketonuria, dehydration and metabolic acidosis. Diabetic ketoacidosis may develop quickly, sometimes even after missing just a single dose of insulin. Ketoacidosis develops predominantly in persons with T1D who have no, or very little, insulin in plasma and, consequently, a very low insulin-to-glucagon concentration ratio. It is rare in T2D, although it may occur after a major stress, such as the myocardial infarction. Untreated ketoacidosis is life-threatening.

Note that there are substantial similarities between metabolism in the fasting state and in diabetes; this is why diabetes was once described as ‘starvation in the midst of plenty’. However, whereas fasting leads just to moderate ketonemia, in diabetes there is accumulation of large amounts of ketone bodies.

Clinical box A 15-year-old girl admitted confused and with breath smelling of acetone

diabetic ketoacidosis

A 15-year-old girl, never known to have diabetes, was admitted to the accident and emergency department. She was confused and her breath had a smell of acetone. She had signs of dehydration with reduced tissue turgor and dry tongue. She also had rapid pauseless respirations. Her blood glucose was 18.0 mmol/L (324 mg/dL) and ketones were present in the urine. Her serum potassium concentration was 4.9 mmol/L (normal 3.5–5.0 mmol/L) and her arterial blood pH was 7.20 (normal 7.37–7.44) (H^+ concentration 63 nmol/L; normal 35–45).

Comment.

This is a typical (if unexpected in this case) presentation of diabetic ketoacidosis. Hyperventilation is a compensatory response to acidosis (see Chapter 24). Diabetic ketoacidosis is a medical emergency. The patient received an intravenous infusion containing physiologic saline with potassium supplements to replace lost fluid, and an infusion of insulin. Note that a substantial proportion of children present with ketoacidosis at the time of diagnosis of diabetes.

Diabetes, obesity, and hypertension are linked with cardiovascular disease

Obesity, insulin resistance and glucose intolerance (or diabetes) may be accompanied by dyslipidemia (Chapter 18) and arterial hypertension. Such cluster of conditions has been described as the **metabolic syndrome**. It is associated with low-grade inflammation affecting the vasculature (Chapter 18), and with an increased tendency to thrombosis (the hypercoagulable state; Chapter 7). Most importantly, it imparts an increased risk of cardiovascular disease.

'Metabolic syndrome' is a concept formulated to increase clinicians' awareness of multiple risk factors for cardiovascular disease when they occur in one individual. The search for a possible common denominator between the major risk factors continues. Increasingly, there is realization that diabetes mellitus and cardiovascular disease may have what some researchers call 'common soil'. The various links between obesity, diabetes and atherosclerosis are illustrated in Figure 21.17.

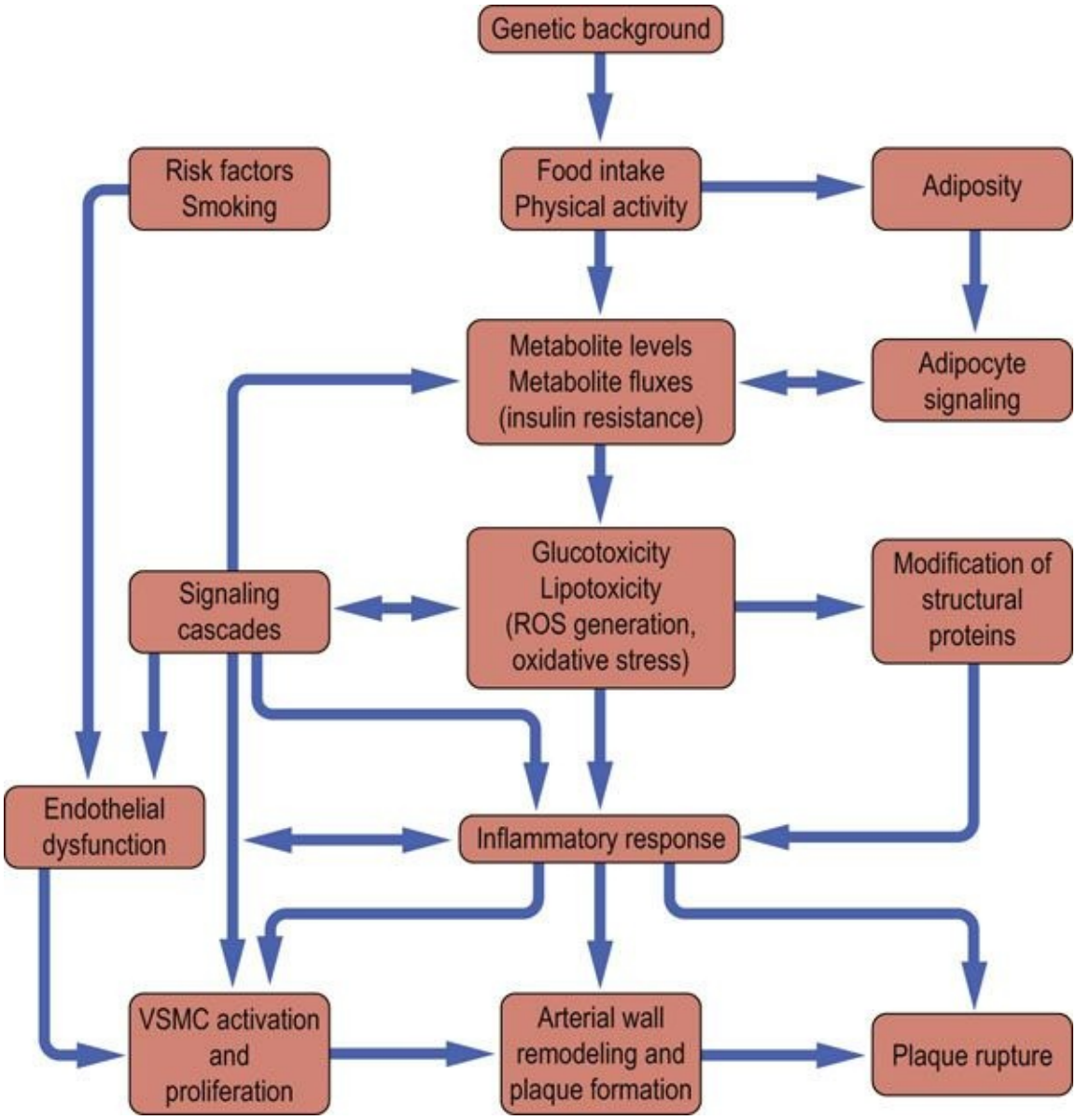


FIG. 21.17 Links between obesity, glucose intolerance, diabetes and atherosclerosis. In obesity the endocrine activity of the adipose tissue is important in the development of insulin resistance, glucose intolerance and type 2 diabetes. Low-grade inflammation and increased oxidative stress can be induced both by obesity and by the classic cardiovascular risk factors. This results in endothelial damage. Once diabetes is present, protein glycation and formation of the advanced glycation (glycoxidation) end products further contribute to vascular damage.

Clinical box A 56-year-old man with effort-related chest discomfort

diabetes and ischemic heart disease

diabetes and ischemic heart disease

A 45-year-old man was referred to the cardiology outpatient clinic for investigation of chest discomfort which he felt when climbing steep hills, and when he was stressed or excited. The patient was 170 cm tall and weighed 102 kg (224 lb). His blood pressure was 160/98 mmHg (upper limit of normal 140/90 mmHg), triglyceride concentration was 4 mmol/L (364 mg/dL) (desirable level, 1.7 mmol/L, 148mg/dL), and fasting plasma glucose was 6.5 mmol/L (117 mg/dL). His resting EKG was normal but an ischemic pattern was observed during exercise testing.

Comment.

This obese man presented with arterial hypertension, hypertriglyceridemia, and impaired fasting glucose. The impaired fasting glucose in this case was due to peripheral insulin resistance. Such a cluster of abnormalities is known as the metabolic syndrome and carries increased risk of coronary heart disease.

The vascular complications of diabetes mellitus

Oxidative stress, advanced glycation (glycoxidation) end products and activity of the polyol pathway contribute to the development of vascular complications

Glucose is toxic in excess. In the presence of transition metals such as copper, it undergoes autooxidation. This generates reactive oxygen species (ROS, Chapter 37). Glucose also attaches nonenzymatically to lysine and valine residues on tissue and plasma proteins in a process known as protein glycation (Fig. 21.18). When glucose interacts with a protein, it first forms a labile compound known as a Schiff base. This, through the so-called Amadori rearrangement, transforms spontaneously to ketoamine. The most widely studied Amadori product is the **glycated hemoglobin** (hemoglobin A_{1c}, HbA_{1c}). Other proteins such as albumin, collagen and apolipoprotein B also undergo glycation. Glycation affects functions of proteins: for instance, their binding to membrane receptors. For example, glycation of apolipoprotein B inhibits the cellular uptake of LDL

particles.

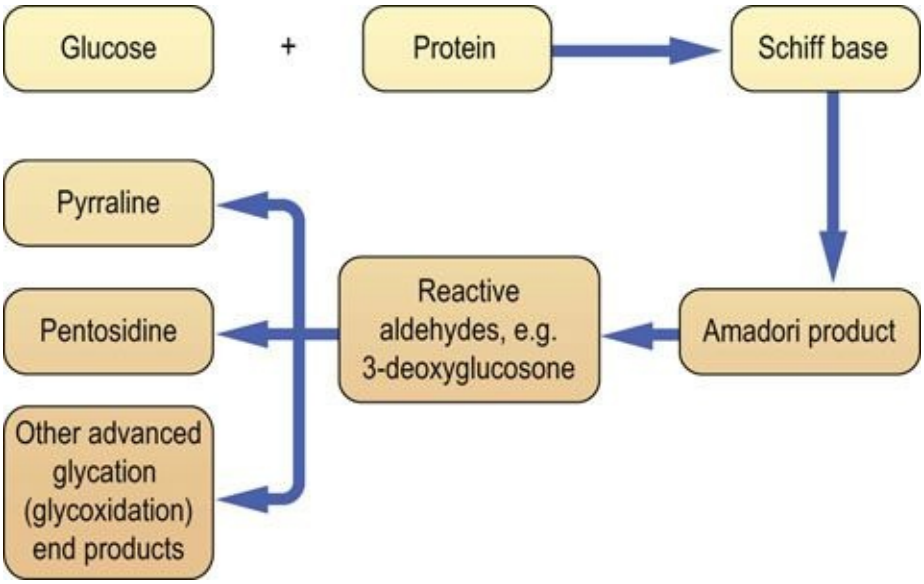


FIG. 21.18 Modification of proteins by glucose: protein glycation and formation of the advanced glycation (glycoxidation) end products.
Nonenzymatic, concentration-dependent, glycation reaction between glucose and protein modifies structure and function of affected proteins. Glycated proteins (Amadori products) are substrates for the formation of the advanced glycation (glycoxidation) end products (AGE). In addition, triose phosphates generated by glycolysis and the increased activity of the polyol pathway can generate reactive precursors of AGE such as methylglyoxal and 3-deoxyglucosone. (Compare Fig. 44.6.)

Further oxidation, rearrangement, dehydration and fragmentation of the Amadori products lead to the formation of a family of compounds known as **advanced glycation (also known as glycoxidation) end products** (AGE). It also generates compounds that possess chemically very active carbonyl groups, such as 3-deoxyglucosone, glyoxal and methylglyoxal (see Fig. 21.18 and also Fig. 44.6). Some of the AGEs are in fact protein crosslinks, forming, for instance, on collagen or myelin, and decreasing elasticity of these proteins. AGE normally accumulate with increasing biological age and this accelerates in diabetes, in the presence of hyperglycemia. AGE bind to their membrane receptors on the endothelial cells, generating oxidative stress. This damages endothelium and stimulates the proinflammatory pathway involving the transcription factor NFκB. NFκB, in turn, controls the expression of a range of inflammatory cytokines such as TNF-α, and IL-1α and IL-6. The result is

chronic low-grade inflammation, known to harm vascular endothelium further. Thus, formation of the AGE is a factor in the development of the microvascular complications of diabetes. It also contributes to atherogenesis. AGE may also be involved in the pathogenesis of other age-related diseases such as Alzheimer's disease.

In addition, the ROS generated during hyperglycemia impair endothelium-dependent relaxation of vascular smooth muscle cells. This is because the vasodilatory nitric oxide (NO) generated by the endothelial cells from arginine is rapidly deactivated by superoxide, forming peroxynitrite radical, which itself is an oxidant. ROS also interfere with signaling cascades, affecting, for instance, the activation of protein kinase C. Hyperglycemia also increases the amount of proton donors within mitochondria, causing increased electrochemical potential difference across the inner mitochondrial membrane and, consequently, increased ROS generation by the respiratory chain (Fig. 37.4). It has been suggested that increased mitochondrial ROS production is the primary cause of long-term diabetic complications.

Increased activity of the polyol pathway is associated with diabetic neuropathy and ocular cataracts

Hyperglycemia alters cellular redox state, by increasing the $NADH/NAD^+$ ratio and decreasing $NADPH/NADP^+$. This directs substrates into the **polyol pathway**, where glucose is reduced to sorbitol by the aldose reductase (Fig. 21.19). Importantly, because the synthesis of NO from arginine requires NADPH, the aldose reductase and nitric oxide synthase compete for it. Sorbitol is further oxidized to fructose by sorbitol dehydrogenase. Aldose reductase has a high K_m for glucose; therefore the polyol pathway is not very active at normal glucose concentrations. However, during hyperglycemia, when glucose concentration in insulin-independent tissues (such as red blood cells, nerve, and lens) increases, the pathway is activated. Similarly to glucose, sorbitol is osmotically active. Its accumulation in the ocular tissue contributes to the development of diabetic cataracts. In the nerve tissue, high concentration of sorbitol decreases the cellular uptake of another alcohol, myoinositol, inhibiting the membrane Na^+/K^+-ATPase, and thus affecting nerve function. The accumulation of sorbitol, hypoxia and reduced nerve blood flow contribute to the development of diabetic neuropathy. Processes that contribute to the long-term diabetic complications are summarized in Fig. 21.20.

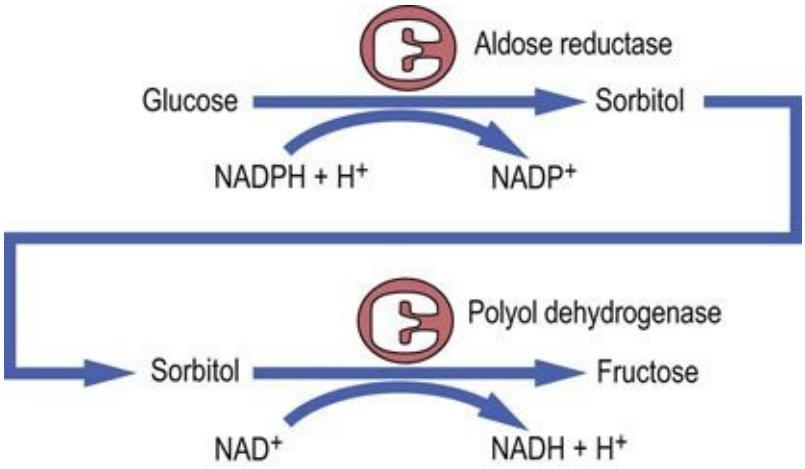


FIG. 21.19 The polyol pathway.
The polyol pathway contributes to the development of diabetic neuropathy. This pathway may be inhibited by inhibitors of its rate-limiting enzyme, aldose reductase.

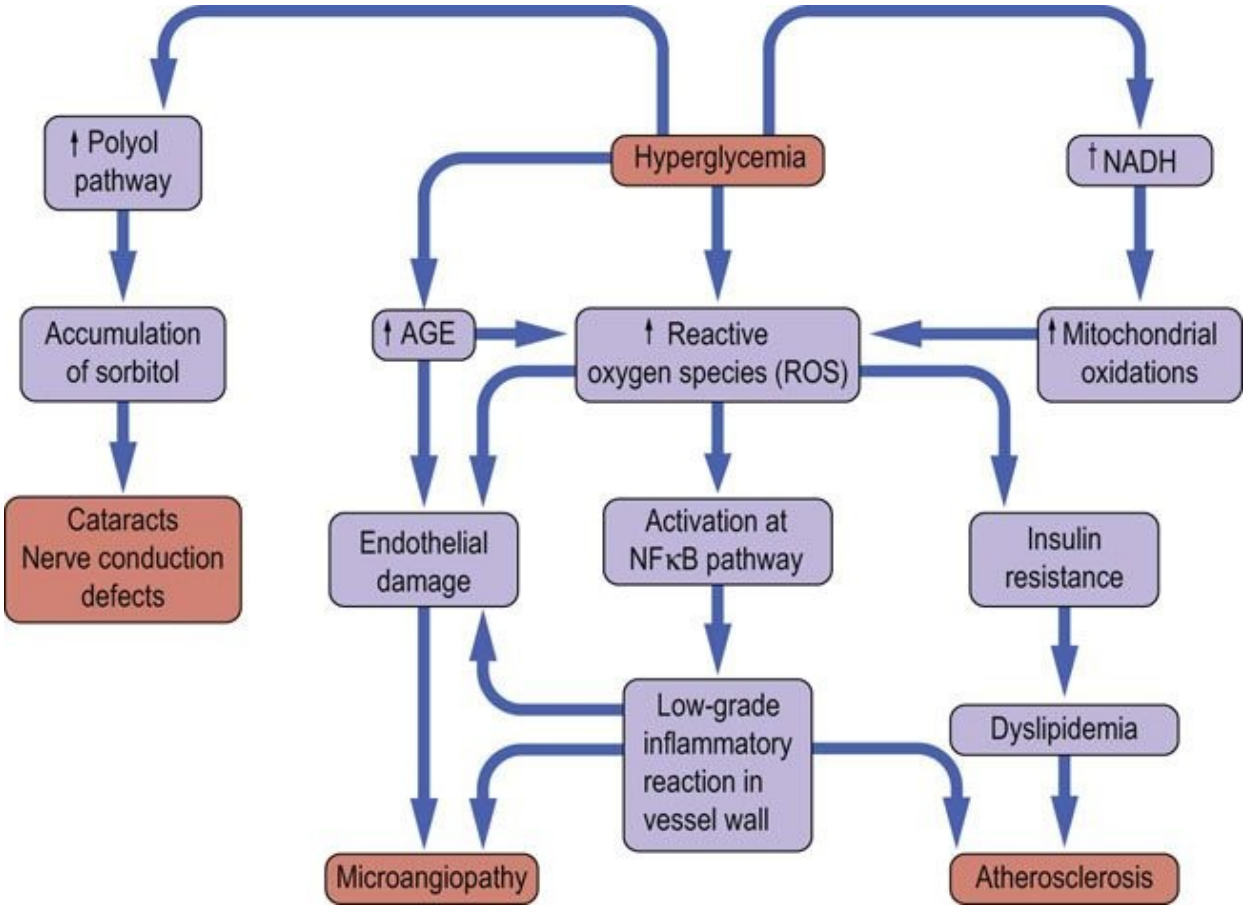


FIG 21.20 Microvascular and macrovascular complications of diabetes mellitus. Poor glycemic control is associated with development of microvascular complications in type 1 and type 2 diabetes, and with increased cardiovascular risk (the latter particularly in type 2 diabetes). Oxidative stress, protein glycation and formation of advanced glycation (glycoxidation) end products (AGE), are the most important candidate mechanisms of development of microvascular complications. Hyperglycemia stimulates generation of the reactive oxygen species (ROS) through increase in the flow of reductive equivalents through the respiratory chain, and through increased formation of AGE. AGE also generate ROS at different stages of their metabolism. ROS toxicity causes structural and functional damage to proteins and stimulates inflammatory phenomena induced through, for instance, the proinflammatory NFκB pathway. ROS damage endothelium and interfere with insulin signaling, contributing to insulin resistance. The low grade inflammation and insulin resistance are particularly important in atherogenesis, causing macrovascular disease (Chapter 18). Note that increased oxidative stress and low-grade chronic inflammation have also been observed in obesity.

Hypoglycemia

Hypoglycemia (the low blood glucose) is blood glucose concentration below 4 mmol/L (72 mg/dL). Low plasma glucose stimulates the sympathetic nervous system. Epinephrine and glucagon are released, initiating the stress response. This manifests itself as sweating, tremor, tachycardia, and a feeling of hunger. The decreased glucose supply to the nervous system (neuroglycopenia) **compromises brain function**: the affected person becomes confused and may lose consciousness (this usually happens when glucose concentration falls below 2.5 mmol/L (45 mg/dL)). Profound hypoglycemia can be fatal.

Hypoglycemia in healthy individuals is usually mild and may occur during exercise, after a period of fasting, or as a result of drinking alcohol. Alcohol increases the intracellular NADH/NAD^+ ratio: this favors conversion of pyruvate to lactate and reduces the amount of pyruvate available for gluconeogenesis. Hypoglycemia may also occur when there is an insufficient amount of counterregulatory hormones to balance the effects of insulin; this happens in adrenal insufficiency (Chapter 39). Another endocrine cause of hypoglycemia is a rare tumor of the β-cells, insulinoma, which may secrete large amounts of insulin. The rare causes of hypoglycemia in childhood are glycogen storage diseases (Chapter 13 and Clinical Box on p. 157). The causes of hypoglycemia are summarized in Fig. 21.21.

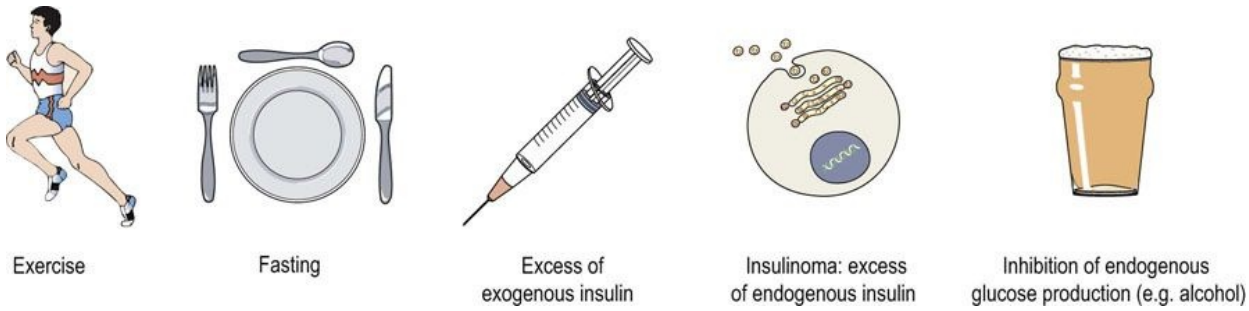


FIG. 21.21 Hypoglycemia.
Hypoglycemia is a plasma glucose concentration below 4 mmol/L (72 mg/dL). Severe hypoglycemia is a glucose concentration below 2.5 mmol/L (45 mg/dL). Hypoglycemia may result from decreased supply of glucose or from an increased insulin secretion. It also results from increased utilization of glucose by tissues (e.g. during exercise).

Hypoglycemia is the most common acute complication of diabetes

It is worth remembering that the most common acute complication of diabetes is not ketoacidosis but hypoglycemia. It can occur in both T1D and T2D. It results from imbalance between insulin dose, carbohydrate supply, and physical activity. Thus, it may occur after taking too much insulin or after missing a meal. Since exercise increases the insulin-independent tissue glucose uptake, diabetic patients, to avoid hypoglycemia, need to decrease their insulin dose before strenuous exercise. Mild hypoglycemia can usually be by taking a sweet drink or eating a few lumps of sugar. Severe hypoglycemia, however, is a medical emergency that requires treatment with intravenous glucose or intramuscular injection of glucagon. Note that the degree of diabetic control achieved during treatment of diabetes is inversely related to the risk of hypoglycemia (the better control, the higher the risk).

Clinical box A 12-year-old diabetic boy WHO lost consciousness on the playing field

hypoglycemia

A 12-year-old diabetic boy was playing with his friends. He received his normal insulin injection in the morning but continued playing through the lunch time without a meal. He became increasingly confused and finally lost consciousness. He was given an injection of glucagon from the emergency kit his father carried, and recovered within minutes. **Severe hypoglycemia is a medical emergency.**

Comment.

An immediate improvement after glucagon injection confirms that this boy's symptoms were caused by hypoglycemia, caused by the combination of the administration of exogenous insulin and insufficient food intake. Recovery from hypoglycemia was due to the action of glucagon. In the hospital, hypoglycemic patients who cannot eat or drink are usually treated with intravenous high concentration glucose. An intramuscular glucagon injection is an

emergency measure that can be applied at home.

Laboratory assessment of fuel metabolism

Diagnosis and monitoring of patients with diabetes mellitus

Measurement of the plasma glucose concentration is the key diagnostic test for diabetes mellitus

The measurement of plasma glucose concentration is the most important test of fuel metabolism. It needs to be interpreted in relation to the feed–fast cycle. The best time to assess carbohydrate metabolism is after an 8–12 h fast (Fig. 21.11), when the fuel metabolism reaches steady state.

Glucose concentration measured irrespective of the meal times is known as the **random plasma glucose**. It is useful for the diagnosis of hypoglycemia or severe hyperglycemia, but it is less helpful in assessing the significance of a mild hyperglycemia.

The measurements that are diagnostic for diabetes are the **fasting glucose** concentration (no caloric intake for approx 10 h) and the concentration measured 2 h after oral ingestion of a standard amount of glucose.

Interpreting the glucose concentration, a clinician wants to know whether it is normal (**normoglycemia**), too high (**hyperglycemia**) or too low (**hypoglycemia**). Further interpretation of hyperglycemia includes the diagnosis of diabetes or identification of the intermediate (prediabetic) stages of metabolism.

A continuum exists between normal, prediabetic and diabetic states

The fasting plasma glucose in an individual is remarkably stable. The prediabetic abnormalities of carbohydrate metabolism are defined as the impaired fasting glucose (IFG) and the impaired glucose tolerance (IGT). The American Diabetes Association (ADA) recommends the diagnosis of IFG. The World Health Organization (WHO) recommends diagnosis of IGT, and this is accepted in Europe.

The diagnostic cut-off points for diabetes mellitus are based on the evidence that higher fasting glucose concentrations constitute a risk for the development of microvascular complications. IGT is also associated with cardiovascular risk, whereas IFG is a risk factor for the future development of diabetes.

Normally, the fasting plasma glucose concentration should remain below 6.1 mmol/L (110mg/dL). The IGT is characterized by a normal fasting level but an elevated concentration 2 h after glucose load. The IFG is defined as an 'intermediate' fasting plasma glucose (higher than 6.0 mmol/L but lower than 7.0 mmol/L (126 mg/dL)). The fasting plasma glucose of 7.0 mmol/L (126 mg/dL) or above, if confirmed, is diagnostic for diabetes. The laboratory diagnosis of diabetes is summarized in Table 21.6.

Table 21.6

The equivalence of units of measurement for glycated hemoglobin (HbA_{1c}) measured using the traditional (DCCT) and reference (IFCC) methods

DCCT units %	IFCC units (mmol/mol)
5	31
6	42
7	53
10	86

DCCT, Diabetes Control and Complications Trial; IFCC, International Federation of Clinical Chemistry and Laboratory Medicine.

Modified from: Misra S, Hancock M, Meeran K, Dornhorst A, Oliver NS. HbA1c: an old friend in new clothes. *Lancet* 377:1476–1477, 2011.

Oral glucose tolerance test (OGTT) assesses blood glucose response to a carbohydrate load

WHO recommends that the OGTT is performed on all individuals whose fasting plasma glucose falls into the IFG category. OGTT must be performed under standard conditions. The patient should attend in the morning, after an approximately 10 h fast. To avoid stress-or exercise-related changes in plasma glucose, the person should sit throughout the test. The test should not be performed during, or immediately after, an acute illness. During the test, the fasting plasma glucose is measured first. The patient is then given a standard quantity of glucose to drink (75 g in 300 mL of water) and the plasma glucose concentration is measured again after 120 min (Fig. 21.22). In some protocols

glucose is measured after 20, 60 and 120 min. Normally, plasma glucose should reach peak concentration after approximately 60 min and should return to a near-fasting state within 120 min. If it remains above 11.1 mmol/L (200 mg/dL/min) in the 120 min sample, diabetes is diagnosed, even if the fasting blood glucose were normal. Nondiabetic fasting blood glucose with the post-load concentration between 6.1 and 7.8 mmol/L (100–140 mg/dL) signifies IGT. Interpretation of the OGTT is summarized in Table 21.7.

Table 21.7

Diagnostic criteria for diabetes mellitus and glucose intolerance

Condition	Diagnostic criteria (mmol/L)	Diagnostic criteria (mg/dL)
Normal fasting plasma glucose	Below 6.1	Below 110
Impaired fasting glucose (IFG)	Equal or above 6.1 but below 7.0	Equal or above 110 but below 126
Impaired glucose tolerance (IGT)	Plasma glucose during OGTT, 2 h after 75 g load 7.8 or above, but below 11.1	Plasma glucose during OGTT, 2 h after 75 g load 140 or above, but below 200
Diabetes mellitus*		
Criterion 1	Random plasma glucose 11.1 or above†	Random plasma glucose 200 or above†
Criterion 2	Fasting plasma glucose 7.0 or above	Fasting plasma glucose 126 or above
Criterion 3	2 h value during OGTT 11.1 or above	2 h value during OGTT 200 or above
Criterion 4	HbA_{1c} > 48 mmol/mol (6.5%)	

*If one of the criteria is fulfilled, diagnosis is provisional. Diagnosis needs to be confirmed next day using a different criterion.

†If accompanied by symptoms (polyuria, polydypsia, unexplained weight loss). These are the criteria proposed by the American Diabetes Association in 1997 (see Further Reading).

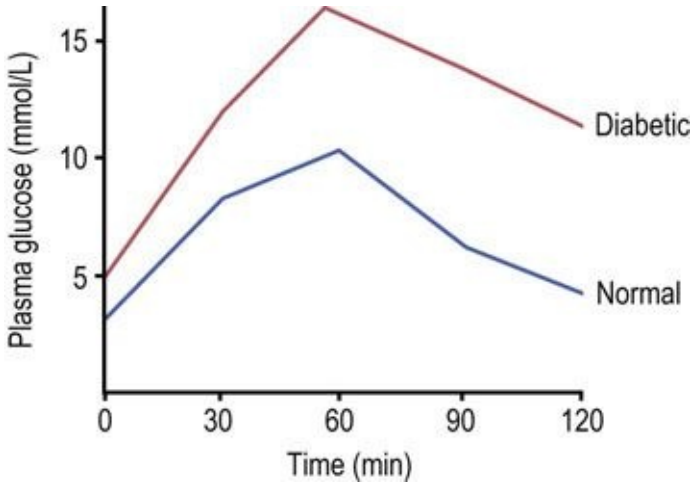


FIG. 21.22 Oral glucose tolerance test (OGTT).
The principle of the test is the measurement of plasma glucose concentration before and after a standard (75 g) oral glucose load. Glucose concentration increases, achieving a peak between 30 and 60 min after the load. It should return to near-fasting values after 2 h. Note higher plasma glucose values at all time points in a diabetic patient.

Glycated hemoglobin (HbA_{1c}) concentration reflects the average concentration of plasma glucose

Knowing an average plasma glucose concentration over a period of time is clinically very relevant, because it is related to the risk of development of late complications of diabetes. Trying to assess the average glycemia by performing multiple plasma glucose measurements is, however, cumbersome. Instead, glycated hemoglobin (hemoglobin A_{1c}, HbA_{1c}) is measured for this purpose. HbA_{1c} forms in the blood erythrocytes at the rate proportional to the prevailing glucose concentration. Because the glycation reaction is irreversible, the formed HbA_{1c} remains in the circulation for the entire life of an erythrocyte. Its amount will reflect the average concentration of plasma glucose (Fig. 21.23) over approximately 8–12 weeks preceding its measurement. The exact time period it reflects is difficult to calculate precisely, because at any time plasma contains

populations of erythrocytes of different ages. Values such as 3–6 and 4–8 weeks have also been reported in the literature. Exposure to glucose over 30 days before measurement contributes approximately 50% to observed changes in HbA_{1c}.

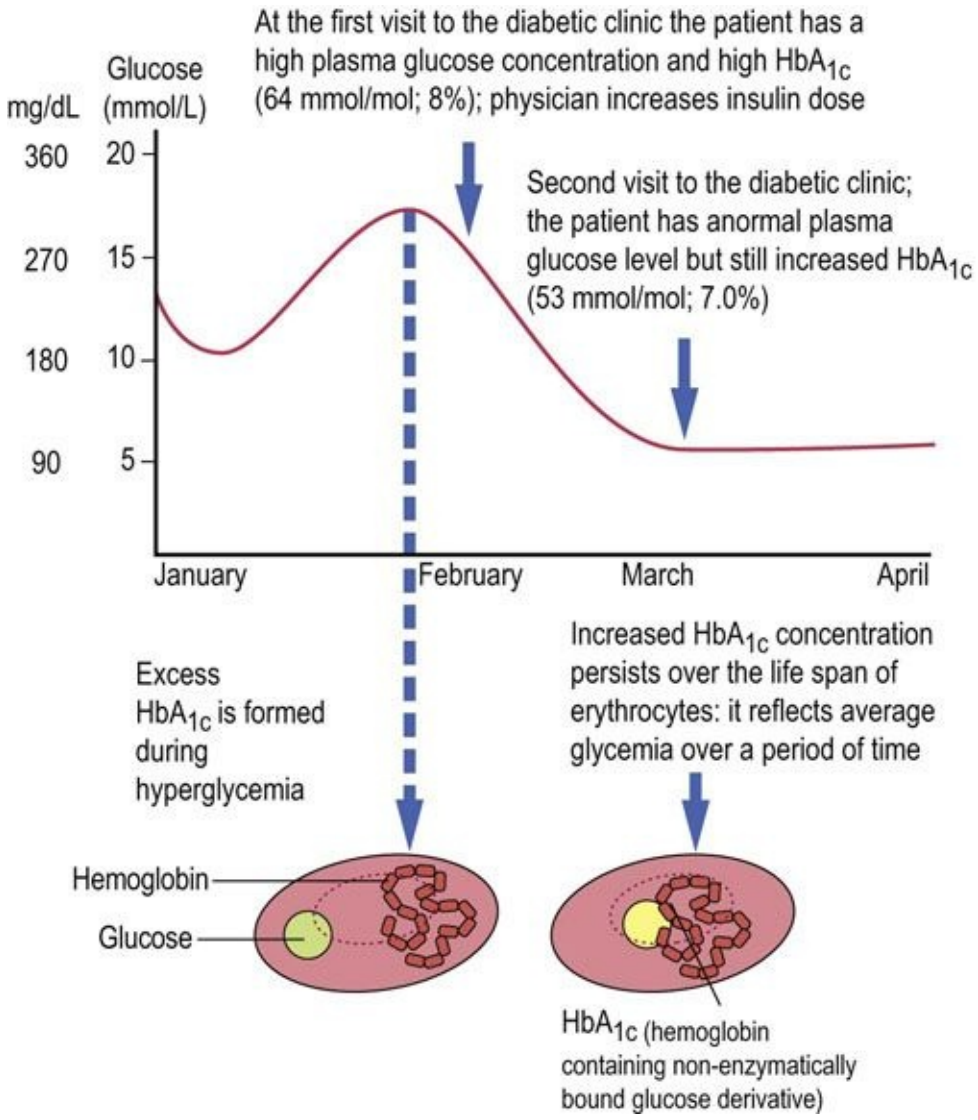


FIG. 21.23 Hemoglobin A_{1c}.
Hemoglobin A_{1c}(HbA_{1c}) is hemoglobin A post-translationally modified by a nonenzymatic glycation. The degree of glycation is proportional to hemoglobin's exposure to glucose during the life span of an erythrocyte. Measurements of HbA_{1c} are used for monitoring of the glycemic control and recently for diagnosis of diabetes. Note that the introduction of a new reference method for the HbA_{1c} resulted in the change of units from the traditional (%) to mmol/mol. The conversion formulas are available (Table

21.6). To obtain glucose concentrations in mg/dL, multiply by 18.

Originally, the HbA_{1c} was measured as a percentage of total hemoglobin. Recently, a reference method capable of measuring the absolute amount of HbA_{1c} has been introduced by the International Federation of Clinical Chemistry and Laboratory Medicine (IFCC). The new method is based on the cleavage of the *N*-terminal hexapeptide from the β-chain of HbA_{1c} by an endopeptidase, and subsequent separation and quantification by mass spectrometry or capillary electrophoresis. The units employed are mmol/mol. The values obtained by the two methods can be compared using a conversion formula (see Appendix 1). Such comparison is shown in Table 21.6. Note that HbA_{1c} concentration may be affected by anemia and the presence of hemoglobin variants.

HbA_{1c} values can be used to diagnose diabetes and to monitor glycemic control

The guidelines developed by ADA (2012) and WHO (2011) take the HbA_{1c} level of 48 mmol/L (6.5%) or above as diagnostic for diabetes. The HbA_{1c} values are also used in clinical practice to set targets for treatment. The normal concentration of HbA_{1c} is below 6%. ADA recommends that in a diabetic patient one should aim to achieve a concentration below 7% (53 mmol/mol). In some patients, particularly in young children and the elderly, this may be difficult because of the risk of hypoglycemia, and the goals must be adjusted to minimize such risk.

Clinical box A boy WHO did not like diabetes treatment

discrepant glucose and HbA_{1c} results

A 15-year-old insulin-dependent boy visited a diabetic clinic for a routine check-up. He told the doctor that he had followed all the dietary advice and never missed insulin injections. Although his random blood glucose was 6 mmol/L (108 mg/dL), HbA_{1c} concentration was 86 mmol/mol (11%) (adequate control: below 53 mmol/mol, 7%). He had no glycosuria or ketonuria.

Comment.

Blood and urine glucose results indicate good control of this boy's diabetes at the time of measurement, but the HbA_{1c} level suggests poor control over the last 3–6 weeks. The probability is that he only complied with treatment days before he was due to come to the clinic. This is not uncommon in adolescents, who find it hard to accept the necessity to adjust their lifestyle to, sometimes demanding, diabetes treatment. Measurement of HbA_{1c} identifies diabetic patients who do not comply with treatment.

Urine glucose is not a diagnostic test for diabetes

At a normal plasma concentration, glucose filtered through the renal glomeruli is reabsorbed in the proximal kidney tubules, and none appears in the urine (Chapter 23). The urinary threshold for glucose reabsorption is approximately 10.0 mmol/L (180 mg/dL). At higher concentrations, the reabsorptive capacity of the renal tubular transport system is exceeded, and glucose appears in the urine (this is known as glucosuria). Note that a healthy person may have a low renal glucose threshold and thus show glucosuria at nondiabetic blood glucose levels. Therefore, diabetes cannot be diagnosed on the basis of urine testing alone.

Ketone bodies in urine of a diabetic person signify metabolic decompensation

High concentration of ketones in urine (ketonuria) reflects high rate of lipolysis. Mild ketonuria may occur in healthy individuals during prolonged fasting or on a high-fat diet. However, in a diabetic patient, **ketonuria is an important sign of metabolic decompensation**, and requires the adjustment of treatment regimen.

Urinary albumin excretion is important in the assessment of diabetic nephropathy

The development of diabetic nephropathy can be predicted by detecting minute amounts of albumin in urine (microalbuminuria). To do this, laboratories employ a method that is more sensitive than the conventional one used for the measurement of serum albumin. The test is positive if more than 200 mg of albumin is excreted in urine over 24 h. Urine protein above 300 mg/day signifies

overt proteinuria. In diabetic patients plasma urea and creatinine concentrations are also routinely checked (Chapter 23).

Increased plasma lactate indicates inadequate oxygenation

The high plasma lactate level indicates increased anaerobic metabolism, and is a marker of inadequate tissue oxygenation (hypoxia; Chapter 5). In extreme situations, such as the cardiac arrest, this causes severe acidosis (the lactic acidosis). In diabetes, measurements of plasma lactate are important in rare instances of hyperglycemic nonketotic coma, a life-threatening condition where very high plasma glucose levels and extreme dehydration occur, but there is no ketoacidosis.

Diabetic patients need regular follow-up

During a periodic assessment of a diabetic patient, the physician would check blood glucose and HbA_{1c} concentrations to assess glycemic control. She would perform an eye examination (looking for signs of retinopathy) and neurologic examination (neuropathy). She would also measure urea and creatinine in plasma and microalbumin/protein in urine (to determine the presence, or assess the risk of, nephropathy), and measure plasma lipids (to check the blood pressure and assess the risk of cardiovascular disease; Chapter 18).

Treating diabetes

The control of glycemia prevents the development of diabetic complications

The goal of treatment in diabetes is **prevention of acute and chronic complications**. Maintaining good glycemic control is fundamental for diabetes care. Two major clinical trials, the Diabetic Control and Complications Trial (DCCT) in T1D and the UK Prospective Diabetes Study (UKPDS) in T2D confirmed that microvascular complications are associated with the severity of hyperglycemia (see Further Reading). There is also a strong evidence that therapeutic interventions that include both glycemic control and the management of cardiovascular risk are optimal for the prevention of long-term complications. Thus, in addition to striving to maintain the plasma glucose concentration close to normal, intensive management of cardiovascular risk factors such as hypertension and dyslipidemia is necessary.

Lifestyle modification is the mainstay of diabetes prevention and a key element of its treatment

Diet and exercise are the key lifestyle factors in the management of diabetes mellitus, and they underpin all drug treatments. They also are essential – and often undervalued – preventive measures. The Diabetes Prevention Program study demonstrated a 58% decrease in development of T2D after lifestyle interventions involving diet and exercise (see Further Reading). Unfortunately, glycemia can be controlled by lifestyle measures alone in less than 20% of all diabetic patients.

Patients with type 1 diabetes are treated with insulin

Insulin remains absolutely necessary for the treatment of T1D. Different preparations of insulin are available and they differ in the duration of their action. The ‘classic’ short-acting insulin is human regular insulin, the intermediate-acting are the Isophane and Lente insulins, and the long-acting is Ultralente. The newer analogs of human insulin are insulin lispro and insulin aspart (short-acting), and insulin detemir and insulin glargine (very long-acting).

Standard insulin treatment protocols involve daily subcutaneous injections throughout life

Patients usually take two subcutaneous injections of intermediate-acting insulin per day, or a mixture of a short-acting and intermediate-acting insulin. The newer so-called basal-bolus approach involves insulin glargine or detemir used as the basal component, and insulin lispro or aspart as boluses added before meals.

The greatest challenge of insulin treatment is replicating normal daily patterns of insulin secretion with insulin injections. Multiple injections of short-acting insulin are used in patients where glycemia is particularly difficult to control. Rarely, a constant insulin infusion is needed: this is delivered by a portable pump programmed to increase the delivery rate at meal times.

Emergency treatment of diabetic ketoacidosis includes intravenous insulin, rehydration and potassium supplementation

Emergency treatment of diabetic ketoacidosis addresses five issues: insulin lack, dehydration, potassium depletion, acidosis and the primary cause of decompensation. Insulin infusion is required to reverse metabolic effect of the excess of anti-insulin hormones, and fluids are infused to treat dehydration. The administered intravenous fluids normally contain potassium to prevent hypokalemia associated with insulin-caused shift of potassium into cells. Such treatment is usually sufficient to control the metabolic acidosis; however, when acidosis is severe, infusion of an alkalinizing solution (sodium bicarbonate) may also be required (see case described on p. 281). Primary cause such as, for instance, infection must also be intensively treated.

Clinical box Diabetic ketoacidosis affects the potassium balance

Insulin increases cellular potassium uptake and the lack of insulin leads to the release of potassium from cells. Since uncontrolled diabetes is also accompanied by osmotic diuresis, the released potassium is excreted in urine. As a result, **most patients admitted with ketoacidosis are potassium depleted** but, paradoxically, often have normal or raised plasma potassium concentration. When

exogenous insulin is given to such patients, it stimulates the entry of potassium into cells and can lead to very low plasma potassium levels (**hypokalemia**). Hypokalemia is dangerous, owing to its effects on the cardiac muscle. Thus, except for patients with very high plasma concentrations, potassium needs to be given during treatment of diabetic ketoacidosis. (See also Chapters 22 and 23.)

Patients with type 2 diabetes are treated with oral hypoglycemic drugs but may require insulin

T2D patients usually do not require insulin treatment because their insulin synthesis is at least partly preserved; they can be treated with hypoglycemic drugs. However, if adequate control cannot be achieved they do require insulin: each year 5–10% of patients treated with hypoglycemic drugs need to commence insulin treatment with insulin.

Advanced concept box Regulatory role of the PPAR-γ transcription factor

Peroxisome proliferator-activated receptors (PPARs) are ligand-activated nuclear receptors belonging to the steroid receptor family. Binding of a ligand induces conformational change which allows a PPAR to form a heterodimer with another receptor, retinoid X receptor (RXR). PPAR may also bind small molecules, coactivators or corepressors. The complex binds to response elements in gene promoters (Fig. 21.9).

There are three types of PPAR: **PPAR-α,** which is discussed in Chapter 18, **PPAR-γ** and **PPAR-β**. PPAR-γ is predominantly expressed in adipose tissue but also in muscle, liver, intestine and heart. It is activated by polyunsaturated fatty acids and by components of the oxidized LDL. It regulates carbohydrate and fatty acid metabolism, inducing, among others, genes coding for lipoprotein lipase (LPL), GLUT-4 glucose transporter and glucokinase. It also induces the ABCA1 transporter, increasing

transfer of cholesterol from cells to the HDL (Chapter 18). It inhibits macrophage activation and the production of cytokines such as tumor necrosis factor (TNF-α), interferon-γ and interleukin-1 (IL-1). PPAR-γ is a target of **thiazolidinediones**, drugs used in the treatment of T2D.

The antidiabetic drugs

The currently used oral hypoglycemic drugs target the three processes: insulin secretion, tissue insulin sensitivity, and absorption and digestion of carbohydrates.

Sulfonylureas, meglitinides and drugs affecting the incretin system stimulate insulin secretion

Sulfonylureas bind to a receptor in the plasma membrane of the pancreatic β-cells. The receptor contains the ATP-sensitive potassium channel. Binding of the drug closes the channel, depolarizes the membrane and opens the calcium channel. Increasing intracellular cytoplasmic calcium concentration stimulates the exocytosis of insulin. Hypoglycemia is a substantial side effect of sulfonylurea treatment.

Meglitinides target the K-ATP channel similarly to the sulfonylureas but bind to a different site.

Drugs affecting the incretin system: GLP-1 receptor agonists and DPP-4 inhibitors. GLP-1 receptor agonists such as exenatide or liraglutide increase insulin secretion. They act via the cyclic AMP–PKA pathway and potentiate insulin secretion induced by raised glucose. Endogenous GLP-1 can also be increased by using DPP-4 inhibitors such as sitagliptin, which prevents GLP-1 degradation and thus increases its effect.

Biguanides and thiazolidinediones sensitize the peripheral tissues to insulin

Metformin, a biguanide, is currently the most common oral treatment in T2D. It reduces hepatic gluconeogenesis, and decreases the endogenous glucose production. It suppresses the effects of glucagon and increases peripheral insulin

sensitivity. It also inhibits glycogenolysis by inhibiting the activity of the Glc-6-Pase. Metformin increases insulin-dependent glucose uptake in the skeletal muscle, and reduces fatty acid oxidation.

Thiazolidinediones, such as pioglitazone, act at the transcriptional level. They improve peripheral glucose utilization and insulin sensitivity. They are ligands of the PPAR-γ transcription factor in the adipose tissue and, to a lesser extent, in muscle. PPAR-γ activation increases transcription of a range of genes responsible for glucose and lipid metabolism such as the lipoprotein lipase, acyl-CoA synthase and the GLUT-4 transporter. Thiazolidinediones also activate the IRS-1/phosphoinositol kinase signaling pathway. They promote expansion of subcutaneous adipose tissue, reduce lipolysis and reduce the adipose tissue inflammation. However, they cause weight increase.

The glitazar group of drugs is currently being tested; these agents stimulate both PPAR-α and PPAR-γ and therefore, apart from thiazolidinedione-like action, they may influence lipid metabolism by raising the HDL and decreasing the plasma triglyceride concentration (Chapter 18).

Acarbose and SGLT2 inhibitors decrease intestinal glucose digestion

Acarbose is an inhibitor of intestinal α-glucosidase, which digests complex sugars. It delays intestinal absorption of glucose.

Amylin, an inhibitor of the SGLT2 glucose transporter, slows down gastric emptying, promotes satiety and inhibits glucagon secretion. It is used as an adjunct to insulin treatment.

Drug combinations used in type 2 diabetes

If glycemia cannot be controlled by diet and exercise, metformin is recommended in patients with T2D as a first-line therapy. Insulin is considered in those with severe symptoms and poor glycemic control. A second oral agent (e.g. thiazolidinedione or sulfonylurea), a GLP-1 receptor agonist, or insulin, may later be added to metformin treatment.

Summary

■ Glucose homeostasis involves the liver, adipose tissue, skeletal muscle and the pancreas.
■ The organism alternates between the fed and fasting state. Metabolite concentrations in blood change during the feed–fast cycle, and are influenced by stress and disease. Therefore, interpretation of metabolite levels needs to be related to meal times and to the overall clinical condition of the patient.
■ Measurement of plasma glucose concentration is part of a routine assessment of every patient admitted to hospital. The measurements of plasma glucose and glycated hemoglobin (HbA_{1c}) are used for diagnosis and monitoring of glycemic control. In diabetic patients, measurements of plasma and urine glucose, urinary ketones, HbA_{1c}, and tests of renal function, including microalbuminuria, are performed.
■ Type 1 diabetes mellitus is an autoimmune disease caused by destruction of pancreatic β-cells.
■ Type 2 diabetes results from the inability of functionally impaired β-cells to compensate for peripheral insulin resistance. Type 2 diabetes is strongly associated with obesity.
■ The short-term complications of diabetes include hypoglycemia and ketoacidosis. The long-term complications include diabetic retinopathy, nephropathy and neuropathy. Cardiovascular disease is linked to diabetic macroangiopathy. Cardiovascular disease is currently the main cause of death in diabetes.

Active learning

1. Describe how insulin causes an increase in cellular glucose uptake.
2. What are the anti-insulin hormones?
3. What is the role of the incretin system in glucose homeostasis?
3. Why would a nondiabetic patient brought to the emergency unit with extensive burns have an increased plasma glucose concentration? Describe her metabolic state.
4. You have asked a patient to come to the outpatient clinic to have plasma triglycerides tested. The patient asks whether he needs to be fasting that day. Please provide an answer and explain your

reasons.
5. Do people with impaired glucose tolerance develop long-term vascular complications?
6. What do obesity and diabetes mellitus have in common?

Further reading

Ahlqvist, E, Ahluvalia, TS, Groop, L. Genetics of type 2 diabetes. *Clin Chem*. 2011; 57:241–254.

American Diabetes Association. Standards of medical care in diabetes – 2012. *Diabetes Care*. 2012; 35(Suppl 1):S11–S63.

Bliss, M. *The discovery of insulin, Edinburgh*. Paul Harris Publishing; 1983.

Daneman, D. Type 1 diabetes. *Lancet*. 2006; 367:847–858.

Diabetes Control and Complications Trial (DCCT) Research Group. The effect of intensive treatment of diabetes on the development and progression of long-term complications in insulin-dependent diabetes mellitus. *N Engl J Med*. 1993; 329:977–986.

Diabetes Prevention Program Research Group. Reduction in the incidence of type 2 diabetes with lifestyle intervention or metformin. *N Engl J Med*. 2002; 24:387–388.

Dominiczak, MH. Obesity, glucose intolerance and diabetes and their links to cardiovascular disease. Implications for laboratory medicine. *CCLM*. 2003; 41:1266–1278.

Drucker, DJ, Nauck, MA. The incretin system: glucagon-like peptide-1 receptor agonists and dipeptidyl peptidase-4 inhibitors in type 2 diabetes. *Lancet*. 2006; 368:1696–1705.

Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. Report of the Expert Committee on the Diagnosis and Classification of Diabetes Mellitus. *Diabetes Care*. 2003; 26:S4–S20.

Jeffcoate, SL. Diabetes control and complications: the role of glycated hemoglobin, 25 years on. *Diabetic Med*. 2004; 21:6576–6665.

Nolan, C, Damm, P, Prentki, M. Type 2 diabetes across generations: from pathophysiology to prevention and management. *Lancet*. 2011; 378:169–171.

Steck, AK, Rewers, MJ. Genetics of type 1 diabetes. *Clin Chem*. 2011; 57:176–185.

Stern, MP. Diabetes and cardiovascular disease. The ‘common soil’ hypothesis. *Diabetes*. 1995; 44:369–374.

Stumvoll, M, Goldstein, BJ, von Haeften, TW. Type 2 diabetes: principles of pathogenesis and therapy. *Lancet*. 2005; 365:1333–1346.

Tuchman, A. Diabetes and the public's health. *Lancet*. 2009; 374:1140–1141.

UK Prospective Diabetes Study (UKPDS) Group. Intensive blood-glucose control with sulphonylureas or insulin compared with conventional treatment and risk of complications in patients with type 2 diabetes (UKPDS 33). *Lancet*. 1998; 352:837–853.

Wong, TY, Liew, G, Tapp, RJ, *et al*. Relation between fasting glucose and retinopathy for diagnosis of diabetes: three population-based cross-sectional studies. *Lancet*. 2008; 371:736–743.

Zimmet, P, Alberti, KGMM, Shaw, J. Global and societal implications of the diabetes epidemic. *Nature*. 2001; 414:782–787.

Websites and downloads

American Diabetes Association. Type 1 diabetes. www.diabetes.org/diabetes-basics/type-1.

American Diabetes Association. Type 2 diabetes. www.diabetes.org/diabetes-basics/type-2/?loc=DropDownDB-type2.

American Diabetes Association. Standards of Medical Care in Diabetes. http://care.diabetesjournals.org/content/36/Supplement_1/S11.full, 2013.

Diabetes.co.u. www.diabetes.co.uk/index.html

Diabetes, UK. Type 1 diabetes. http://www.diabetes.org.uk/Guide-to-diabetes/Type-1-diabetes/IDF Diabetes Atla. www.idf.org/diabetesatlas/5e/Update2012.

Diabetes UK blog. http://blogs.diabetes.org.uk

NHS choices. Diabetes type 2 – symptoms. www.nhs.uk/Conditions/Diabetes-type2/Pages/Symptoms.aspx.

WHO Consultation. Abbreviated Report. Use of glycated haemoglobin (HbA1c) in the diagnosis of diabetes mellitus. www.who.int/diabetes/publications/report-hba1c_2011.pdf.

CHAPTER 22

Nutrition and Energy Balance

Marek H. Dominiczak and Jennifer Logue

Learning objectives

After reading this chapter you should be able to:

- Describe mechanisms controlling food intake.
- Describe the role of AMP-activated kinase in maintaining cellular energy balance.
- Identify the main categories of nutrients and essential nutrients within these categories.
- Relate your knowledge of energy metabolism to current nutritional recommendations.
- Characterize malnutrition and obesity.
- Discuss nutritional assessment.

Introduction

From the biochemical point of view, nutrition is an essential interaction of the organism with the environment. Nutrition underpins health and affects susceptibility to disease; malnutrition and obesity are both associated with health risks.

Nutritional status is determined by biological, psychologic and social factors

Factors that determine nutritional status of an individual are the genetic background, the environment, the phase of the life cycle, the level of physical activity and the presence or absence of illness (Fig. 22.1). Nutritional status is also influenced by the availability of food, its palatability and variety. Nutritional deficiencies may result from dietary inadequacies or from genetically determined metabolic errors.

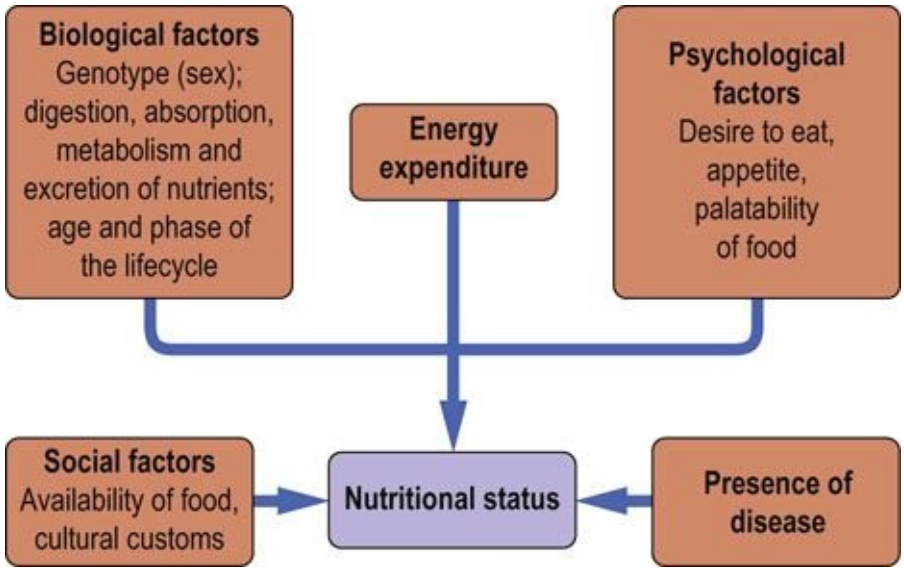


FIG. 22.1 Factors that determine the nutritional state.

Regulation of food intake

Food intake is controlled by hunger (a desire to eat) and appetite (a desire for a particular food)

The main centers regulating appetite are located in the hypothalamic arcuate and paraventricular nuclei in the central nervous system (CNS). In humans the arcuate nucleus area is known as the infundibular nucleus. The brain regulates energy homeostasis and is also the primary regulator of body weight (Fig. 22.2). Signals controlling energy intake originate from the adipose tissue and are sent to the central nervous system. These signals are mediated by the adipokine leptin and by insulin. In response, the brain sends efferent signals through a complex network of neuropeptides. These regulate appetite and hunger. The neurons in the arcuate nucleus express two neuropeptides: catabolic proopiomelanocortin (POMC), and anabolic neuropeptide Y (NPY). POMC is cleaved, yielding melanocortins such as α-MSH, that decrease food intake. On the other hand, NPY expression increases when adipose tissue is depleted, and when there is a decrease in leptin. NPY links to neurons expressing melanin-concentrating hormone (MCH) and orexins A and B. They, in turn, are involved in the control of food intake by acting on brainstem neurons. These neurons connect with the brain cortex (the satiety center) to promote hunger and to stimulate yet another set of hormones such as thyreoliberin (TRH), corticoliberin (CRH) and oxytocin. Thyreoliberin increases thermogenesis and food intake, whereas corticoliberin decreases food intake and, through the sympathetic activity, increases energy expenditure. Further signals that control food intake are conveyed by gastrointestinal peptides such as glucagon, cholecystokinin, glucagon-like peptide, amylin and peptide YY. Ghrelin, secreted by the stomach, stimulates NPY-expressing neurons. It is the only known appetite-stimulating peptide. Gastric stretch also affects food intake. Finally, hypoglycemia decreases the activity of the satiety center.

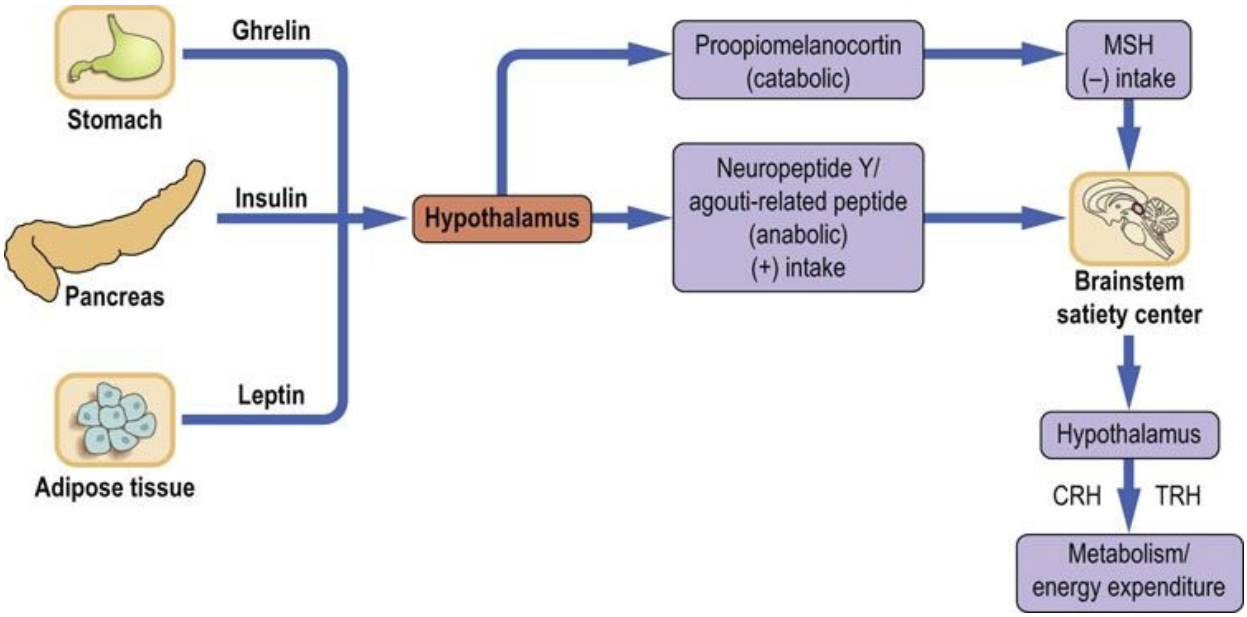


FIG. 22.2 Regulation of food intake.
Regulation of food intake is accomplished by signals generated in the adipose tissue, pancreas, stomach and the brain. The hypothalamus translates signals related to the energy balance into eating behavior through secretion of a range of neuropeptides. The (+) sign means action leading to increase of appetite and food intake, and the (−) sign to a decrease. CRH, corticoliberin; TRH, thyreoliberin; MSH, melanocyte-stimulating hormone.

The endogenous cannabinoid system

The hypothalamus and brainstem translate the information about energy balance into eating behavior. This involves the endogenous cannabinoid system. Endocannabinoids are compounds synthesized from membrane phospholipids. They include Δ^9-tetrahydrocannabinol and anandamide formed as a result of hydrolysis of *N*-arachidonylphosphatidyl ethanolamine by phospholipase D. Endocannabinoids are released at the synapses and bind to the synaptic receptors called CB1. The receptors are present in the central nervous system and also in gut, adipose tissue, liver, muscle and pancreas. They are coupled to G-proteins and adenylate cyclase, and also regulate potassium and calcium channels. The binding of endocannabinoids to the receptors modulates release of neurotransmitters such as GABA, noradrenaline, glutamate and serotonin. Hypothalamic levels of endocannabinoids increase during food deprivation.

Clinical box The ABC of emergency treatment

Eating and drinking, like breathing, link living organisms with the environment. To survive we need **oxygen, water and nutrients**. One can exist without oxygen for minutes only. Without water the survival time is days. With these two supplied, a human can survive without food for between 60 and 90 days.

Consequently, these considerations determine the urgency of treatment in critical situations. Re-establishment of oxygen supply and circulating volume is the first priority **(The ABC of resuscitation: Airway, Breathing, Circulation).** Repletion of lost fluid and electrolytes is also necessary within hours to days, depending on the state of the patient. Provision of other nutrients becomes important as soon as the life-saving measures have been taken. The rule of thumb is that patients unable to eat would need nutritional support if they have been (or will be) unable to take food for more than 7 days. This period is shorter in hypercatabolic persons.

Regulation of energy balance

Adipose tissue is an active endocrine organ

Adipose tissue, far from being an inert depot of storage fat, is an active endocrine organ (Fig. 22.3). Its products are known as adipokines. This endocrine activity influences development of obesity and conditions such as insulin resistance. The two major adipokines secreted by the adipose tissue are **leptin** and **adiponectin**.

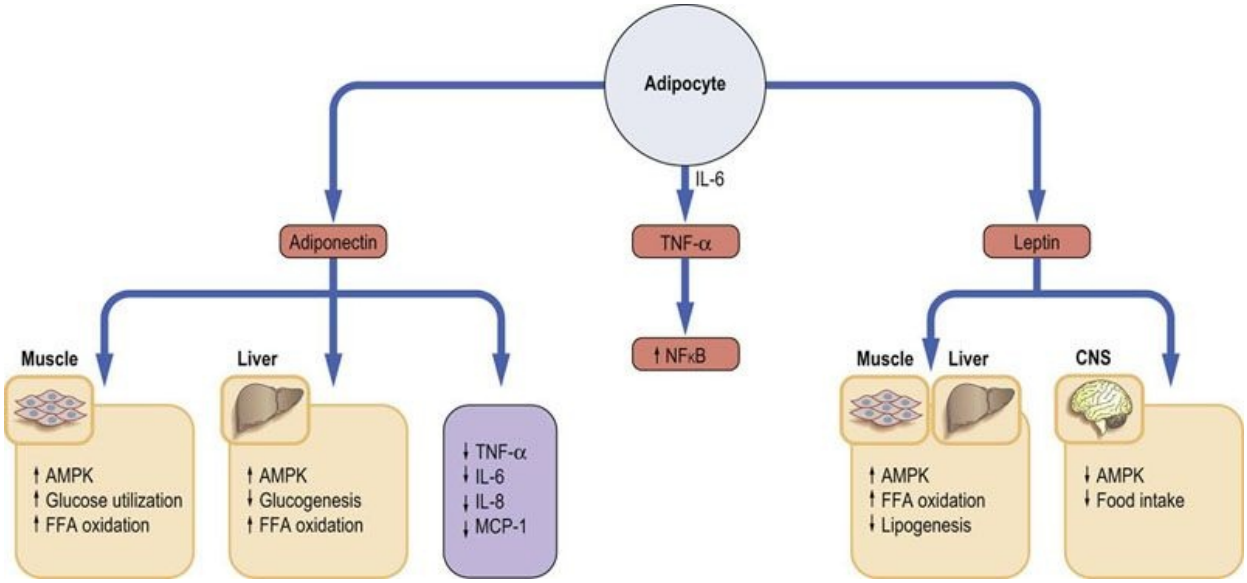


FIG. 22.3 Endocrine activity of the adipose tissue.
Adipose tissue produces adipokines. Leptin and adiponectin play an essential role in adjusting metabolism to body energy needs. They activate the energy sensor enzyme, AMP-activated kinase (AMPK). Adipocytes also secrete a range of proinflammatory cytokines such as TNF-α or IL-6. TNF-α is known to stimulate NFκB, a key transcription factor in the induction of genes involved in the inflammatory response. FFA, free fatty acids; CNS, central nervous system; IL, interleukin; MCP-1, monocyte chemoattractant protein 1; TNF-α, tumor necrosis factor α.

Leptin regulates adipose tissue mass and responds to the energy status

Leptin is a 16 kDa protein. Its secretion is linked to the adipose tissue mass and

to the size of adipocytes. Acting in the central nervous system it **decreases food intake**. It also acts on the skeletal muscle, liver, adipose tissue and pancreas. Leptin gene expression is regulated by food intake, energy status, hormones and inflammatory state. It affects metabolism by stimulating fatty acid oxidation and by decreasing lipogenesis. Importantly, it also decreases ectopic deposition of fat in liver or muscle.

Leptin signals through a membrane receptor that has an extracellular binding domain and intracellular tail. Its signaling pathways involve the Janus kinase and signal transducer and activator of transcription (JAK/STAT; Chapter 40). Mitogen-activated protein kinase and phosphatidyl-inositol 3′ kinase are also involved, as is the AMP-activated kinase (AMPK).

Adiponectin increases insulin sensitivity. Its lack leads to insulin resistance

Adiponectin is a 244 amino acid protein with a structural homology with collagens type VIII and X and with complement factor C1q. Several forms exist in the circulation: globular, trimeric and high molecular weight. Adiponectin stimulates glucose utilization in muscle and increases fatty acid oxidation in muscle and liver, thus increasing insulin sensitivity. It also decreases hepatic glucose production. Low adiponectin levels are linked with insulin resistance and with hepatic steatosis. Adiponectin downregulates secretion of proinflammatory cytokines interleukins 6 and 8 (IL-6 and IL-8) and monocyte chemoattractant protein-1 (MCP-1).

Physical training increases adiponectin expression and upregulates its receptors in the skeletal muscle. On the other hand, its concentration decreases in obesity and in type 2 diabetes. Low levels of adiponectin are also associated with low-grade inflammation, oxidative stress and endothelial dysfunction. Adiponectin receptors activate the AMPK, p38 mitogen-activated protein kinase, and PPARα which in turn regulates fatty acid metabolism (Chapter 21).

Adipose tissue secretes proinflammatory cytokines

Adipose tissue secretes the proinflammatory cytokines-tumor necrosis factor α (TNF-α) and IL-6. TNF-α is highly expressed in obese animals and humans, and it also induces insulin resistance and type 2 diabetes. TNF-α activates the proinflammatory NFκB pathway.

AMP-stimulated kinase (AMPK) responds to cellular energy levels

The AMPK is a serine-threonine kinase. It is a heterotrimer encoded by three genes: it has a catalytic subunit α and two regulatory subunits, β and γ. It is activated by phosphorylation by a kinase known as LKB1, which is a tumor suppressor molecule. The key activator of the AMPK is cellular accumulation of 5′-AMP and the increase the ratio of 5′-AMP/ATP. A high creatine/phosphocreatine ratio also activates the enzyme. High 5′-AMP concentration induces allosteric changes and the phosphorylation of the AMPK catalytic subunit. The AMP is generated in the myokinase (adenylate kinase) reaction:

$$\mathrm{ADP} + \mathrm{ADP} \rightleftharpoons \mathrm{ATP} + \mathrm{AMP}$$

AMPK stimulates energy-producing pathways and suppresses energy-utilizing ones

In the skeletal muscle, AMPK activation promotes glucose transport, glycolysis and fatty acid oxidation. On the other hand, It inhibits fatty acid synthesis by phosphorylating (and inactivating) acetyl-CoA carboxylase. This leads to a decrease in malonyl-CoA, and to disinhibition of carnitine palmitoyl transferase 1 and consequent facilitation of fatty acid transport into mitochondria. The increased fatty acid oxidation prevents tissue lipid accumulation. In the liver, AMPK inhibits lipogenesis and cholesterol synthesis (the latter by phosphorylation and inactivation of HMG-CoA reductase); its activation suppresses the sterol regulatory element-binding protein 1c (SREBP1c; Chapters 16 and 18).

In the skeletal muscle, AMPK is activated during exercise and is involved in contraction-stimulated glucose transport and fatty acid oxidation (exercise increases insulin sensitivity). In the heart it is activated by ischemia.

The AMPK also acts at the hypothalamic level. The expression of AMPK in the hypothalamus reduces food intake. Leptin and adiponectin stimulate AMPK in the skeletal muscle, liver and adipose tissue. Insulin and leptin inhibit it in the hypothalamus. The antidiabetic drugs such as thiazolidinediones and biguanides act in part through AMPK (Chapter 21). The effects of AMPK activation are summarized in Figure 22.4 (see also Box on p. 98).

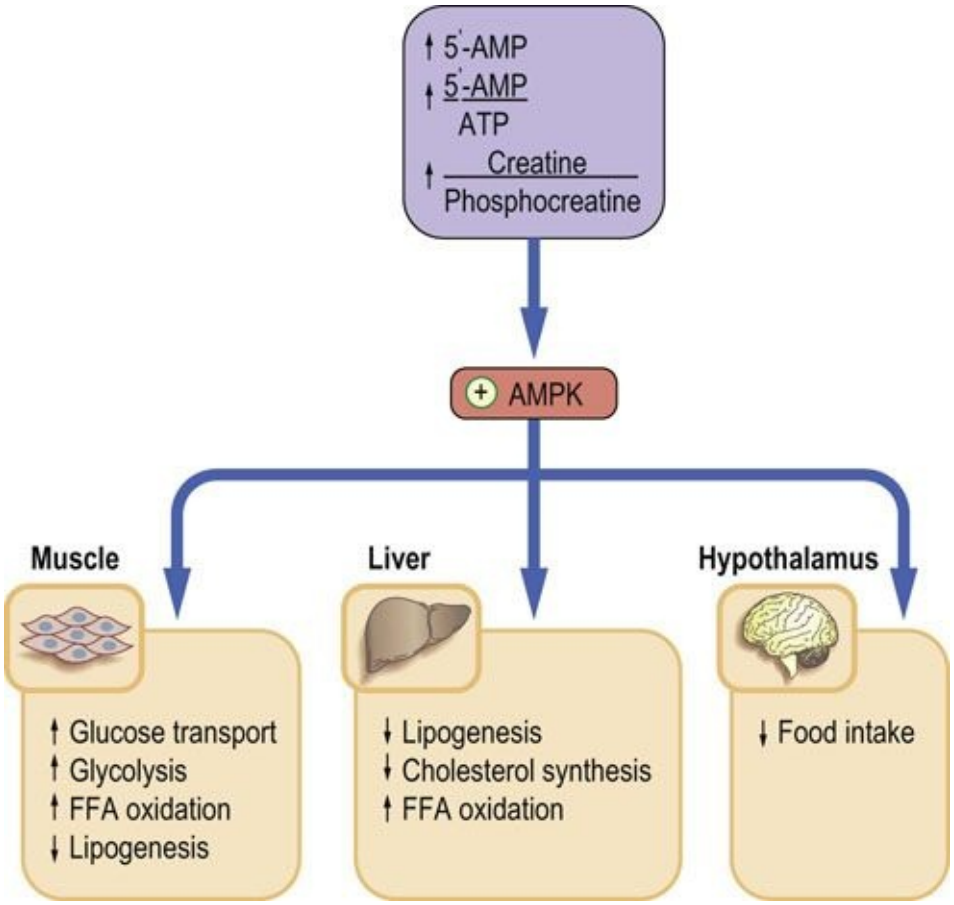


FIG. 22.4 AMP-activated kinase (AMPK).
AMPK fulfills the role of a sensor of cellular energy levels. Its main stimulator is the increase in 5′-AMP. An increased 5′-AMP/ATP ratio signals low cellular energy level. Activated AMPK phosphorylates rate-limiting enzymes in energy-producing and energy-using pathways. The final effect is inhibition of energy-utilizing pathways such as lipogenesis or cholesterol synthesis, and stimulation of pathways that generate energy: glycolysis and FFA oxidation. FFA, free fatty acids.

Energy expenditure

Total daily energy expenditure is a sum of the basal metabolic rate, the thermic effect of food and the energy used up during physical activity.

Energy expenditure can be measured by direct calorimetry, which relies on measurements of heat production. Indirect calorimetry is based on the measurement of the oxygen consumption rate (VO_2). The ratio of VCO_2 to VO_2 is known as the respiratory exchange rate (RER) or respiratory quotient. For carbohydrates, RER = 1; for fat, RER = 0.7.

Basal metabolic rate (BMR) is energy expenditure required

to maintain body function at complete rest

The BMR depends on sex, age and body weight. At rest, energy is required for membrane transport (30% of the total), protein synthesis and degradation (30%), and for maintaining temperature, physical activity and growth. Certain organs use particularly high amounts of energy: in a 70-kg person, brain metabolism constitutes approximately 20% of basal metabolic demand, liver 25% and muscle 25%. On the other hand, in very low birth weight babies, the brain is responsible for as much as 60% of the BMR, liver for 20% and muscle for only 5%.

In health, physical activity is the most important changeable component of energy expenditure

The level of activity is normally expressed as metabolic equivalents of task (METs). METs use a reference value for resting metabolism of 1 kcal/kg/h and activities are multiples of these. As both at rest and during activity the metabolism is dependent on the mass of the individual, which allows the intensity and energy expenditure to be compared between people of different weight (the ratio of energy expenditure during the activity to resting metabolism will remain unchanged). Examples of energy expenditure associated with different activities are given in Table 22.1. Energy requirement also depends on sex and age (Table 22.2).

Table 22.1

Energy expenditure

Physical activity ratio	Example of activity
1.3	Watching TV, reading, writing
2.0	Dressing and undressing, making beds, walking slowly
2.3	Washing dishes, ironing
2.5	Dusting and cleaning, cooking
4.5	Cleaning windows, golf, carpentry
6.5	Jogging, digging and shoveling
8.0	Climbing stairs, cycling, football, skiing

Energy expenditure is expressed as metabolic equivalents of task (METs, *i.e.* multiples of the expenditure at complete rest.).

Table 22.2

Daily energy requirements

EAR, kcal/day (mJ)		
Age	**Males**	**Females**
0–3 months	545 (2.28)	515 (2.16)
4–6 months	690 (2.89)	645 (2.69)
4–6 years	1715 (7.16)	1545 (6.46)
15–18 years	2755 (11.51)	2110 (8.83)
19–50 years	2550 (10.60)	1940 (8.10)
75+ years	2100 (8.77)	1810 (7.61)

Estimated average requirements (EAR) for energy for selected age and sex groups. Data from Dietary Reference Values for Food Energy and Nutrients for the United Kingdom; Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy, London: TSO 2003.

Nutrigenomics

Individual response to nutrients is to a substantial extent determined by genetics. Genes influence digestion and absorption of nutrients, their metabolism and excretion. Perceptions such as taste or satiety are also, to an extent, genetically determined. This has consequences for nutritional guidelines: because the gene pool varies between populations, optimal nutritional guidelines should be population-specific, rather than general. Nutrigenomics has large potential implications for future nutritional interventions. It is analogous to pharmacogenomics: it aims to exploit the knowledge accumulated by the Human Genome Project, and the ability to monitor the expression of a large number of genes, to devise individual dietary treatments customized to genetic background. Metabolomics, the monitoring of metabolic response patterns to nutrients, offers further opportunities to determine individual nutrition profiles (Chapter 36).

Genotype influences plasma concentrations of nutrients

An example of the genotype effect on nutrient intake is the response of plasma cholesterol concentration to its dietary content. Approximately 50% of individual variation in plasma cholesterol is genetically determined. Response to a cholesterol-containing diet is associated with the apoprotein E (apoE) genotype. ApoE is a protein synthesized in the liver, and is the main metabolic driver of the lipoprotein remnant particles (Chapter 18). It exists in several isoforms coded by alleles designated ε2, ε3 and ε4. It has been observed that plasma cholesterol concentration increases on low-fat/high-cholesterol diet in people with the E4/4 but not E2/2 phenotype.

There are many examples of nutrients affecting gene expression. For instance, the activities of key hepatic enzymes differ in persons on a long-term high-fat diet compared to a high-carbohydrate diet. The amount of dietary cholesterol affects the activity of HMG-CoA reductase. Polyunsaturated fatty acids inhibit the expression of fatty acid synthase and the ω-3 fatty acids (see below) reduce the synthesis of mRNA coding for the platelet-derived growth factor (PDGF) and the inflammatory cytokine interleukin 1 (IL-1). In essential hypertension, sensitivity to dietary salt is controlled, at least to an extent, by the angiotensinogen gene variants. Only 50% of patients are sensitive to salt intake: 30–60% of blood pressure variation is genotype-related.

Genetic regulation of food intake and energy expenditure

Obesity is concordant in 74% between monozygotic, and 32% between dizygotic twins. The current obesity epidemic has prompted a search for genes which control energy expenditure and food intake. Large population studies have been used to look for a relationship between changes in genes and body weight. As there is a close relationship between the body mass of members of the same family, and especially identical twins, it is estimated that 40–70% of the differences in the predisposition to obesity can be explained by genetics.

Changes within two genes have been found to be associated with obesity: fat mass and obesity-associated protein (FTO) and melanocortin-4 receptor (MC4R). FTO is expressed in the arcuate nucleus in response to hunger, and is therefore thought to have an effect on body mass via the control of food intake. MC4R is a ligand for melanocortins in the arcuate nucleus that control food intake. These effects have been shown in a study where participants who had variations in their FTO gene which were associated with obesity picked food with a higher energy content from a lunch buffet, compared to participants who did not have these genetic variants. However, in the general population, variation in FTO and regions near MCR4 genes only account for an increase in body mass of 0.39 kg/m^2 and 0.23 kg/m^2, respectively, so go no way towards explaining the current obesity epidemic. No genes associated with energy expenditure have been found to date.

Nutrition, life cycle and metabolic adaptation

The demand for nutrients is affected by both physiology and disease. **Pregnancy, lactation, and growth** (in particular the intensive growth in utero, growth during infancy and the adolescent growth spurt) are the three most important physiologic states associated with increased demand for nutrients.

Pregnancy is an example of metabolic adaptation termed expansive adaptation

Here the body of the mother adapts to carrying the fetus and supplying it with nutrients. Around the time of conception the mother's body prepares for the metabolic demands of the fetus. In early pregnancy the mother sets up the ‘supply capacity’ and later in pregnancy such supply takes place. Ninety percent of fetal weight is gained between the 20th and 40th weeks of pregnancy and the

steepest growth is established between the 24th and 36th weeks. The total amount of energy stored during pregnancy is about 70,000 kcal (293,090 kJ), amounting to approx 10 kg of weight.

Nutrient intake changes during the life cycle

After delivery there is a transition from feeding through the placenta to breastfeeding and then, gradually, the baby adapts to a free diet. Up to the **breastfeeding stage**, nutrition is controlled by substrates and the infant is entirely dependent on the mother for nutrition. Later, the growth hormone assumes a major role in directing development. At **school age**, new eating and activity patterns emerge as a child learns to be independent from its parents. This continues during **adolescence.** At this stage, sex hormones begin to play a prominent developmental role. In **adulthood,** muscle mass increases between 20 and 30 years of age and at that point the level of physical activity stabilizes. Thereafter, muscle mass starts to decline and the fat mass starts to increase. This accelerates after the age of 60. The bone mass also declines with age.

When nutrients are in short supply, either because of increased nutritional need or reduced availability of food, the so-called **reductive adaptation** takes place: the metabolic rate falls and the desire to eat decreases. This limits weight loss.

Main classes of nutrients

The main nutrients are **carbohydrates** (including fiber), **fats, proteins, minerals and vitamins**. Carbohydrates, proteins, fat, fiber and some minerals are macronutrients. Vitamins and trace metals are **micronutrients** (Chapter 11). Caloric value of the main nutrients is given in Table 22.3. The functions of nutrients are summarized in Figure 22.5.

Table 22.3

Caloric content of nutrients

	Energy	
Nutrient	**kJ/g**	**kcal/g**
Starch	17	4
Glucose	17	4
Fat	37	9
Protein	17	4
Alcohol	30	7

Caloric content of nutrients (1 kJ = 239 cal; 1 kcal = 4.184 kJ).

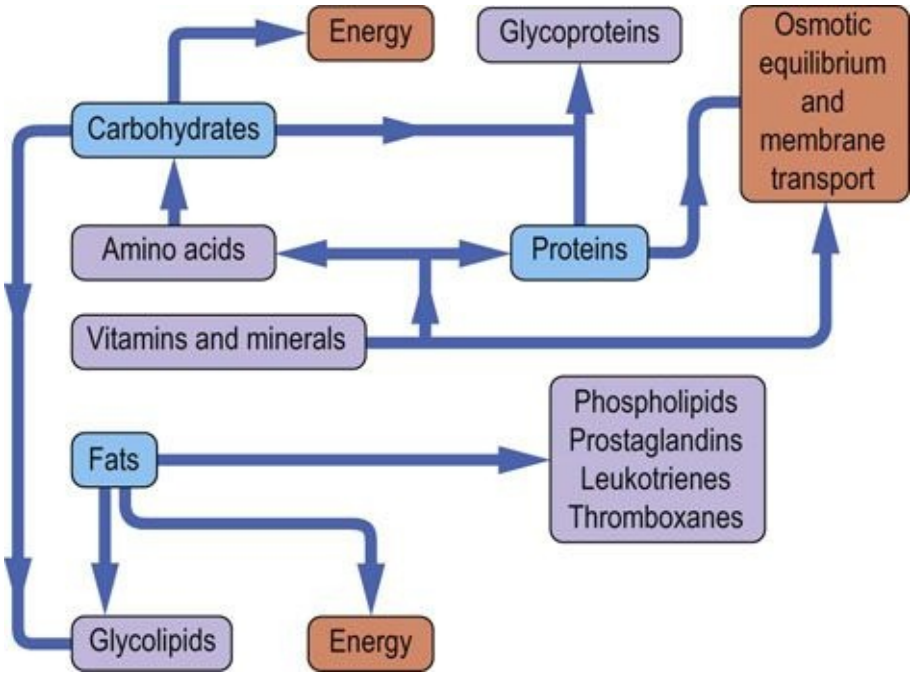


FIG. 22.5 The functions of nutrients.
All main classes of nutrients can be used to produce energy, and all contribute to the synthesis of more complex compounds. The main role of vitamins and minerals is participation in enzymatic reactions either as cofactors or components of enzyme prosthetic groups.

Carbohydrates

Carbohydrates and fats are the most important **energy sources**. Carbohydrates also are precursors of glycoproteins, glycolipids, and mucopolysaccharides. Dietary carbohydrates include refined carbohydrates such as sucrose in sweets, drinks and fruit juices, and complex carbohydrates such as starch, present in grains and potatoes. **Fiber** is carbohydrate which is indigestible by the human gut, such as cellulose, hemicellulose, lignin, pectin, and β-glucan. Fiber is present in unprocessed cereals, legumes, vegetables, and fruits. Its main role is to regulate gut motility and transit.

Glycemic index and glycemic load provide quantitative and qualitative insight into handling of carbohydrate-containing foods

The glycemic index (GI) is the system of ranking of the carbohydrate-containing foods according to the degree of increase in blood glucose which takes place

after their ingestion. Procedure that underpins the ranking is similar to the oral glucose tolerance test. The effect of a standard dose (25 g or 50 g) of a particular food on plasma glucose concentration is tested and compared to the reference nutrient (e.g. glucose). The comparison is based on the ratio of the area under curve (AUC) for the tested nutrient and glucose.

$$\text{GI} = (\text{AUC tested nutrient} / \text{AUC glucose}) \times 100 \quad \textbf{(1)}$$

The GI is expressed on a scale of 1 to 100 (low GI is 0–55, moderate 56–69, and high >69). Foods that are rapidly absorbed and digested have a high GI. Slower absorption and digestion yield low GI. The GI is affected by the nature of food, and the type of starch, but also by the cooking method (i.e. the GI of lightly cooked spaghetti will have lower GI lower than that cooked for a longer period). The **low GI foods** control postprandial glycemia and insulinemia, and are beneficial for people with diabetes, and better for weight control. Note that low GI foods tend to be high in fat, and low in carbohydrate and fiber.

The derivative of the GI is the **glycemic load**. It translates the qualitative information contained in the GI into qualitative, which can be used to calculate the carbohydrate content of a given portion of food.

$$\text{GL} = \text{GI} / 100 \times \text{CHO (grams per serving)} \quad \textbf{(2)}$$

Proteins

Dietary proteins are digested to their component amino acids. **Amino acids** are then used as a **material to build the host's own proteins**. They also serve as the ‘last resort’ energy substrate: the catabolic state is typically associated with release of muscle amino acids, and thus **muscle wasting**. Due to the different composition of animal and plant proteins, eating no animal products at all may lead to nutrient deficiencies such as those of vitamin B_{12}, calcium, iron, and zinc. Similarly to energy requirements, protein requirements change during the life cycle (Table 22.4). Increased demand for dietary protein is associated with pregnancy, lactation and the adolescent growth spurt.

Table 22.4

Daily protein requirements

Age	g/day males	g/day females
0–3 months	12.5	12.5
10–12 months	14.9	14.9
4–6 years	19.7	19.7
15–18 years	55.2	45
19–50 years	55.5	45
>50 years	53.3	46.5

Protein requirements are age and sex dependent. Here reference nutrient intakes (RNI) for selected age groups are shown. Data from Dietary Reference Values for Food Energy and Nutrients for the United Kingdom; Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy, London: TSO 2003.

Fats

Fats are the most important nutrients used for **energy storage**. Lipids also provide thermal insulation for the organism. They are essential components of biological membranes and serve as substrates for synthesis of glycolipids and glycoproteins, and also phospholipids, prostaglandins, leukotrienes, and thromboxane. They are particularly important for the development of the brain and the retina (see also Table 3.2).

Fatty acids also serve as **signaling molecules** and are activating ligands for transcription factors. They, for instance, stimulate accumulation of diacylglycerol (DAG), activate the δ isoform of protein kinase C (PKC) in the liver, and decrease tyrosine phosphorylation of insulin receptor substrate proteins 1 and 2 (IRS-1 and 2). Free fatty acids (FFA) also stimulate the NFκB pathway.

Long-chain fatty acids are not soluble in water but short- (C4–6) and medium-chain fatty acids (C8–10) are. Short-and medium-chain fatty acids are transported in plasma bound to albumin rather than in chylomicrons.

Fats are classified into saturated and unsaturated (the latter are either mono-or polyunsaturated). The most common saturated fatty acid is palmitic acid (C16)

Others are stearic (C18), myristic (C14), and lauric (C12). All animal fats (beef fat, butterfat, lard) are highly saturated. Saturated fats are also present in palm oil, cocoa butter and coconut oil.

Oleic acid (ω-9) is the only significant dietary monounsaturated fatty acid

Monounsaturated fatty acids are present in all animal and vegetable fats. Olive oil is a particularly rich source of monounsaturated fats. Monounsaturated *trans* fatty acids (Fig. 22.6), the isomers of the *cis*-oleic acid, are byproducts of the hydrogenation process of liquid vegetable oils. *Trans* fatty acids arising from hydrogenation of fats are associated with increased risk of coronary disease.

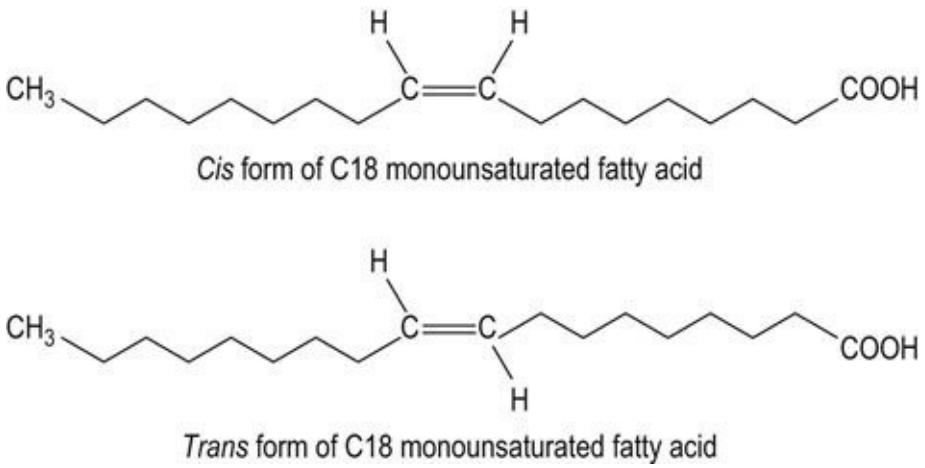


FIG. 22.6 The example of *cis*- and *trans*-monounsaturated fatty acid (18-carbon oleic acid).
Trans fatty acids are produced during hydrogenation of liquid vegetable oils.

Polyunsaturated fatty acids include ω-3 and ω-6 acids

The ω-3 fatty acids are α-linolenic (ω-3, C-18:3, $\Delta^{9,12,15}$), eicosapentaenoic (ω -3, C-20:5, $\Delta^{5,8,11,14,17}$), and docosahexaenoic (ω-3, C-22:6, $\Delta^{4,7,10,13,16,19}$) acids. They are present primarily in fish, shellfish, and phytoplankton and also in some

vegetable oils such as olive, safflower, corn, sunflower, and soybean and leafy vegetables.

The ω-6 acids are arachidonic acid (ω-6, C-20:4, $\Delta^{5,8,11,14}$) and its precursor linoleic acid (ω-6, C-18:2, $\Delta^{9,12}$). The ω-6 fatty acids are present in soybean and canola oils and in fish oils (particularly in fatty fish such as salmon, sardines, and pilchards).

Definitions in nutrition science

Diet is the total of all the foods and drinks ingested by an individual. The **food or foodstuff** is the particular food that is ingested. **Nutrients** are chemically defined components of food required by the body.

Dietary intake is not easy to assess

Available data are based on (sometimes incomplete) population surveys. Sets of values derived from these describe suggested minimal, average, and adequate intakes of particular nutrients. Different values are used in different countries, and there has been a degree of confusion and overlap between various definitions (see Box on p. 305). Currently, the estimates of nutrient intake are based on the dietary reference intakes (DRI), developed by the Food and Nutrition Board (FNB) at the Institute of Medicine (IOM) of the National Academies in the USA. They constitute the sets of values that describe the intake of a given nutrient in a population (Fig. 22.7).

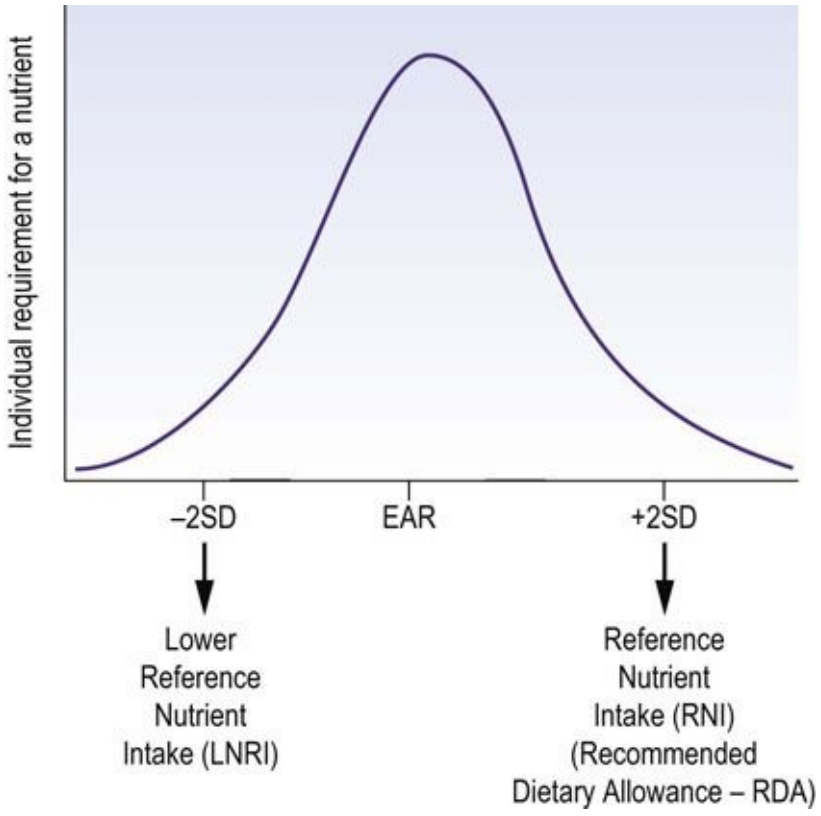


FIG. 22.7 Dietary reference intakes (RDA): the component values.
While the Estimated Average Requirement (EAR) reflects the intake adequate for half of a population, the RNI (or RDA) are values representing intake adequate for the great majority of individuals. SD:standard deviation.

Clinical box A woman with renal failure and weight loss

A 59-year-old woman was admitted to the renal unit with recurrent infections and recurrent endocarditis. She had been treated with hemodialysis for 5 years and over a period of 2 years had lost 33% of her body weight. She had a very poor appetite and had been taking two cartons (1.5 kcal/mL and 6.2 kJ/mL) of milkshake sip feeds daily. She was anuric and her fluid intake was restricted to 1000 mL daily. Her height was 1.68 m and she weighted 52.7 kg; BMI was 18. Her most recent biochemistry results revealed persistent hyperkalemia between 5.7 and 6.2 mmol/L.

Comment.

Because of her poor nutritional status and continuous weight loss she needs to continue taking nutritional supplements to minimize further weight loss. However, due to the hyperkalemia, a low-electrolyte sip feed should be considered. She was initially taking 22 mmol of potassium from the 1.5 kcal/mL milkshake. Subsequently this was changed to a low-electrolyte sip feed, which provided 12 mmol in total. The UK Renal Nutrition Group Standards recommend that daily potassium intake in this type of patient should be less than 1 mmol/kg body weight. Low-electrolyte/volume sip feeds should always be considered for renal patients and the serum electrolyte concentrations need to be closely monitored.

Essential (limiting) nutrients

Essential nutrients cannot be synthesized in the human body and therefore need to be supplied from outside. They include:

- essential amino acids
- essential fatty acids (EFA)
- some vitamins and trace elements.

Note that carbohydrates are not essential nutrients.

Some plant proteins are relatively deficient in essential amino acids, whereas animal proteins usually contain a balanced mixture

Essential amino acids are phenylalanine (tyrosine can be synthesized from phenylalanine), the branched-chain amino acids valine, leucine, isoleucine, threonine, and methionine and lysine (Chapter 19).

The essential fatty acids (EFA) are linoleic acid and α-linolenic acid

Arachidonic acid, eicosapentaenoic acid, and docosahexaenoic acid can be made in limited amounts from EFA; however, they become essential nutrients when EFA are deficient.

Vitamins and trace metals are important for the catalysis of chemical reactions

They act as coenzymes and form functionally important prosthetic groups of enzymes. They are discussed in detail in Chapter 11.

Assessing nutritional STATUS

Such assessment includes dietary habits and dietary history, a range of body (anthropometric) measurements, and biochemical and hematologic laboratory tests. Since there is no single definitive marker of nutritional status, assessment relies on the interpretation of a range of variables.

Dietary history should include more than the details of food intake

Dietary habits include meal patterns and the amount and composition of food. Individual diet is determined by biological, psychologic, sociologic and cultural factors. Biological factors involved are the state of the systems responsible for the intake, digestion, absorption and metabolism of nutrients. Enzyme deficiencies, such as, for instance, that of lactase (Chapter 10), cause impaired absorption of foodstuffs (in this case, milk).

Taking dietary history

Psychologic factors play an important role in determining food intake; eating disorders such as anorexia nervosa and bulimia nervosa may lead to severe malnutrition. Sociologic factors include availability and price of food, and measures taken by society to improve diets such as, for instance, school meals or subsidized meals for the elderly or disabled persons. Cultural factors also determine eating patterns and the type of preferred foodstuffs. All the above are important when taking the nutritional/dietary history. Individual food intake can be assessed by food frequency questionnaires, 24-h dietary recalls, food records, and also by direct analysis of foods and by metabolic balance studies. The newer method of diet research is the so-called structured assessment of dietary patterns.

Body weight and the body mass index

Body weight in relation to height is the most commonly used measurement in nutritional assessment. Relationship between the two is expressed as the body mass index (BMI), calculated according to the formula:

$$\text{BMI} = \text{weight (kg)} / \text{height (cm)}^2 \quad \textbf{(3)}$$

The BMI is used to categorize the nutritional status, as shown in Table 22.5. As mentioned above, it should not be used as a sole/definitive indication of the nutritional status.

Table 22.5

Body mass index (BMI) and nutritional status

BMI (kg/m^2)	Interpretation
<18.5	Malnourished
18.5–20	Underweight
21–25	Desirable
26–30	Overweight
>30	Obese

Other measurements used in nutritional assessment are the **waist-to-hip ratio**, the **mid-arm circumference** and the **skinfold thickness** measured with carefully calibrated calipers. The simple measurement of waist circumference correlates with the amount of visceral fat and is used in the diagnosis of the metabolic syndrome (Chapter 21). More detailed analysis includes assessment of total body water, analysis of body bioelectrical impedance and the measurements of lean body mass using dual-energy X-ray absorptiometry (DEXA). Some of these measurements allow calculation of variables such as body fat content and composition. Functional measurements such as grip strength or peak expiratory flow are also relevant to the nutritional assessment.

Biochemical markers of nutritional status

Urinary nitrogen excretion helps to assess nitrogen balance

Nitrogen balance relates to body protein requirements. It is a difference between the intake of nitrogen and its excretion. Positive nitrogen balance means that the intake exceeds loss. Negative nitrogen balance signifies that the loss exceeds intake. The 24-h urinary nitrogen excretion is an estimate of the quantity of proteins metabolized by the body. Ninety percent of the excreted nitrogen appears in the urine (80% of this as urea). The rest is excreted in the stool, hair, and sweat. Nitrogen excretion adjusts to protein intake over 2–4 days. Measurement of urinary nitrogen (or urea) excretion is the most reliable way of

assessing daily protein requirements. Currently, it is rarely used outside the research setting. As a guide, most people require 1–1.2 g protein/kg body weight/day. Age-related protein requirements are listed in Table 22.4.

Specific plasma proteins have been used as markers of nutritional status

Concentration of a protein in plasma may reflect the nutritional status over the time period related to its half-life. Proteins most commonly used for this purpose are **albumin** and **transthyretin** (prealbumin). Many studies confirm the link between liver albumin synthesis (albumin half-life is approximately 20 days (Chapter 4) and the nutritional status. Transthyretin, which has a half-life of 2 days, has also been used in nutritional assessment. It is synthesized in the liver and forms a complex with retinal-binding protein in plasma.

Unfortunately, interpretation of plasma concentrations of nutritionally relevant proteins is often difficult, because they are not exclusively determined by the state of nutrition. For instance, albumin concentration in plasma also depends on the state of **hydration**; it decreases in overhydrated patients (Chapter 24). In addition, albumin and transthyretin are affected by the **acute phase response** (transthyretin concentration increases during the acute phase reaction and albumin concentration decreases). Thus, in critically ill patients, albumin is not a useful marker of nutritional state.

General laboratory tests provide information supplementing nutritional assessment

Measurement of hemoglobin may uncover iron deficiency. Checking the liver (Chapter 30) and kidney (Chapter 23) function, the measurements of plasma sodium, potassium, chloride, bicarbonate, calcium, phosphate and magnesium, and the assessment of iron metabolism, all provide useful additional information. Assessing daily fluid intake and loss (Chapter 24) is essential in patients who are being considered for nutritional support.

Full assessment involves measurements of vitamins and trace metals

This is particularly important in patients who remain on long-term parenteral nutrition.

Simplified assessment of nutritional status

The Malnutrition Universal Screening Tool (MUST) has been introduced by the British Association for Parenteral and Enteral Nutrition (BAPEN) to enable fast assessment of nutritional state in adults. It is a five-step method of identifying individuals at risk (Fig 22.8).

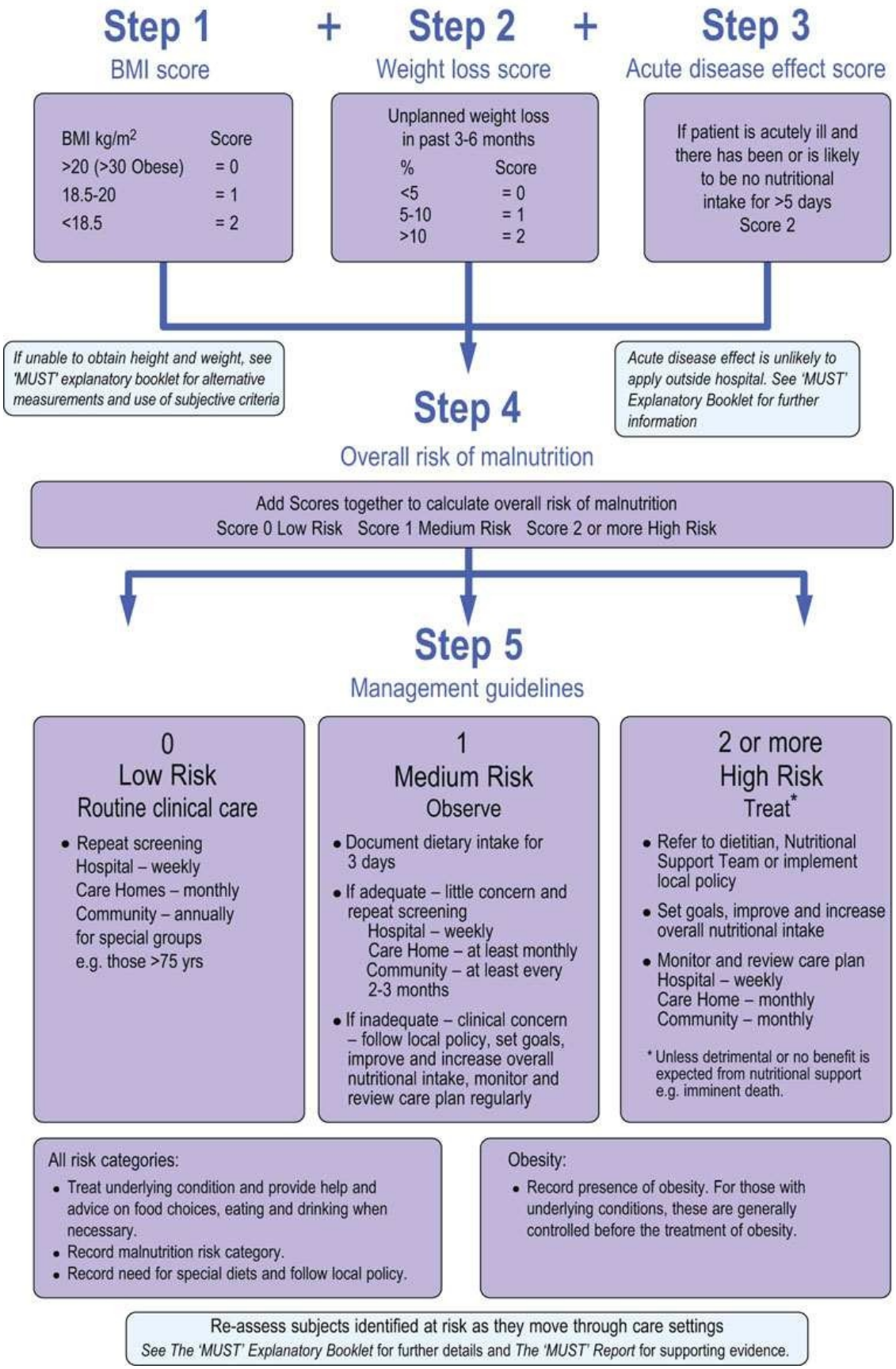

Step 1
BMI score
+
Step 2
Weight loss score
+
Step 3
Acute disease effect score
BMI kg/m2 Score
>20 (>30 Obese) = 0
18.5-20 = 1
<18.5 = 2
Unplanned weight loss in past 3-6 months
% Score
<5 = 0
5-10 = 1
>10 = 2
If patient is acutely ill and there has been or is likely to be no nutritional intake for >5 days
Score 2
If unable to obtain height and weight, see 'MUST' explanatory booklet for alternative measurements and use of subjective criteria
Acute disease effect is unlikely to apply outside hospital. See 'MUST' Explanatory Booklet for further information
Step 4
Overall risk of malnutrition
Add Scores together to calculate overall risk of malnutrition
Score 0 Low Risk Score 1 Medium Risk Score 2 or more High Risk
Step 5
Management guidelines
0
Low Risk
Routine clinical care
• Repeat screening
Hospital – weekly
Care Homes – monthly
Community – annually
for special groups
e.g. those >75 yrs
1
Medium Risk
Observe
• Document dietary intake for 3 days
• If adequate – little concern and repeat screening
Hospital – weekly
Care Home – at least monthly
Community – at least every 2-3 months
• If inadequate – clinical concern – follow local policy, set goals, improve and increase overall nutritional intake, monitor and review care plan regularly
2 or more
High Risk
Treat*
• Refer to dietitian, Nutritional Support Team or implement local policy
• Set goals, improve and increase overall nutritional intake
• Monitor and review care plan
Hospital – weekly
Care Home – monthly
Community – monthly
* Unless detrimental or no benefit is expected from nutritional support e.g. imminent death.
All risk categories:
• Treat underlying condition and provide help and advice on food choices, eating and drinking when necessary.
• Record malnutrition risk category.
• Record need for special diets and follow local policy.
Obesity:
• Record presence of obesity. For those with underlying conditions, these are generally controlled before the treatment of obesity.
Re-assess subjects identified at risk as they move through care settings
See The 'MUST' Explanatory Booklet for further details and The 'MUST' Report for supporting evidence.

FIG. 22.8 Malnutrition Universal Screening Tool (MUST) developed by the British Association for Enteral and Parenteral Nutrition.
Steps involved in nutritional assessment. (Reproduced with permission from www.Bapen.org.uk). Refer to the BAPEN website for further information.

Malnutrition

Malnutrition is a gradual decline in nutritional status, which in its more advanced stages leads to a decrease in functional capacity and to other complications

Protein energy malnutrition (PEM) is defined as poor nutritional status due to inadequate nutrient intake.

Reduced food intake leads to reductive adaptation, which includes a decrease in nutrient stores, changes in body composition and the more efficient use of fuels such as the use of ketone bodies by the brain (metabolic changes that accompany starvation are described in Chapter 21).

Malnutrition is one of the key issues faced by public health in the developing world, and needs to be viewed from not only medical but also social and economic perspectives. Mortality in malnourished patients (BMI between 10 and 13) is four times higher compared to well-nourished people. The effects of malnutrition are summarized in Table 22.6. Worldwide, malnutrition contributes to 54% of the 11.6 million deaths annually among children below 5 years of age.

Table 22.6

Consequences of protein-calorie malnutrition

Decreased protein synthesis
Decreased activity of Na^+/K^+-ATPase
Decreased glucose transport
Fatty liver, liver necrosis, liver fibrosis
Depression, apathy, mood changes
Hypothermia
Compromised ventilation
Compromised immune system: impaired wound healing
Risk of wound breakdown
Decreased cardiac output
Decreased renal function
Loss of muscle strength
Anorexia
Decreased mobility

In the developed world, malnutrition is a problem in hospitalized patients who are unable to eat because of their primary problem: for instance, stroke or cancer. Gastrointestinal problems, particularly colon pathology and celiac disease (see case, Chapter 10), or postoperative state, are associated with specific nutritional problems. Malnutrition also affects a large group of older individuals.

In the UK 25–34% of people admitted to hospital are at risk of malnutrition and 20–40% of critically ill patients have evidence of protein-energy malnutrition. Malnutrition is associated with increased morbidity and mortality, with longer hospital stay, and with increased rate of complications. In addition to malnutrition, specific deficiencies such as those of vitamin D, iron, and vitamin C may occur.

Markers of malnutrition risk

BMI of below 18.5 kg/m^2 suggests significant risk of malnutrition and so does unintentional loss of 10% of body weight within preceding 3–6 months. In the course of acute illness, inability to eat for more than 5 days poses the risk.

There are two types of protein-calorie malnutrition: marasmus and kwashiorkor

Marasmus results from a prolonged inadequate intake of calories and protein. It is a chronic condition, which develops over months or years. It is characterized by loss of muscle tissue and subcutaneous fat with the preservation of the synthesis of visceral proteins such as albumin. There is a clear loss of weight.
Kwashiorkor is a more acute form of undernutrition, which may also occur on the background of marasmus. It also develops because of inadequate nutrient intake after trauma or infection. In kwashiorkor, in contrast to marasmus, visceral tissues are not spared: the hallmark of kwashiorkor is edema due to the low concentration of plasma albumin and the loss of oncotic pressure (Chapter 24). The edema may mask the weight loss. Complications of kwashiorkor are dehydration, hypoglycemia, hypothermia, electrolyte disturbances and septicemia. These patients have impaired immunity and wound healing, and are prone to infection.

The WHO classification of malnutrition is based on anthropometry and the presence of bilateral pitting edema. Another classification has been proposed which distinguishes complicated from uncomplicated malnutrition (Table 22.7). Marasmus and kwashiorkor are terms rarely used in the hospital practice in

developed countries; **malnutrition** and **complicated malnutrition** are probably more appropriate.

Table 22.7
Classification of malnutrition

	Moderate	Severe	Complicated
Weight for height (% of median)	70–80	<70	<80
– or pitting edema	No	Yes	Yes
– or mid upper arm circumference	110–125	<110 mm	<110 mm
Appetite, clinically well, alert	Yes	Yes	No*

Note that, in the classic classification of malnutrition, the presence of edema is also the main differentiating feature between marasmus and kwashiorkor.

*Patients with complicated malnutrition may develop anorexia, high fever, anemia and dehydration.

After Collins S, Yates R: The need to update the classification of acute malnutrition, *Lancet* 362:249, 2003.

Refeeding syndrome develops as a consequence of inappropriate feeding of a malnourished person

The inpatient treatment of malnutrition in famine areas includes standard preparations such as Formula 100 therapeutic milk (F100). F100 is a liquid diet with an energy content of 100 kcal/100 mL. It includes dried skimmed milk, oil, sugar and a mix of vitamins and minerals (without iron). In the areas of famine, community feeding programs use the so-called life-sustaining general rations (at least 2100 kcal; 8786 kJ/day) containing grains, legumes, and vegetable oil. During the treatment of malnutrition, this needs to be combined with providing adequate water, sanitation, and basic health care.

It is important to take time to replete a starved person nutritionally. Too quick a replacement may be dangerous due to a possibility of a major shift between intracellular and extracellular fluid. This is known as the **refeeding syndrome**,

and is characterized by a severe decrease in concentrations of serum magnesium, phosphate and potassium (the latter because of the stimulation of insulin secretion). Also, if thiamine deficiency is present, carbohydrate feeding can precipitate the **Wernicke–Korsakoff syndrome** (Chapter 11). Frequent simple meals at short intervals are recommended during famine relief and, in a hospital setting, gradual introduction of nutritional support and close monitoring are required.

Clinical box Obese man with type 2 diabetes, coronary disease and arthritis

Mr K is a 55-year-old man with type 2 diabetes and coronary disease. He suffers from angina on effort and from severe arthritic knee pain. When he initially presented to the out-patient clinic his weight was 140 kg and his height 1.80 m (BMI 43). Within a year he managed to lose 12 kg by dieting. He was prescribed a lipase inhibitor, which he tolerated well. However, his arthritis worsened and he was increasingly less able to exercise. As a result his weight increased again to 137 kg. He was referred to the surgeons and is now being considered for gastric banding surgery.

Comment.

This patient illustrates multiple problems associated with obesity and in particular the way a concomitant disease may interfere with weight reduction programs. Weight loss is, to a substantial extent, dependent on the level of exercise. This patient lost weight initially but the maintenance of lower body weight was compromised by decreased mobility caused by arthritis.

Nutritional support

Nutritional support is required for a substantial number of hospitalized patients, and ranges from simple assistance with meals, through enriched or special-consistency diets, to enteral nutrition and total parenteral nutrition (Table 22.8).

Table 22.8

Forms of nutritional support

Special diets
Assistance with eating
Enteral nutrition (feeding through different feeding tubes: nasogastric, nasoduodenal; gastrostomy and jejunostomy)
Parenteral (intravenous) nutrition

Enteral nutrition entails feeding a person through special tubes placed in the stomach or jejunum

Enteral nutrition is appropriate when there are difficulties with taking food orally but the gastrointestinal tract functions properly. Standard enteral feeds contain carbohydrate, protein, fat, water, electrolytes, vitamins, and minerals, including trace elements. Predigested feeds contain short peptides or free amino acids.

When the gastrointestinal tract does not function because of, for instance, intestinal obstruction, or when large parts of it have been surgically removed, total parenteral nutrition is appropriate

Parenteral nutrition means **intravenous feeding**. Parenteral nutrition solutions contain fluids, glucose (dextrose), amino acids, and fats given as lipid emulsion (in the US derived from soybean oil and in Europe also from fish oil, olive oil and medium-chain triglycerides). Vitamins, minerals and electrolytes are also included. Total parenteral nutrition, while in many instances life-saving, is a treatment potentially associated with complications caused by intravenous line infections (it requires strictly sterile procedures) as well as metabolic problems. For this reason, parenteral nutrition treatment in hospitals is managed by multidisciplinary teams that include specialist nurses, surgeons, gastroenterologists, dieticians, pharmacists, and laboratory medicine physicians.

Obesity

Obesity has emerged as a major health problem worldwide

Worldwide obesity has increased by more than 70% since 1980. In the US, 35.7% of adults and 16.9% of children are obese (US National Health and Nutrition Examination Survey 2009). Main causes of this seem to be the wide availability of highly caloric food, and the decrease in physical activity both at work and during leisure time.

Clinical box A 46-year-old man with hypertriglyceridemia and fatty liver

Mr A is a 46-year-old man of South Asian descent. He has recently been gaining weight – 'middle-age spread' as he puts it. This weight gain is concentrated on his abdomen. His wife is concerned, as his father had had type 2 diabetes and died from a heart attack aged 60, and insists he goes to the doctor for a check-up. He attends the doctor and has blood samples taken – these show moderately raised total cholesterol, high triglyceride concentration, and low HDL-cholesterol Chapter 18). He has impaired fasting glucose and his liver function tests show a small rise in his transaminases. The liver ultrasound is consistent with intrahepatic fat deposition (the fatty liver).

Comment.

The combination of Mr A's sex, ethnicity and family history has predisposed him to gaining weight centrally rather than evenly through his body subcutaneously. This central obesity is due to fat within his abdominal cavity known as the visceral fat. This fat is deposited round his liver and pancreas, and increases his risk of developing type 2 diabetes. His plasma glucose is raised but he is not yet at the level classified as diabetes (Chapter 21). There is a degree of inflammation in his liver due to the layer of fat across it and, although rare, this can later lead to liver cirrhosis in some

people. His cholesterol profile is in keeping with insulin resistance; this is mostly due to the ectopic triglyceride deposition in muscle and liver.

Obesity is associated with an increased risk of medical and surgical problems

Obesity is associated with an increased risk of diseases in every system in the body (Table 22.9). In particular, it is a risk factor for type 2 diabetes mellitus (Chapter 21); the increased incidence of diabetes worldwide parallels that of obesity. Insulin resistance, which develops in obesity, is an important common denominator between obesity and diabetes. Obesity and insulin resistance carry an increased risk of cardiovascular disease.

Table 22.9

Health risks associated with obesity

System	Conditions associated with obesity
Cardiovascular	Coronary heart disease Phlebitis/venous ulceration High blood pressure High plasma cholesterol
Endocrine	Type 2 diabetes Polycystic ovary syndrome Infertility
Gastrointestinal	Nonalcoholic fatty liver disease Esophageal reflux Esophageal cancer Gallstones Hepatocellular cancer
Respiratory	Obstructive sleep apnea Asthma
Central nervous system	Idiopathic intracranial hypertension Stroke
Locomotor	Osteoarthritis Gout
Genitourinary	Cervical cancer Endometrial cancer Renal cancer Prostate cancer
Other	Breast cancer Cataracts Psoriasis Complications of pregnancy

Attempting weight loss to reverse the consequences of

obesity

Losing weight increases life expectancy, decreases blood pressure, decreases visceral fat deposition, improves plasma lipid concentrations, increases insulin sensitivity and normalizes glycemia, improves clotting and platelet function, and enhances the quality of life.

To lose weight, one needs to change the balance between energy intake and expenditure, *i.e.* between food intake and physical activity

However, the process involves many other factors, such as motivation, available time, cost, and access to appropriate weight reduction programs. Low-calorie diets contain approximately 1200–1300 kcal/day and very low-calorie diets around 800 kcal/day. Generally, a combination of **diet** and **exercise** plus **behavior interventions** (e.g. goal-setting, relapse avoidance) is more effective in inducing weight loss than diet alone. However, currently there are no evidence-based interventions that can induce a weight loss greater than 5% of body weight and maintain this long term, other than **surgical treatments** (bariatric surgery).

Low-carbohydrate diets for weight reduction

Low-carbohydrate, high-fat diets appear to be effective in weight reduction, at least in the short term, and they consistently lower plasma glucose concentration. As with many other aspects of dietary management, they are not completely evidence based. There are persisting concerns that include potential risk of abnormalities of liver and kidney function, and the long-term effects of higher fat intake on cardiovascular risk. Research in the field continues.

Surgical treatment of obesity results in large weight loss

Bariatric surgery includes **gastric bypass surgery**, where the stomach is made smaller and the first part of the small intestine is bypassed, and **gastric banding**, where a band is inflated around the part of the stomach. Surgery can result in a 20% weight loss and has been shown to reduce the future risk of cardiovascular disease and cancers, as well as improving mental health and general wellbeing. It can have such a profound effect on glycemic control that patients with diabetes achieve blood glucose concentrations well below the threshold at which diabetes is diagnosed.

Clinical box A 46-year-old man with hypertriglyceridemia and fatty liver: the benefits of weight loss

Mr A is worried about his father's history of type 2 diabetes and cardiovascular disease. He decides to try and lose weight. He thinks carefully about where he goes wrong in his diet and realizes that evening snacking on chocolate biscuits and sugary drinks is his main weakness. He decides the best option is to not buy any of these snacks at the supermarket so he can avoid temptation. He also decides he will join a gym and attend a circuit class twice per week. Six months into this regimen he has lost 7 kg, is feeling well and enjoying his new lifestyle. He returns to the doctor to have his bloods rechecked. His liver function tests have returned to within normal limits. However while his triglycerides are also lower, they still remain elevated, and his HDL-cholesterol is still low. His blood glucose is lower than before but is still classed as impaired fasting glucose. Mr A is upset that all the results have not normalized and feels that his efforts have been in vain.

Comment.

Mr A has done very well to lose 7 kg of weight. When a person with excess visceral fat loses weight, it is this visceral fat they lose first; this is why Mr A's liver function tests have improved as the liver is no longer ‘irritated’ by fat deposits. However, obesity is only one component of risk: in Mr A's case his sex, age, ethnicity and strong family history are risk factors that he cannot modify. That does not make lifestyle intervention futile in this case, as it will be good for general health and wellbeing. Nevertheless, Mr A may still go on to develop type 2 diabetes in the next 10 years. However, being less obese and participating in regular exercise may delay the onset of diabetes compared to what it would have been otherwise. Also, if he does develop diabetes, it should be simpler to treat.

Advanced concept box Definitions in nutrition science

The definitions below (given in bold) are used by the Food and Nutrition Board, Institute of Medicine (IOM) of the National Academies in the United States. Definitions used in the UK are also mentioned.The requirements are described by the **Dietary Reference Intakes (DRI)**. The DRI are described by:

- **Estimated Average Requirement (EAR):** average daily nutrient intake estimated to meet the requirement of half of the healthy individuals in a particular gender group at a particular stage of life. The EAR is complemented by:
  - **Recommended Dietary Allowance (RDA):** in the UK the *Reference Nutrient Intake (RNI)*. It describes the average daily nutrient intake level sufficient to meet the nutrient requirement of nearly all (97–98%) healthy individuals. The *Lower Reference Nutrient Intake (LNRI)* used in the UK: the daily intake observed at the low end of intake distribution in a population (about 2%). If intake falls below this, a deficiency may occur.
- **Adequate Intake (AI):** recommended average daily nutrient intake based on estimates of nutrient intake by a group of healthy people that are assumed to be adequate – used when an RDA cannot be determined.
- **Tolerable Upper Intake Level (UL):** highest average daily intake likely to pose no health risk to almost all individuals in a particular gender group at a particular stage of life. As intake increases above the UL, the risk of adverse effects increases.

Note that DRI are intended for healthy people, and a DRI established for any one nutrient presupposes that requirements for other ones are being met.

Advanced concept box Adipose tissue and disease

Adipose tissue can be stored in three distinct depots in the body: **subcutaneous** fat under the skin; **visceral** fat within the abdominal cavity; and the **ectopic** fat, for example the cardiac fat pad, or the excess deposition in the liver itself, and in myocytes. The distribution of fat in a person depends on their sex and ethnicity, combined with family history. For example, women are classically thought of as 'pear-shaped', with large subcutaneous adipose stores on their hips and thighs, while men are 'apple-shaped' with rounded abdomens due to excess visceral fat in their abdominal cavity. People of south-Asian, Chinese and Japanese descent are also generally more prone to depositing excess visceral rather than subcutaneous fat.

Subcutaneous fat is generally benign; it serves as an energy store. Health problems related to subcutaneous fat deposition are mechanical when a person gets so large they cannot function normally. Visceral fat, however, is an active endocrine organ that produces a number of mediators including interleukin 6 which lead to accelerated atherogenesis and premature cardiovascular disease, and facilitate development of type 2 diabetes. It also promotes a procoagulant state (Chapter 7) by secreting plasminogen activator inhibitor 1 (PAI-1), which inhibits fibrin degradation.

One of the major problems associated with visceral adiposity is the **nonalcoholic fatty liver disease (NAFLD)**. This is due to excess fatty acids being deposited in the liver. It leads to hepatic insulin resistance and is a risk factor for type 2 diabetes, and it can also lead to hepatitis and cirrhosis of the liver.

Ectopic fat pads are less well understood. The main fat pads associated with disease are the cardiac fat pad and the pharyngeal fat pad. The cardiac fat pad is thought to have several characteristics that can be mediators of cardiac disease. The first is simple mechanical obstruction of the heart, which impairs cardiac function. There is also the possibility that these cells also release procoagulant and inflammatory mediators having a local effect on

cells. There will also be toxic lipid accumulation in cardiac myocytes. The pharyngeal fat pad is associated with obstructive sleep apnea; in this condition, the pharynx is obstructed during sleep, leading to episodes of hypoxia and apneas (temporary cessation of breathing). People with this condition have very poor sleep quality and are at a high risk of developing cardiovascular disease, including hypertension. It is not known whether the pharyngeal fat pad has effects beyond obstructing the pharynx.

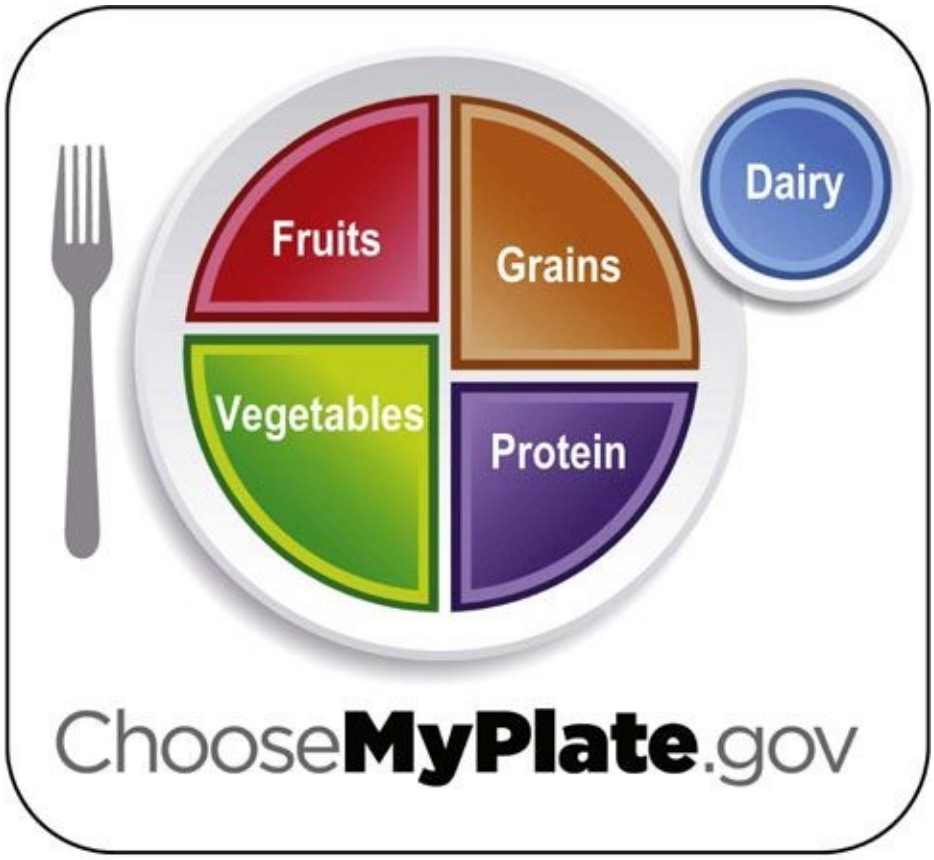
Visceral and ectopic fat is preferentially depleted during weight loss. This means that a person may not have to return to a 'normal' weight in order to substantially decrease their visceral and ectopic fat and the health risks these pose.

Healthy eating and dietary prevention of disease

Current dietary recommendations for the general population focus on balanced eating

Current recommendations stress the **balanced diet**. They unequivocally recommend high intake of fruit and vegetables. Starchy foods are another major part of the diet. With regard to carbohydrates, whole grains are probably the healthiest type of food (as opposed to, for example, white bread or white rice). The intake of dairy products is recommended in moderation. Fish, poultry, beans and nuts are regarded as healthy sources of proteins and are preferred over red meats. With regards to fat, vegetable cooking oils (olive and canola oil) are regarded as healthy fats, as opposed to saturated fats and particularly the *trans* fats. The diet should be low in unrefined carbohydrates, saturated fat and cholesterol. Restriction of sugar and caloric sweeteners is recommended, as well as sodium restriction and moderation in alcohol consumption. A healthy diet should be combined with an **active lifestyle**.

Figure 22.9 contains vignettes relating to major recommendations. My Plate contains recommendations of the US Department of Agriculture, which superseded the well known Food Pyramid. The critical comments and the alternative named Healthy Eating Plate has been recommended by the Harvard School of Public Health. The Eatwell Plate is the recommendation developed by the National Health Service in the UK.


Fruits
Grains
Dairy
Vegetables
Protein
ChooseMyPlate.gov

A

FIG. 22.9 Healthy eating.
Current recommendations emphasize balanced diet.
(A) My Plate contains the current recommendations on healthy eating developed by the US Department of Agriculture. It substituted the previously used Food Pyramid.
(B) The eatwell plate developed by the Food Standards Agency UK (www.food.gov.uk).

Diet is important in the prevention of coronary disease

Nutrients which affect atherosclerosis are dietary cholesterol, saturated fat and *trans* fatty acids. Excessive caloric intake and consequent obesity also facilitate atherosclerosis. High dietary cholesterol and a diet high in high saturated fat decrease the LDL receptor expression mediated through an increase in the intracellular cholesterol content in the hepatocytes: this leads to an increase in plasma cholesterol concentration (Chapter 17). Foods containing saturated fatty acids (full-fat milk, cheese, butter, and red meats) are atherogenic in excess. The rationale behind low-cholesterol diets is that LDL receptor expression can be increased if dietary cholesterol content is sufficiently low. The typical Western

diet contains approximately 400–500 mg of cholesterol daily. The intake required to achieve reduction in plasma cholesterol is much lower – below 200 mg or even below 100 mg daily.

Nutrients that seem to be protective against atherosclerosis include polyunsaturated ω-6-rich fatty acids (contained in vegetable oils) and ω-3-rich fish and fish oils, monounsaturated fat (contained in *e.g.* olive oil), and soluble fiber. Polyunsaturates also reduce plasma cholesterol concentration. The ω-3 fatty acids reduce triglycerides by decreasing VLDL synthesis and increasing its catabolism. They are also antithrombotic: fish oil inhibits thromboxane and PDGF synthesis, reduces blood viscosity and enhances fibrinolysis. Interestingly, it seems to electrically stabilize the heart muscle, preventing arrhythmias and decreasing the incidence of sudden death in people who have suffered myocardial infarction. Monounsaturated fats increase the HDL concentration, increase insulin sensitivity and decrease plasma triglyceride concentration.

Diets low in fat and cholesterol normally result in an approximately 10% decrease in serum cholesterol if applied during clinical trials, and a lesser decrease (usually below 5%) when applied to the general population. Alcohol seems to be protective against coronary disease but taken in excess it carries many other risks, obesity included. Unfortunately, low-fat diets seem to decrease HDL-cholesterol by 10–20%.

Also, low-fat diets, if the protein content is kept constant, tend to be high in carbohydrates. In spite of recommendations to use predominantly complex carbohydrates, in many populations refined sugars constitute too high a proportion of carbohydrate intake, and this promotes obesity. High-carbohydrate diets may also lead to hypertriglyceridemia.

Low-carbohydrate diet and the Mediterranean diet are alternatives to low-fat diet with regard to cardiovascular prevention and weight loss

For many years the **low-cholesterol, low-fat diet** was the hallmark of cardiovascular prevention. However, the emergence of the epidemics of obesity, and the greater understanding of the role of carbohydrates in the development of insulin resistance and metabolic syndrome, focused the attention on diets containing **low carbohydrates** and generally containing foods with a low glycemic index (see above).

Two types of diet seem to have significant health benefits. **The low-carbohydrate diets** (the Atkins diet and its derivatives) stress low carbohydrate content together with unrestricted caloric intake and are used primarily as means of weight reduction and maintenance. Typically such diet contains only 20 g carbohydrates per day during the 2-month induction phase. Subsequently, the carbohydrate content gradually increases to up to 120 g. Most regimens also encourage consumption of fat and protein from vegetarian sources and avoidance of the *trans* fats. People who take the low-carbohydrate diet are particularly prone to ketonuria.

The Mediterranean diet contains moderate amount of fat and a high proportion of monounsaturated fats. It is rich in vegetables, low in red meat and instead contains poultry and fish, as well as low-fat grains, fruits and legumes. To achieve weight loss, caloric restriction is necessary. Research studies suggest that both the low-carbohydrate diets and the Mediterranean lead to a greater weight loss than the low-fat diet. In terms of the effect on cardiovascular risk, these diets are at least alternative to the conventional low-fat diet.

Lifestyle and diet may influence the incidence of cancer

High intake of fruit and vegetables seems to be protective against cancer, possibly due to the presence of antioxidants and fiber. Dairy products seem also to be protective. There were suggestions that high-fat diets may increase cancer risk, but epidemiologic evidence for this is weak. In clinical studies β-carotene, retinoids (Chapter 11), fiber, and calcium were not found to be effective in cancer prevention. Massive weight loss with **bariatric surgery** has shown a reduction in the incidence of cancer.

Low calcium intake may be related to risk of osteoporosis

Osteoporosis is associated with low bone mass and bone fragility, which results in an increased risk of fractures (Chapter 26). It is particularly prominent in the elderly. Calcium is required for mineralization of bone and it is recognized that sufficient intake of **calcium**, alongside **vitamin D** in older adults is associated with a lower risk of osteoporosis. High alcohol intake and low body mass are associated with an increased risk of osteoporosis. Numerous other nutritional factors have been related to bone health, though with less evidence, including zinc, copper, manganese, boron, vitamin A, vitamin C and vitamin K; a diet containing high amounts of fruit and vegetables is recommended for optimum

bone health.

Summary

■ Appropriate nutrition underpins health and wellbeing, and poor nutrition increases susceptibility to disease.
■ Food intake is controlled by a neuroendocrine system responding to signals generated in adipose tissue.
■ Genotype, food availability, state of health, and physical activity are factors which determine nutritional status.
■ Nutritional needs change during the life cycle.
■ The main categories of nutrients are carbohydrates, fats, proteins, and vitamins and minerals. Water balance is closely associated with nutrition.
■ Assessment of nutritional status is an important part of general clinical assessment. It includes the assessment of current diet, dietary history, clinical examination and a range of biochemical and hematologic tests.
■ Malnutrition affects large areas of the developing world, and in the developed world is an issue among disadvantaged social groups and also in hospitalized persons.
■ Obesity has become a major health problem worldwide.
■ Nutritional support includes graded assistance with nutrient intake, ranging from assistance with meals to total parenteral nutrition.

Active learning

1. Outline the processes that maintain energy homeostasis.
2. Describe the role of different classes of fatty acids in nutrition.
3. List the principles of a weight reduction program.
4. Discuss instances when increased nutritional demand can precipitate malnutrition.
5. What diet would you recommend for a diabetic patient?

Further reading

Brand-Miller, J, Buyken, AE. The glycemic index issue. *Curr Opin Lipidol*. 2012; 23:62–67.
Collins, S, Yates, R. The need to update the classification of acute malnutrition. *Lancet*. 2003; 362:249.
Cota, D. CXB1 receptors: emerging evidence for central and peripheral mechanisms that regulate energy

balance, metabolism, and cardiovascular health. *Diabetes/Metabol Res Rev*. 2007; 23:507–517.

Crowley, VEF. Overview of human obesity and central mechanisms regulating energy homeostasis. *Ann Clin Biochem*. 2008; 45:245–255.

Dietary Reference Values for Food Energy and Nutrients for the United Kingdom. *Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy, Department of Health*. London: TSO; 2003.

Eckel, RH. Nonsurgical management of obesity. *N Engl J Med*. 2008; 358:1941–1950.

Elamin, E. Dietary and pharmacological management of severe catabolic conditions. *Am J Med Sci*. 2011; 342:513–518.

Gastadelli, A. Role of beta-cell dysfunction, ectopic fat accumulation and insulin resistance in the pathogenesis of type 2 diabetes mellitus. *Diab Res Clin Pract*. 2011; 93S:560–565.

Gidden, F, Shenkin, A. Laboratory support of the clinical nutrition service. *Clin Chem Lab Med*. 2000; 38:693–714.

Ramachandrappa, S, Farooqi, IS. Genetic approaches to understanding obesity. *J Clin Invest*. 2011; 121:2080–2086.

Shai, I, Schwarzfuchs, D, Henkin, Y, *et al*. Weight loss with a low-carbohydrate, Mediterranean, or low-fat diet. *N Engl J Med*. 2008; 359:229–241.

Vincent, HK, Innes, KE, Vincent, KR. Oxidative stress and potential interventions to reduce oxidative stress in overweight and obesity. *Diabetes Obesity Metabol*. 2007; 9:813–839.

Willett, WC, Ludwig, DS. The 2010 dietary guidelines – the best recipe for health? *N Engl J Med*. 2011; 365(17):1563–1565.

Websites and downloads

BAPEN. What is malnutrition. www.bapen.org.uk/about-malnutrition/introduction-to-malnutrition.

BAPEN Malnutrition Universal Screening Tool. www.bapen.org.uk/pdfs/must/must_full.pdf.

Burani, J. Gushers and Tricklers. Practical use of the glycemic index. www.glycemicindex.com/cmsAdmin/uploads/glycemicindex.ppt, 2006.

ChooseMyPlate.gov. United States Department of Agriculture.

www.choosemyplate.gov/

www.cnpp.usda.gov/Publications/MyPlate/GettingStartedWithMyPlate.pdf

Dietary Reference Intakes. Applications in Dietary Planning (2003). Food and Nutrition Board, Institute of Medicine (IOM) of the National Academies. www.iom.edu.

Eight tips for healthy eating – Live Well – NHS Choices. www.nhs.uk/Livewell/Goodfood/Pages/eight-tips-healthy-eating.aspx.

Eatwell plate. www.nhs.uk/Livewell/Goodfood/Documents/Eatwellplate.pdf.

International chair on cardiometabolic risk. www.myhealthywaist.org.

UK Food Standards Agenc. www.Foodstandards.gov.uk

World Health Organization. Obesity. www.who.int/topics/obesity/en/.

World Health Organization, Management of Severe Malnutrition: A Manual for Physicians and Other Senior Health Workers. WHO, Geneva, 1999. www.who.int/nutrition/publications/severemalnutrition/9241545119/en/

The glycemic index. www.glycemicindex.com.

The Nutrition Source. Healthy Eating Plate vs. USDA's MyPlate. www.hsph.harvard.edu/nutritionsource/healthy-eating-plate/healthy-eating-plate-vs-usda-myplate/index.html.

CHAPTER 23

Role of Kidneys in Metabolism

Marek H. Dominiczak and Mirosława Szczepańska-Konkel

Learning objectives

After reading this chapter you should be able to:

- Describe the handling of sodium and water in the nephron.
- Discuss links between renal oxygen consumption and sodium reabsorption.
- Explain the role of sodium/potassium ATPase in the nephron.
- Explain why urine analysis can provide clinically important information.
- Describe clinical assessment of the glomerular filtration rate.
- Comment on the possible mechanisms of proteinuria.

Introduction

The human kidneys weigh about 300 g. They have a rich blood supply, receiving about 25% of the cardiac output. About 80% of the blood is distributed in the renal cortex and 20% in the renal medulla. Almost all the blood passes through the glomerular capillaries, which act as high-pressure filters. Every day about 180 L of plasma containing several kilograms of plasma proteins, sodium chloride and other electrolytes, and metabolites, are filtered through the glomerular filtration area of 0.5–2 m^2. More than 99.9% of the plasma proteins are retained by the filter, whereas almost all filtered water and sodium chloride and other solutes are retained by transport systems in the renal tubules. Under normal conditions, the kidneys form 1–2 liters of urine per day. Urine composition is summarized in Table 23.1.

Table 23.1

Daily excretion of nitrogen compounds and main ions in urine (mmol/24h)

Urea	Uric acid	Creatinine	Ammonia
250–500	1–5	7–15	30–50
Sodium	**Potassium**	**Chloride**	**Phosphate**
100–250	30–100	150–250	15–40

Urea is a major contributor of the urine nitrogen excretion. It is the final product of protein catabolism in humans. Daily urea excretion also reflects nutrition status and strongly depends on protein intake.

Uric acid excretion depends mainly on endogenous purine degradation, but might be elevated on a purine-rich diet.

Creatinine is derived from the skeletal muscle phosphocreatine.

At metabolic steady state, urinary excretion of **nitrogen** compounds strictly depends on kidney function.

In **renal failure** urine output falls and this leads to an increase in plasma urea and creatinine concentrations.

Urinary excretion of sodium, potassium and chloride reflect their intake. Excessive sodium intake

or impaired elimination might lead to hypertension.

Ammonia is generated in the kidney by deamination of glutamine and glutamate, and is excreted as the ammonium ion. Daily excretion of ammonia and phosphate depends on hydrogen ion excretion in urine (Chapter 25).

Approximate values in an average adult person are given.

The key functions of the kidney are filtration of plasma, and subsequent tubular excretion and reabsorption of ions, low-molecular-weight substances and water

The excretory function of the kidneys involves filtration of the plasma in the glomeruli, transport of water and solutes from the tubular lumen back to the blood (tubular reabsorption), and secretion of different substances from the tubular cells into the lumen.

The kidneys remove, in urine, products of metabolism such as urea, uric acid and creatinine, and retain substances such as glucose, amino acids, and proteins. They also metabolize and remove drugs and toxins. They play a critical role in the regulation of extracellular fluid volume and composition and in maintaining the acid–base balance (Chapters 24 and 25). These functions are regulated hormonally by vasopressin (antidiuretic hormone, ADH) produced in the posterior pituitary, and by the renin–angiotensin system. Kidney function is also controlled by norepinephrine and dopamine released from nerve terminals in the kidney itself.

The kidneys also produce calcitriol (1,25-dihydroxycholecalciferol, $1,25(OH)_2D_3$), a vitamin involved in calcium homeostasis (Chapter 26) and erythropoietin, which controls the production of erythrocytes.

The functional unit of the kidney is the nephron

Each kidney consists of approximately 1 million nephrons composed of glomerulus and an excretory tubule (Fig. 23.1). Glomeruli, located in the kidney cortex, are biological filters, interfacing the plasma with the tubules that reabsorb and excrete substances. The segments of each tubule (starting from the glomerulus end) are known as the **proximal tubule**, the **loop of Henle**, the **distal tubule**, and the **collecting duct**.

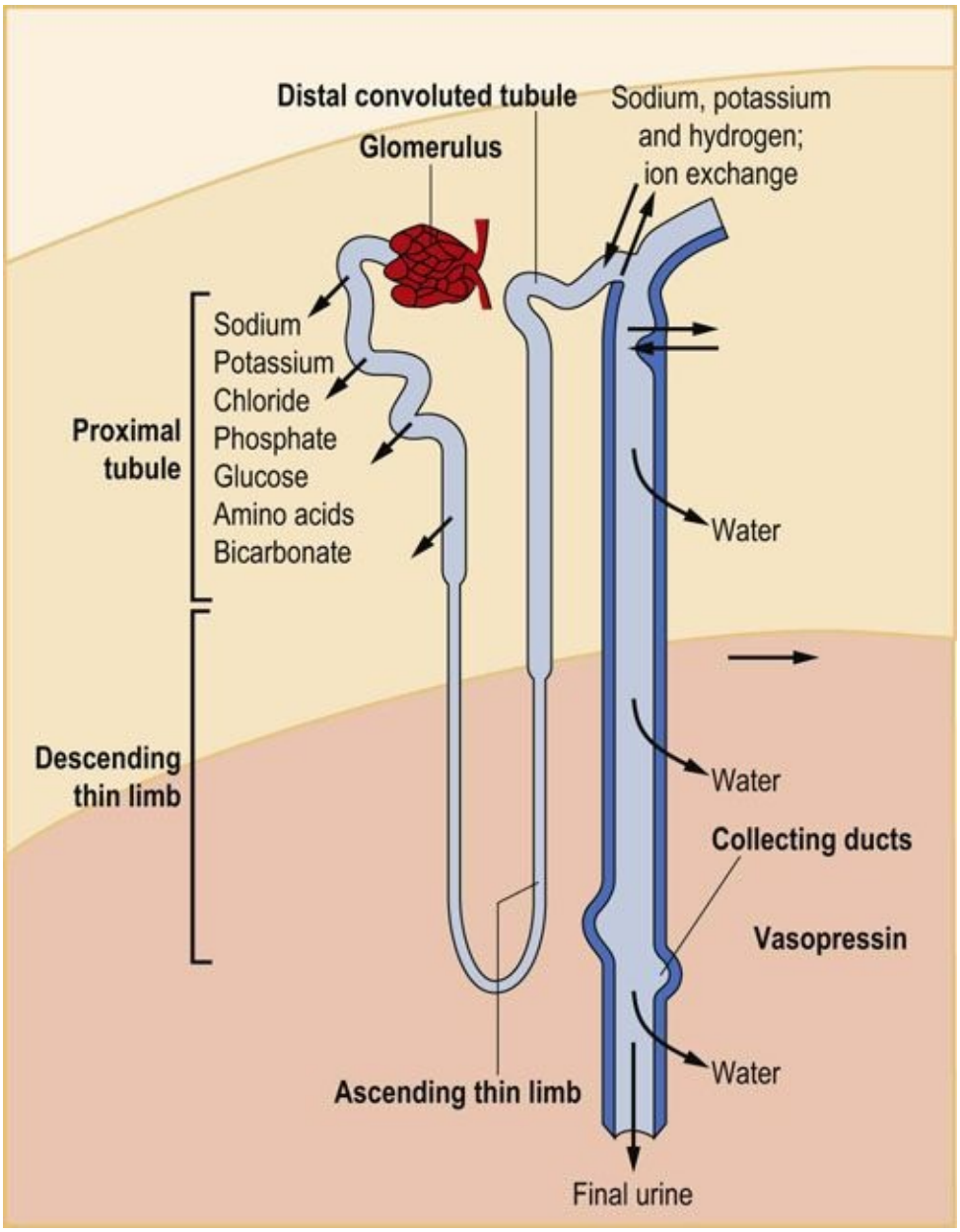


FIG. 23.1 The nephron and its major transport sites.

The glomerular filtration barrier

Plasma from the glomerular capillaries is filtered into the interior of the glomerulus known as the Bowman's space. The filtration barrier includes a layer of endothelial cells that line glomerular blood vessels, the basement membrane, and the epithelial cells (podocytes) with characteristic foot processes (Fig. 23.2). Glomerular filtration depends on the filtration surface and on the permeability of

the barrier.

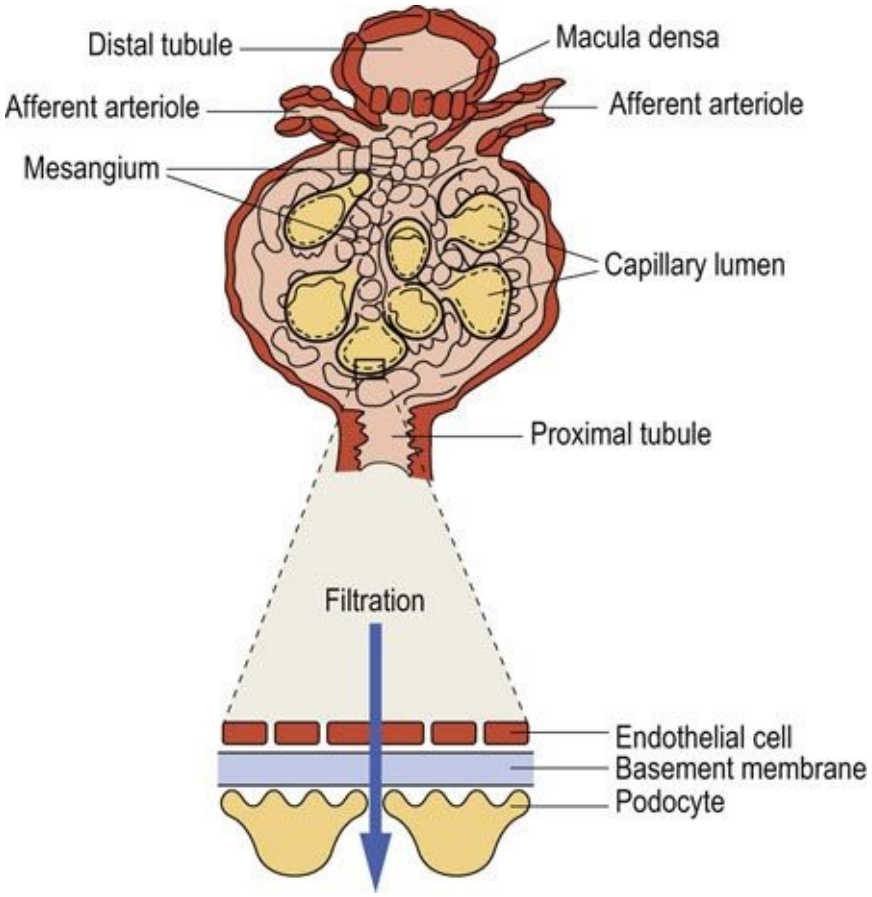


FIG. 23.2 Glomerular filtration barrier.
The barrier consists of the endothelial cells, the basement membrane, and the podocytes. The *macula densa* cells are part of the juxtaglomerular apparatus: they sense chloride concentration in the distal tubule and adjust the diameter of afferent arterioles, regulating the glomerular blood flow.

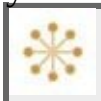
The main component of the glomerular basement membrane is type IV collagen, which forms a network of filaments providing elasticity and resistance to hydrostatic pressure. The membrane also contains laminin, fibronectin and proteoglycans with negatively charged heparan sulfate groups, which form an electrostatic barrier for proteins filtered from plasma.

The podocytes and mesangial cells possess receptors for a range of vasoactive substances such as angiotensin II, vasopressin, bradykinin, ATP, endothelin, prostaglandins, dopamine, natriuretic peptides and adenine nucleotides.

The fenestrations in the endothelial layer and the spaces between foot

processes of the podocytes form a sieve that filters water and small molecules. Filtration of larger molecules is limited by their size, shape, and electric charge. For instance, at pH 7.4, most plasma proteins are negatively charged, and so is the filtration barrier; this hinders filtration of even the smallest proteins, such as myoglobin (molecular mass 17 kDa), and almost completely prevents filtration of the larger (69 kDa) albumin.

Glomerular filtration is driven by the hydrostatic pressure in the glomerular capillaries, which is approximately 50 mmHg. The hydrostatic pressure is counteracted by the oncotic pressure of the plasma and the back-pressure (approximately 10 mmHg) of the filtrate in the glomerular capsule. Changes in the glomerular filtration rate alter the total amount of filtered water and solute, but not the composition of the filtrate. A decrease in the blood pressure in the afferent arteriole of the glomerulus is sensed by the group of cells known as the juxtaglomerular apparatus. This stimulates renin secretion and activates the renin–angiotensin system. An increase in the intracapillary blood pressure and/or hyperfiltration may induce glomerular cells' injury and may lead to severe kidney damage.



Advanced concept box

Hyperglycemia and mechanical stress resulting from glomerular hypertension contribute to the development of diabetic nephropathy

Mechanical stress and high glucose concentration trigger multiple signaling pathways that may accelerate glomerular cells' damage. Podocytes, the cells covering the outer aspect of glomerular basement membrane, are subjected not only to the load of filtered glucose but also to diverse mechanical forces. Both high glucose and mechanical stress may impair the protein systems anchoring the podocyte foot processes in the glomerular basement membrane (integrin, agrin), therefore blunting resistance of these cells to mechanical forces, and causing their detachment from the membrane (and appearance in urine). Modulation by these factors of expression and activity of numerous structural and functional proteins such as actin, nephrin, podocin, and other compounds

such as phospholipase C, growth factors, cytokines, and chemokines results in the inflammatory responses, dysfunction, and apoptosis or necrosis of the podocytes. Loss of the podocytes is irreversible due to their inability to proliferate and to replenish damaged cells. Podocytes are injured early in the course of diabetic nephropathy.

Glomerular filtrate: formation of urine

Kidneys consume large amounts of oxygen, mainly to support active sodium transport

Most of the metabolic processes in the kidneys are aerobic, and consequently their oxygen consumption is high: it approximately equals that of the cardiac muscle, and is three times greater than that of the brain. Such a high metabolic activity is required to maintain tubular reabsorption; about 70% of the oxygen consumed by the kidney is used to support active sodium transport, which in turn determines reabsorption of glucose and amino acids.

Reabsorption of sodium. Approximately 80% of the filtrate is reabsorbed in the proximal tubule. Sodium is reabsorbed by several mechanisms: through specific ion channels, in exchange for the hydrogen ion, and in co-transport with glucose, amino acids, phosphate, and other anions. The movement of sodium causes reabsorption of water (Fig. 23.3). The entry of sodium into the proximal tubular cells is passive. This is possible because the Na^+/K^+-ATPase maintains low sodium concentration in the cytoplasm of the tubular cells. More sodium is reabsorbed in the distal tubule and in the collecting duct in exchange for potassium or hydrogen ion. This process is controlled by aldosterone (Fig. 23.3 and 23.4).

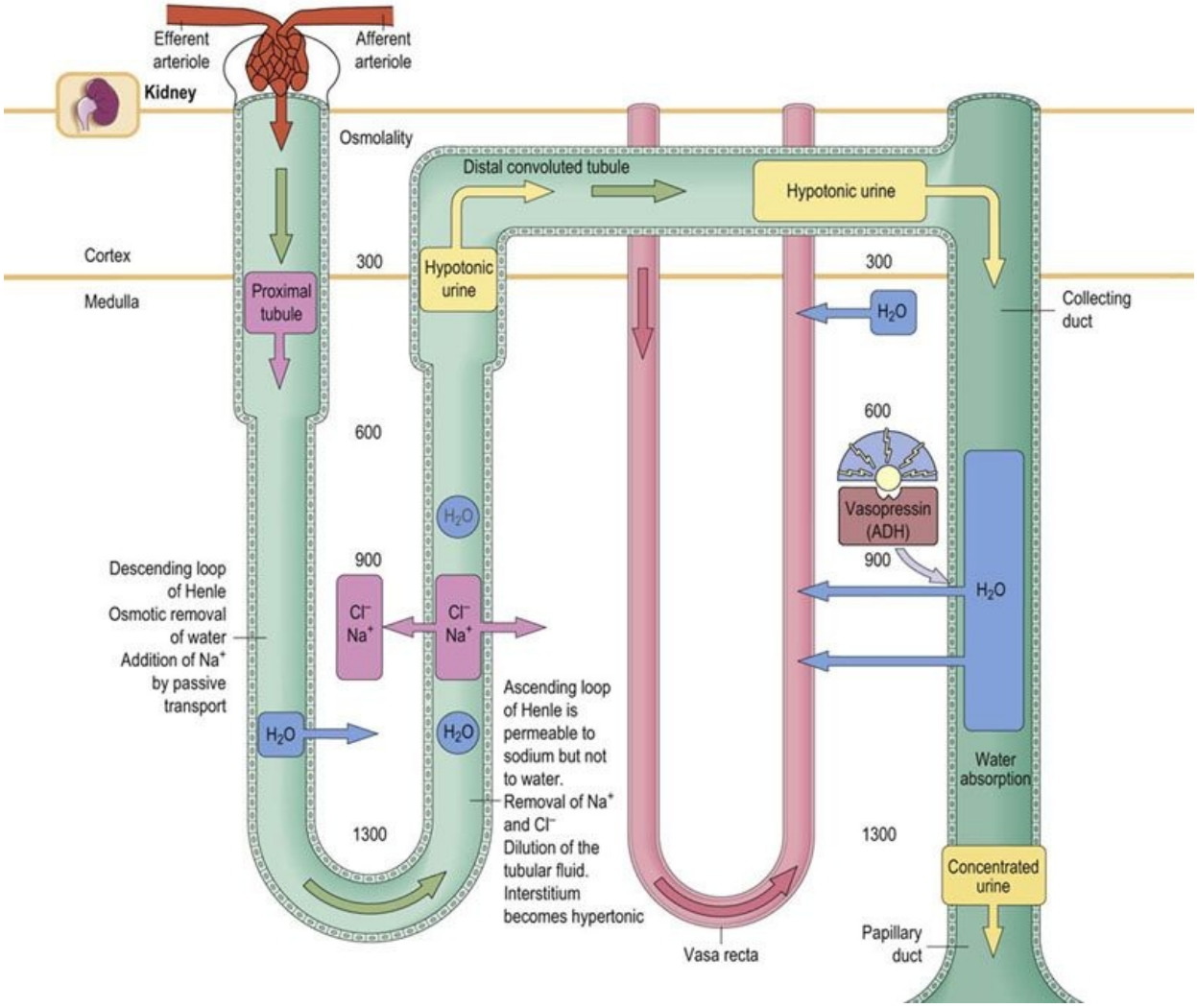


FIG. 23.3 Counter-current exchange and multiplication in the renal tubules. The counter-current mechanism is essential for the formation of urine, and for reabsorption of water in the distal tubule. In the **ascending arm of the loop of Henle**, sodium and chloride ions are pumped into the interstitial fluid. They then diffuse freely into the lumen of the **descending limb**, creating a functional loop, which perpetuates the increase in osmolality of the filtrate reaching the ascending limb. This is known as **counter-current multiplication**. As a result of this, while the osmolality of the renal cortex is similar to that of plasma (300 mmol/L), in the medulla it might be as high as 1300 mmol/L. High osmolality of the medulla later facilitates reabsorption of water in the collecting ducts. This is known as **counter-current exchange**. The amount of reabsorbed water is controlled by vasopressin.

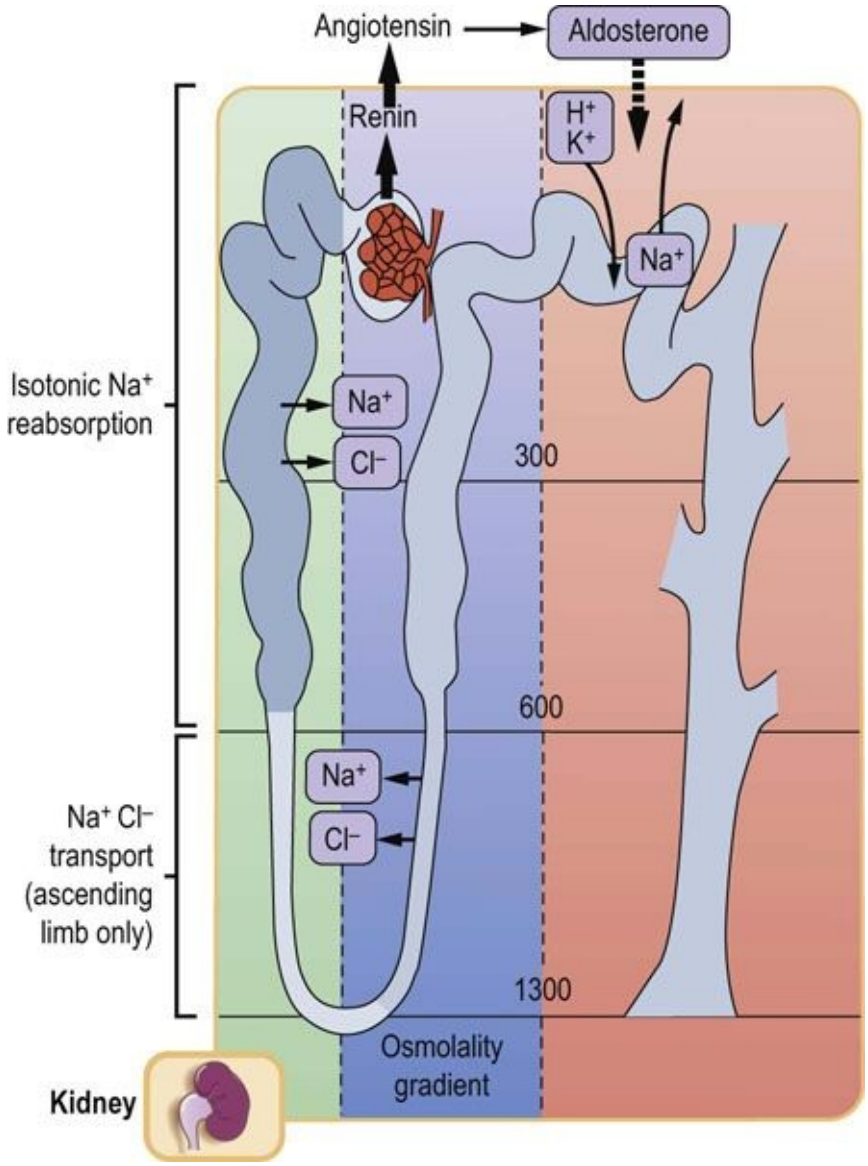


FIG. 23.4 Sodium reabsorption in the renal tubules.
More than 80% of filtered sodium is actively reabsorbed in the proximal tubule. Sodium and chloride ions are also reabsorbed in the ascending limb of the loop of Henle. A different mechanism operates in the distal tubule, where sodium reabsorption is stimulated by **aldosterone** and is coupled with the secretion of hydrogen and potassium ions. Aldosterone causes sodium retention and an increase in potassium excretion.

Reabsorption of water. The fluid leaving the proximal tubule is isotonic. Different permeability of the ascending and descending limbs of the loop of Henle maintains high osmolality of the medulla. This is essential for subsequent reabsorption of water in the collecting duct (Fig. 23.5). The fluid that leaves the loop of Henle is diluted (hypotonic). Water reabsorption in the collecting duct is controlled by vasopressin (Fig. 23.3).

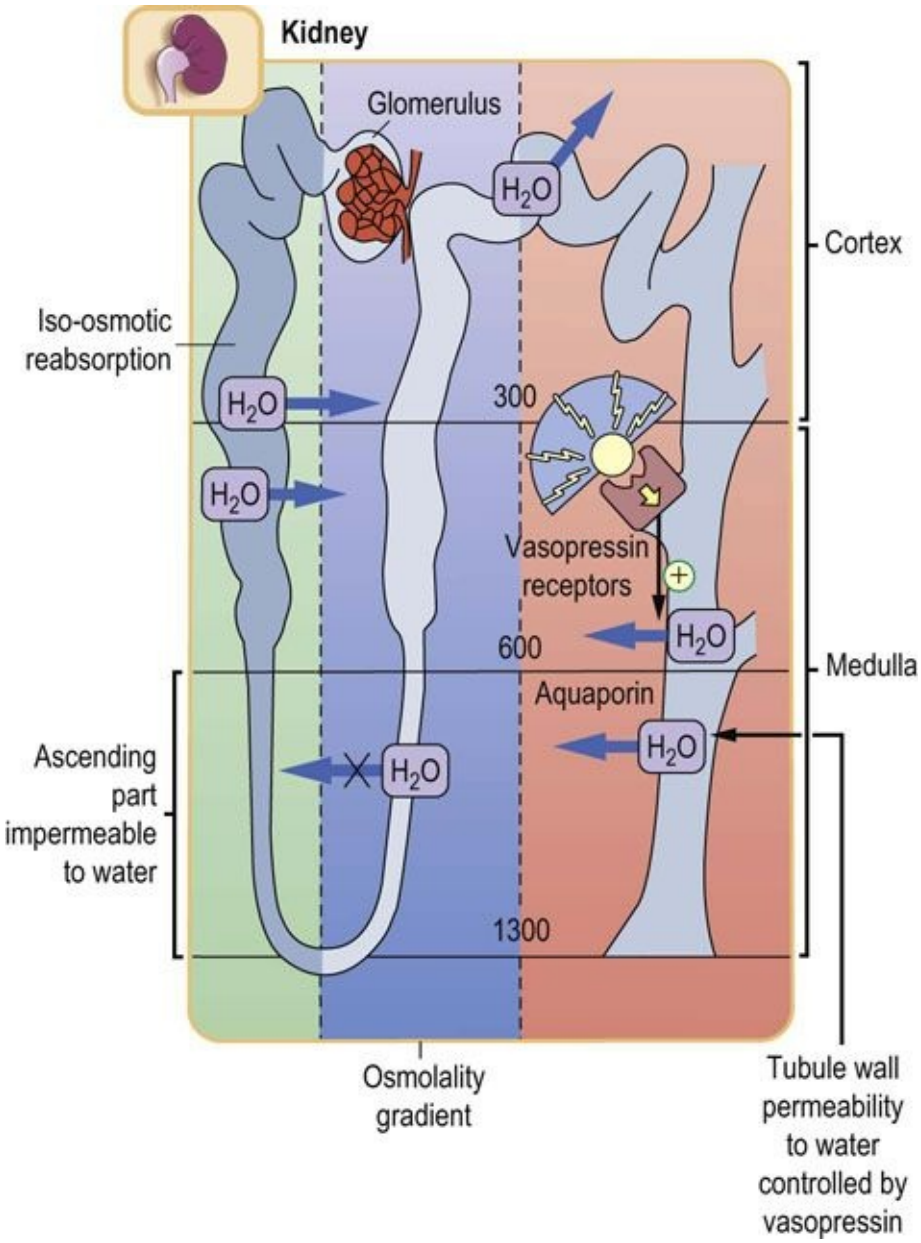


FIG. 23.5 Renal handling of water.
Permeability of the tubular walls to water differs along the nephron. About 80% of filtered water is reabsorbed in the proximal tubule, by iso-osmotic reabsorption. The ascending loop of Henle is impermeable to water. In the collecting duct, **vasopressin** controls the water reabsorption through **aquaporin water channels** (Chapter 24).

Role of kidneys in glucose homeostasis

Kidneys contribute to glucose homeostasis via three different mechanisms: the release of glucose into the circulation via **gluconeogenesis**, the **uptake of**

glucose from the circulation to satisfy its own energy needs, and the **reabsorption of glucose** from the glomerular filtrate.

Renal gluconeogenesis

In humans, only the liver and kidney are capable of gluconeogenesis (Chapter 13). After an overnight fast, 75–80% of glucose released into the circulation derives from the liver and the remaining 20–25% derives from the kidney. Surprisingly, after a meal, renal gluconeogenesis increases approximately twofold and accounts for *60% of endogenous glucose release in the postprandial period (hepatic gluconeogenesis decreases by* 80% at that time). **Glutamine** and **lactate** are preferential gluconeogenic precursors in the kidney.

The key enzymes of gluconeogenesis, pyruvate carboxylase, phosphoenol pyruvate carboxykinase, fructose-1,6-bisphosphatase, and glucose-6-phosphatase, are found mainly in the renal cortical cells. The rate of gluconeogenesis depends on the glucose concentration, substrate availability and hormonal control. Insulin suppresses, and epinephrine increases, glucose release by the kidney. Glucagon has no effect on renal glucose release. The produced glucose provides the brain and other organs with a needed energy substrate, particularly when liver gluconeogenesis is impaired: *e.g.* after liver transplantation, in hepatic failure, and during prolonged fasting, hypoglycemia and acidosis.

Renal glucose utilization

The metabolic fate of glucose is different in different regions of the kidney. Glucose is an essential energy substrate for the renal medulla, because of low oxygen tension and low levels of oxidative enzymes there. Consequently, in the medulla, the lactate is the main metabolic end product of glucose metabolism.

In contrast, the renal cortex contains high levels of oxidative enzymes. The main energy substrates in the renal cortex are thus fatty acids, lactate, glutamate, citrate, and ketone bodies. Of these, the fatty acids are the main source of energy.

Renal tubular cells with a high Na^+/K^+-ATPase activity possess multiple mitochondria situated close to the plasma membrane, so that the released ATP is easily accessible.

Neuronal control of kidney function

The kidneys communicate with the central nervous system via the sensory (afferent) nerves. An increase in renal afferent activity directly influences sympathetic outflow to the kidneys via efferent nerves. Norepinephrine released from the nerve terminals, together with co-transmitters (such as ATP, other adenine mono-and dinucleotides, and neuropeptide Y) modify kidney function (Fig. 23.6). Norepinephrine increases sodium reabsorption directly via tubular Na^+/K^+-ATPase activation, and indirectly via the stimulation of the renin–angiotensin system (Fig. 23.7). Increased renal sympathetic activity has been identified as a major contributor to the complex pathophysiology of **hypertension**. On the other hand, kidney denervation leads to blood pressure lowering.

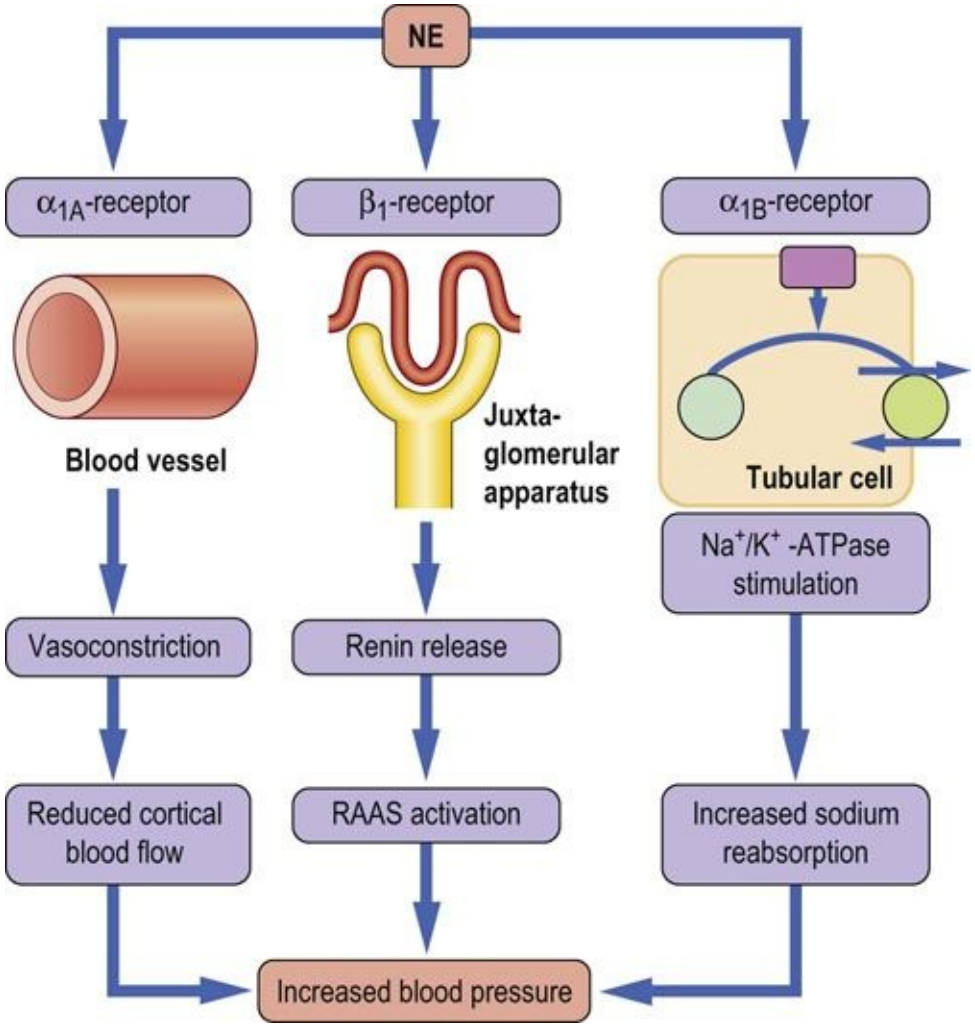


FIG. 23.6 Renal effects of norepinephrine.

Kidneys have dense sympathetic innervation, which terminates in the blood vessels, the juxtaglomerular apparatus, and the renal tubules. Stimulation of the renal sympathetic nerves causes norepinephrine release, which increases renin secretion rate by activation of β_1-adrenoceptors, and increases the renal sodium reabsorption by activation of α_{1B} receptors and tubular Na^+/K^+-ATPase, and decreases the renal blood flow by activation of α_{1A}-adrenoceptors. Overall, the activation of renal adrenergic receptors leads to increased blood pressure. NE, norepinephrine; RAAS, renin-angiotessin-aldosterone system.

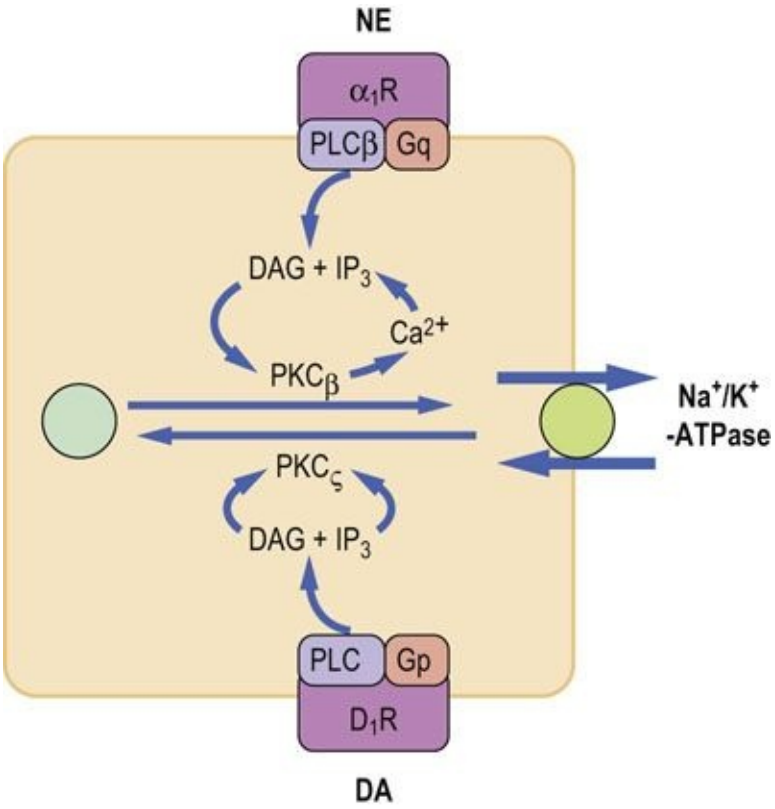


FIG.23.7 Norepinephrine and dopamine regulate sodium reabsorption in the renal proximal tubule.

Norepinephrine and dopamine control Na^+/K^+-ATPase. Norepinephrine stimulates Na^+/K^+-ATPase via α_{1B}-adrenergic receptors, and through G-proteins stimulates phospholipase C_β and protein kinase C_β, leading to enzyme phosphorylation and its translocation to the cell membrane. On the other hand, dopamine inhibits Na^+/K^+-ATPase. Dopamine binds to its receptors and through G-proteins stimulates phospholipase C, which in turn activates protein kinase C_ζ. PKC_ζ phosphorylates membrane Na^+/K^+-ATPase and induces its translocation to cytoplasm. NE, norepinephrine; DA, dopamine; PKC_β,protein kinase C_β; PKC_ζ, protein kinase C_ζ; α_1R, α_{1B}-adrenergic receptors; D_1R, Dopamine receptor 1.

Membrane transport systems in the kidney

Na^+/K^+-ATPase

Na^+/K^+-ATPase activity in the kidney is several thousand times higher than in other tissues. In the kidney, its major function is sodium reabsorption by catalyzing the extrusion of sodium to the interstitial fluid. There is a close relationship between the amount of Na^+/K^+-ATPase and sodium reabsorptive capacity of the different segments of the nephron. In the renal tubular cells, as in all sodium-reabsorbing epithelial cells, Na^+/K^+-ATPase is present in the basolateral membrane. Its activity in the nephron is controlled by a number of hormones, mainly by aldosterone, angiotensin II, and by neurotransmitters norephinephrine and dopamine (Chapter 24). In humans, the kidney reabsorbs about 18 moles of sodium per day and utilizes about 6 moles of ATP for this process. Thus, Na^+/K^+-**ATPase is an energy transducer** that converts metabolic energy into ionic gradients.

Sodium transport system in the renal tubules

Sodium reabsorption occurs along the nephron, with the exception of the descending limb of the loop of Henle. The driving force for the reabsorption of sodium and other solutes is the electrochemical gradient generated by the Na^+/K^+-ATPase in the basolateral membrane of tubular cells. The transport of sodium ions across cellular membrane, similarly to other polar molecules, requires the involvement of specific membrane proteins.

Understanding the renal transport systems elucidates the action of diuretics

In the proximal tubule the sodium ions enter into the cell from the luminal side via the sodium–hydrogen exchanger (NHE3), through ion channels and via co-transporters with glucose, phosphates and amino acids. The sodium–bicarbonate co-transporter (known as NBC1) is located in the basolateral membrane.

In the thin ascending limb of the loop of Henle the sodium ions move into the cell via sodium–potassium–chloride co-transporter (known as NKCC2), which is inhibited by **furosemide.** In this segment, potassium ions are secreted

into the lumen by the ATP-sensitive rectifier potassium channel.

In the distal tubule the reabsorption of sodium ions involves sodium–chloride co-transporter (NCC), which is **thiazide**-sensitive.

In the collecting duct the sodium ions are reabsorbed by **amiloride**-sensitive epithelial sodium channel (ENaC). Aldosterone, through the mineralocorticoid receptor stimulates both ENaC expression and the activity of the Na^+/K^+-ATPase. The pharmacologic antagonist of aldosterone is the diuretic **spironolactone**.

Reabsorption of sodium in turn drives the movement of water. The inhibitors of sodium reabsorption, *e.g.* thiazide diuretics (hydrochlorothiazide), loop diuretics (e.g. furosemide, torasemide), amiloride and spironolactone, are extensively used in clinical practice to induce natriuresis and diuresis.

Clinical box Diuretics are used for treatment of edema, cardiac failure and hypertension

Diuretics are drugs that stimulate water and sodium excretion. **Thiazide diuretics**, *e.g.* bendrofluazide, decrease sodium reabsorption in the distal tubules by blocking sodium and chloride co-transport. **Loop diuretics**, such as frusemide, inhibit sodium reabsorption in the ascending loop of Henle. **Spironolactone**, a potassium-sparing diuretic, is a competitive inhibitor of aldosterone: it inhibits sodium–potassium exchange in the distal tubules, and decreases potassium excretion. An osmotic diuresis may be induced by the administration of the sugar alcohol, **mannitol**.

The net effect of treatment with diuretics is increased urine volume and loss of sodium and water. Diuretics are important in the treatment of edema associated with circulatory problems such as heart failure, in which impaired cardiac function may lead to a severe breathlessness caused by pulmonary edema. They are also essential in the treatment of hypertension.

Glucose reabsorption

Normally, at the average plasma glucose concentration around 5 mmol/L (~100 mg/dL), approximately 200 g of glucose are filtered by the kidney each day. In healthy individuals, all filtered glucose is reabsorbed into the circulation and the urine is glucose-free. Reabsorption of glucose from the filtrate in the proximal convoluted tubules occurs with the participation of sodium–glucose co-transporters (SGLTs). There are two sodium-dependent–glucose co-transporter proteins: SGLT2 is a low-affinity, high-capacity transporter reabsorbing 90% of filtered glucose, and SGLT1 is a high-affinity low-capacity transporter. Glucose transport mediated by the SGLTs is an active process, utilizing energy derived from the sodium electrochemical gradient maintained by the Na^+/K^+-ATPase. Glucose reabsorbed into the tubular epithelial cells via SGLTs is released into the circulation by the specific glucose transporters (GLUT-1 and GLUT-2) located in the basolateral membranes.

Reabsorption of filtered glucose increases linearly until the maximal capacity of the tubules (T_{max}) is exceeded

The point of saturation of the transport system, the so-called renal threshold, is reached at the plasma glucose concentration of 11.0 mmol/L (198 mg/dL) in healthy adults. Above this concentration, the percentage of filtered glucose that is reabsorbed decreases and glucosuria appears. The threshold is decreased in individuals with a rare condition known as **familial renal glucosuria**, caused by a range of mutations to the SLC5A2 gene, which encodes the SGLT2. Depending on the nature of the mutations, affected individuals have varying degrees of glucosuria. In the most severe form they can lose more than 100 g of glucose per day in the urine.

Renal handling of amino acids

In humans, the total plasma concentration of amino acids is approximately 2.5 mmol/L (~25 mg/dL). Free plasma amino acids are filtered in the renal glomeruli (approximately 50 g per day). In parallel, amino acids are transported into the kidney cells and are metabolized. For example glutamine, transported through both luminal and basolateral membranes, is deaminated by glutaminase. The product of this reaction, ammonia, is secreted into the tubular lumen where it buffers the hydrogen ion (Chapter 25). In the kidney glutamine is also a substrate for gluconeogenesis.

The near-complete reabsorption of filtered amino acids is a fundamental transport function of the proximal tubule. Their transepithelial transport from the tubular lumen, across the cell and into the circulation, is driven by sodium electrochemical gradient. On the luminal membrane of tubular cells, there are three types of sodium-dependent amino acid co-transporters (belonging to solute carrier transporters family; SLC): one for **neutral amino acids** (glycine, alanine, proline), one for **acidic amino acids** (glutamate, aspartate) and one for **basic amino acids** (cystine, arginine, ornithine, lysine). The transport out of the cell involves both co-transporters and antiporters (the latter also belonging to the SLC family). Individual amino acids can be transported by more than one transporter, providing backup capacity if the transport is partially impaired. Elevated urine amino acid excretion may be caused by mutation in the gene(s) encoding for a particular amino acid transporter, as happens in **cystinosis**.

Only small amounts of amino acids (0.7 g/24 h) are normally present in the urine. Each substance, which is reabsorbed in the renal tubules, has its own renal transport maximum (T_{max}). T_{max} can be exceeded either when the amount of filtered substance becomes too large to handle or when the tubular cells do not function properly. Thus, aminoaciduria may result from the accumulation of amino acids such as phenylalanine, leucine, isoleucine, and valine in the plasma, or from an impaired tubular function. Aminoaciduria may also be symptom of proximal tubule damage (the **Fanconi syndrome**).

Renal handling of phosphate

Homeostasis of inorganic phosphate (Pi) depends mainly on the balance between the intestinal absorption of Pi and its urinary excretion. Adaptation of renal Pi handling to changing dietary Pi content is also well documented. The renal handling of Pi is regulated by hormonal and nonhormonal factors. Changes in urinary excretion of Pi are usually due to changes in the activity of the renal tubular transport system.

Approximately 90% of plasma phosphates are ultrafiltrable in the glomeruli (10% of Pi is bound to plasma albumin). Reabsorption of the filtered Pi occurs in the proximal, and also to some extent in the distal, tubules. The transport of Pi across tubular cells depends on the magnitude of sodium gradient formed by Na^+/K^+-ATPase in the basolateral membrane. However, the rate-limiting factor for Pi transport is the abundance of specific membrane transport proteins.

The transport of Pi across brush border membrane of the renal proximal tubules is mediated by the two Na^+-dependent Pi co-transporter proteins, NaPi-

IIa and NaPi-IIc (NaPi-IIb is also highly abundant in the brush border membrane of the small intestine). Na-Pi IIa is electrogenic, because the transport of one divalent Pi ion is coupled with three sodium ions ($3Na^{+}$:HPO_4^{2-}). NaPi-IIc, in contrast, is electroneutral and couples two sodium ions with one divalent Pi ion ($2Na^{+}$:HPO_4^{2-}). The expression of NaPi-IIa and NaPi-IIc adjusts the renal reabsorption of Pi to the organism's needs. For example, dietary restriction of Pi intake increases the quantity of NaPi-IIa and NaPi-IIc proteins in the luminal membrane. In contrast, high Pi diet leads to a decrease in the transporter expression. The membrane expression of NaPi-IIa is regulated by **parathyroid hormone (PTH)**. PTH receptors are located on both the basolateral and luminal membranes of the tubular cells. PTH activates adenylate cyclase via receptors located on basolateral membrane and increases the intracellular cyclic AMP. cAMP, in turn, activates protein kinase A, which phosphorylates NaPi-IIa transporter, leading to its internalization (a shift into the cytosol). Also, through the binding to luminal receptors, PTH causes phosphorylation of NaPi-IIa by phospholipase C and protein kinase C, and, again, the subsequent internalization of NaPi-IIa transporter. All this leads to a decreased availability of transporters for the filtered Pi, and its urinary excretion increases.

Clinical box Inherited nephron transport disorders

Gitelman's syndrome is a result of inactivating mutations in the gene encoding the thiazide-sensitive sodium–chloride co-transporter (gene SLC12A3). It is an autosomal-recessive disorder. Homozygous individuals are generally normotensive. The observed biochemical abnormalities are similar to thiazide-induced side effects (e.g. hypochloremic metabolic alkalosis, hypokalemia, hypocalciuria and sometimes hypomagnesemia).

Bartter's syndrome is a group of inherited defects in ion transport along the thick ascending limb of the loop of Henle. The neonatal Bartter's syndrome is linked to mutation in the furosemide-sensitive sodium–potassium–chloride co-transporter gene (SLC12A2) or the thick ascending limb potassium channel gene (ROMK/KCNJ1). The classic Bartter's syndrome results from chloride channel gene (CLCNKB) mutation. Clinical symptoms include polyuria and polydypsia, and there also is hypokalemia and alkalosis.

Cystinosis is a result of autosomal-recessive inactivating mutations in the genes SLC3A1 or SLC79A related to the transport of dibasic amino acids. SLC3A1 encodes dibasic amino acids transporter (rBAT), which forms a heterodimer with SLC7A9 gene product (B0, +AT1). In effect, there is a liposomal storage of cystine, and impairment of proximal tubule function, which lead to a decreased glucose and phosphate reabsorption.

Drug transport by organic anion transporters

The kidney eliminates waste products through filtration and secretion into the tubular fluid. It also eliminates water-soluble xenobiotics, *e.g.* drugs and their metabolites. The transport takes place in the proximal tubules, and is facilitated by substrate-specific tubular transporters. The most important in this are the organic anion transporters (OATs) which belong to the families of ATP-binding cassette (ABC) transporters or the solute carrier transporters (SLC).

The ABC transporters are transmembrane proteins that utilize the energy of ATP hydrolysis to carry various substrates across membranes, *e.g.* metabolites, inorganic ions, lipids, sterols, peptides and drugs. Classification of the ABC transporters is based on their amino acids sequence and the organization of the ABC domain (Chapter 17 and 18). The ABC transporters are expressed in various organs – *e.g.* the kidneys, liver and the intestine – and play a role in tumor resistance, **cystic fibrosis** (Chapter 10) and other inherited diseases, as well as in the development of resistance to multiple drugs (such as colchicine). Competition of drugs for the transporters may lead to toxic effects.

The SLC superfamily includes more than 300 proteins, which possess the ability to exchange the extracellular organic ions for intracellular ones. The SLC transporters are subclassified into organic cation transporters (OCTs – *e.g.* nicotine, quinine), organic anion transporters (OATs – transporting, e.g., lactate and succinate), and organic zwitterion/cation transporters (OCTNs). These proteins are expressed on basolateral and luminal membranes of various epithelial tissues, including the kidney and the liver.

Urine

The kidneys excrete from 0.5 L to more than 10 L of urine daily, the average daily volume being 1–2 L. The minimum volume necessary to remove the products of metabolism (mainly nitrogen, excreted as urea) is approximately 0.5 L/24 h. The osmolality of the glomerular filtrate is about 300 mmol/L and the osmolality of urine varies from about 80 to 1200 mmol/L. Thus, the maximal urine concentration, is approximately fourfold. Conversely, to excrete excess water, urine may be diluted to below the osmolality of plasma.

Urine analysis provides clinically important information

Urine analysis (urinalysis) in clinical laboratories includes testing for the presence of protein, glucose, ketone bodies, bilirubin, and urobilinogen, and for traces of blood. Measuring urinary osmolality assesses the concentrating capacity of the kidney. The urine is also tested for the presence of leukocytes and various crystals and deposits. Specialist investigations include analysis of urinary amino acids, hormones and other metabolites.

Only traces of protein are normally detectable in the urine. This increases when the glomeruli are damaged: the presence of significant amounts of protein in urine is an important sign of **renal disease**. Even a minimal amount of albumin in the urine (microalbuminuria) predicts the development of diabetic nephropathy (Chapter 21). Larger proteins such as immunoglobulins appear in the urine when the damage is more extensive: the immunoglobulin light chains (the Bence Jones protein) are present in urine in **multiple myeloma** (Chapter 4). In a **hemolytic anemia**, urine may contain free hemoglobin and urobilinogen. The presence of myoglobin is a marker of muscle damage **(rhabdomyolysis)**. The measurement of urine glucose and ketones is important in the assessment of glycemic control in **diabetic** patients (Chapter 21). Measurements of urobilinogen and bilirubin help to assess **liver function** (Chapter 29).

Assessment of renal function

Glomerular filtration rate is the most important characteristic describing kidney function

The renal clearance is the volume of plasma (in milliliters) that the kidney clears of a given substance every minute. The glomerular filtration rate (GFR) is the most important characteristic describing kidney function. GFR could be estimated by measuring the clearance of a substance, such as the polysaccharide inulin, that is neither secreted nor reabsorbed in the renal tubules. The amount of inulin filtered from plasma (i.e. its plasma concentration, P_{in}, multiplied by GFR) equals the amount recovered in the urine (i.e. its urinary concentration, U_{in}, multiplied by the urine formation rate, V):

$$P_{in} \times \mathrm{GFR} = U_{in} \times V \quad \textbf{(1)}$$

From this we calculate the GFR:

$$\mathrm{GFR} = U_{in} \times V / P_{in} \quad \textbf{(2)}$$

The average GFR is 120 mL/min in men and 100 mL/min in women. The renal clearance of inulin equals the GFR.

Serum creatinine and urea are first-line tests in the diagnosis of renal disease

To administer inulin intravenously every time one would want to assess the GFR is impractical. In clinical practice, we use the creatinine clearance instead.

Creatinine is derived from skeletal muscle phosphocreatine

The clearance of creatinine is similar to that of inulin. Although some creatinine is reabsorbed in the renal tubules, this is compensated by an equivalent tubular

secretion. To calculate creatinine clearance, one needs a sample of blood, and urine collected over 24 hours. The concentrations of creatinine in serum (P_{Cre}) and urine (U_{Cre}) are measured first. Urine excretion rate is calculated by dividing the urine volume by the collection time (*V*, above). The creatinine clearance is then calculated according to the formula:

$$\text{Creatinine clearance} = U_{Cre} \times V / P_{Cre}$$

Serum concentration of creatinine is 20–80 mmol/L (0.28–0.90 mg/dL). An increase in serum creatinine concentration reflects the decrease in GFR: serum creatinine concentration doubles when the GFR decreases by 50%. Another test used to assess kidney function is measurement of serum **urea** concentration. However, because urea is the end product of protein catabolism, its level in plasma is also dependent on factors such as the dietary protein intake and the rate of tissue breakdown.

In clinical practice, serum urea and creatinine are first-line tests in the diagnosis of **renal failure** (Fig. 23.8). Renal failure leads to a decrease in urine volume and creatinine clearance, and to an increase in serum urea and creatinine. The refinements in laboratory testing include standardization of creatinine measurements by making methods used in clinical laboratories traceable to the reference method, which is isotope dilution mass spectroscopy.

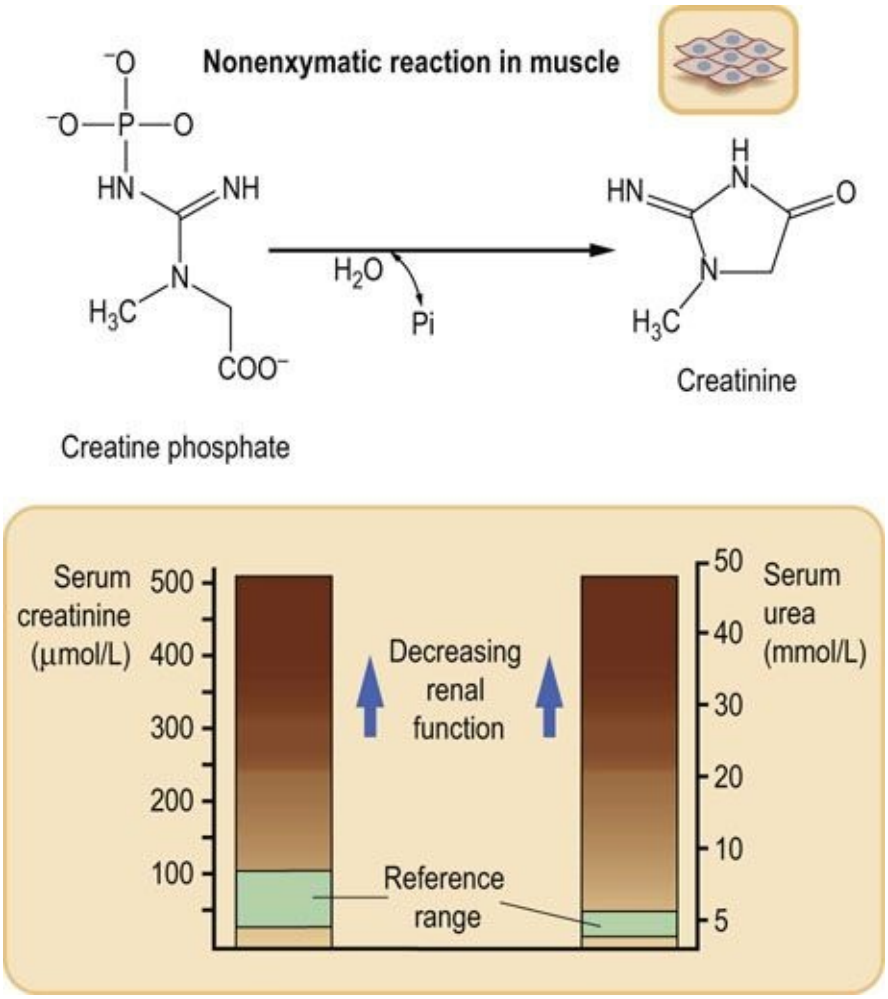


FIG.23.8 Serum urea and creatinine concentrations are important markers of renal function.
The upper panel shows the conversion of muscle phosphocreatine to creatinine. Loss of 50% of nephrons results in approximate doubling of serum creatinine concentration.

Serum cystatin C concentration is another marker of the GFR

Cystatin C is a 122-amino acid, 13-kDa protein belonging to the family of cysteine proteinase inhibitors. It is a product of the housekeeping gene expressed in all nucleated cells, and is produced at a constant rate. Because of its small size and basic isoelectric point, cystatin C is freely filtered through the glomerulus. It is not secreted by the tubules and although it is reabsorbed, it is subsequently catabolized and therefore does not return to plasma. Its serum concentration is not significantly affected by age, and therefore it is a preferential marker of GFR

in children. However, other factors, independent of the GFR, such as the inflammatory phenomena, may affect serum cystatin C concentration.

Neutrophil gelatinase-associated lipocalin is a novel biomarker of acute kidney injury

Neutrophil gelatinase-associated lipocalin (NGAL) is a 178 amino acid, 25-kDa protein belonging to lipocalin family (Greek *lipos*, 'fat, grease' and *calyx*, 'cup'). This protein is produced in various epithelial cells (kidney, liver, lung, intestine) and in myelocytes, and is a component of innate immunity. It also plays a role in tissue remodeling, especially after heterodimerization with the metalloproteinase 9 (MMP-9, collagenase IV). NGAL synthesis and secretion into the bloodstream is immediately increased after tissue damage.

Serum NGAL is freely filtered through the glomeruli and is reabsorbed in the proximal tubules. The rising serum NGAL during acute kidney injury may reflect reduced GFR. Though the major source of the urinary NGAL is the distal nephron, a small fraction may come from the filtered pool, escaping tubular reabsorption, due to proximal injury.

Recent studies demonstrated a quantitative link between cellular stress, expression of NGAL in the kidney and excretion of NGAL in urine. The cells in the distal nephron are activated to express NGAL within a few hours following damage (e.g. by ischemia, hypoxia, drug toxicity), before another nephron segment may be affected. Urine NGAL is also being evaluated as a sign of the earliest stage of acute kidney injury, before there are evident clinical abnormalities. The limitations of urine NGAL measurements as biomarker of an acute renal injury are that it may be affected by coexisting chronic kidney disease, renal inflammation and urinary tract infections.

Clinical box A 25-year-old man admitted after a motorcycle accident: acute renal failure

A 25-year-old man was admitted to hospital unconscious after a motorcycle accident. He had evidence of shock with hypotension and tachycardia, a fractured skull and multiple injuries to his limbs. Despite treatment with intravenous colloid and blood, he showed persistent oliguria (urine output 5–10 mL/h; oliguria is <20 mL/h).

On the third day, his serum creatinine concentration had risen to

300 µmol/L (3.9 mg/dL) and his urea concentration to 21.9 mmol/L (132 mg/dL). eGFR was 22 mL/min/1.73m^2. Reference values are:
Serum creatinine: 20–80 µmol/L (0.23–0.90 mg/dL)
Serum urea: 2.5–6.5 mmol/L (16.2–39 mg/dL).
eGFR: See table 23.2

Comment.

This young man has developed acute renal failure due to acute tubular necrosis as a consequence of hypovolemic shock. He subsequently underwent emergency hemofiltration. Renal function started to recover after 2 weeks with an initial increase in urine volume, the so-called ‘diuretic phase’.

Clinical box Diabetes often leads to impairment of renal function

A 37-year-old woman with a 12-year history of type 1 diabetes came for a routine visit to the diabetic clinic. Her glycemic control was poor and glycated hemoglobin (HbA_{1c}) was 8% (64 mmol/mol). Blood pressure was mildly raised at 145/88 mmHg. A quantitative measurement of albumin in urine revealed protein concentration of 5 mg/mmol creatinine, indicating microalbuminuria. Reference values are:
HbA_{1c}: desirable value below 7% (53 mmol/mol).
Urine microalbumin: less than 3.5 mg/mmol creatinine.

Comment.

This patient had mildly impaired renal function and raised blood pressure as a result of glomerular damage from diabetes. The presence of microalbuminuria predicts future overt diabetic nephropathy.

Estimated GFR (eGFR)

Creatinine clearance changes with age, body surface and gender, and also varies with race. Also, the relationship between GFR and creatinine concentration may differ between healthy population and patients with renal disease. In current practice, the GFR values are being adjusted using formulas that include, apart from serum creatinine concentration, factors such as age, gender, weight and race.

Different equations provide useful estimates of GFR: in adults, the equations used are the one developed by the Modification of Diet in Renal Disease Study Group (MDRD) and the Cockcroft–Gault equation; in children, the Schwartz and Counahan–Barratt equations are used. The so-called "abbreviated MDRD formula" includes four variables: serum creatinine, age, sex and race. The original MDRD six-variable equation included serum albumin and serum urea in addition to these. A detailed discussion of these formulas is beyond the scope of this text – the reader is referred to the Further Reading section.

The eGFR equations are useful in the detection of renal impairment. Calculation of the estimated GFR is recommended for the identification and classification, as well as for screening and monitoring, of chronic kidney disease. The severity of chronic kidney disease is classified into six stages (Table 23.2).

Table 23.2

The staging of chronic kidney disease

Stage	Description	eGFR (mL/min/1.73 m^2)
1	Normal kidney function but urine findings or structural abnormalities of the kidney*	≥90
2	Slightly decreased GFR	60–89
3	Moderate decrease in GFR	30–59
4	Severe decrease in GFR	15–29
5	End-stage kidney failure or dialysis	<15

The severity of chronic kidney disease is categorized in six stages. The eGFRs used in this classification has been derived from the abbreviated MDRD equation. *Proteinuria, albuminuria, hematuria lasting for at least 3 months and/or structural abnormalities

Reference: KDOQI Clinical Practice Guidelines for Chronic Kidney Disease: Evaluation, Classification, and Stratification. (see Further Reading for details)

Kidney Disease Outcome Quality Initiative, *Am J Kidney Dis* 39(suppl. 1):S1–S266, 2002.

One should note though, that in individuals with exceptional dietary intake (vegan, vegetarian diet, creatinine supplements) or abnormal muscle mass (resulting from amputation, malnutrition, muscle wasting), the renal function should still be assessed by measuring creatinine clearance.

Summary

■ The kidneys maintain electrolyte and water homeostasis and thus play a critical role in the regulation of the composition and volume of the extracellular fluid. Sodium handling by the kidney is also a major determinant of the arterial blood pressure.

■ The kidneys are exposed to mechanical stress induced by hypertension, to hyperglycemia, and to various nephrotoxins, which may damage renal cells and impair kidney function. Therefore, assessment of renal function is an important part of clinical examination.

■ Serum concentrations of urea and creatinine are key tests in the assessment of renal function. For more accurate assessment of renal function, the clearance of creatinine is assessed.

■ Proteinuria is a symptom of renal filtration barrier damage. Microalbuminuria is an early marker of nephropathy; NGAL in urine is an early marker of acute kidney injury.

Further reading

Broer, S. Amino acid transport across mammalian intestinal and renal epithelia. *Physiol Rev*. 2008; 88:249–286.

Burckhardt, G. Drug transport by Organic Anion Transporters (OATs). *Pharmacol Thera*. 2012; 136:106–130.

Chadha, V, Alon, US. Hereditary renal tubular disorders. *Semin Nephrol*. 2009; 29:399–411.

Devarajan, P. Neutrophil gelatinase-associated lipocalin: a promising biomarker for human acute kidney injury. *Biomark Med*. 2010; 4:265–280.

Haraldsson, B, Nystrom, J, Deen, WM. Properties of the glomerular barrier and mechanisms of proteinuria. *Physiol Rev*. 2008; 88:451–487.

Mather, A, Pollock, C. Glucose handling by the kidney. *Kidney Int*. 2011; 79(suppl 120):S1–S6.

Murer, H, Biber, J. Phosphate transport in the kidney. *J Nephrol*. 2010; 23(Suppl 16):S145–S151.

Schlaich, MP, Sobotka, PA, Krum, H, *et al.* Renal denervation as a therapeutic approach for hypertension. *Hypertension*. 2009; 54:1195–1201.

Websites and downloads

KDOQI Clinical practice guidelines for chronic kidney disease. evaluation, classification, and stratification. Part 5. Evaluation of laboratory measurements for clinical assessment of kidney disease. Guideline 4. Estimation of GFR www.kidney.org/professionals/kdoqi/guidelines_ckd/p5_lab_g4.htm. Kidney

Disease Outcome Quality Initiative. *Am J Kidney Dis*. 2002; 39(suppl. 1):S1–S266.

KDOQI Clinical practice guidelines for chronic kidney disease: evaluation, classification, and stratification. Part 5. Evaluation of laboratory measurements for clinical assessment of kidney disease. Guideline 4. Estimation of GFR www.kidney.org/professionals/kdoqi/guidelines_ckd/p5_lab_g4.htm.

MDRD eGFR calculator. www.nephron.com/MDRD_GFR.cgi.

UniProt Protein knowledge base. www.uniprot.org/.

CHAPTER 24

Water and Electrolyte Homeostasis

Marek H. Dominiczak and Mirosława Szczepańska-Konkel

Learning objectives

After reading this chapter you should be able to:

- Describe the body water compartments in the adult and the composition of the main body fluids.
- Explain the role of albumin in the movement of water between plasma and interstitial space, including the consequences of proteinuria.
- Describe how sodium influences the movement of water between the extracellular and intracellular space.
- Explain why sodium/potassium ATPase is essential for normal cell hydration, and comment on the consequences of its inhibition.
- Describe factors affecting plasma potassium concentration.
- Discuss the links between sodium and water homeostasis.
- Describe the clinical assessment of water and electrolyte status.

Introduction

Water and electrolytes are constantly exchanged with the environment, and their body content depends on the balance between intake and loss

Water is essential for survival and accounts for approximately 60% of the body weight in an adult person. This changes with age: it is about 75% in the newborn and decreases to below 50% in older individuals. Water content is greatest in the brain tissue (about 90%) and least in the adipose tissue (10%).

The stability of subcellular structures and activities of numerous enzymes are dependent on adequate cell hydration. The maintenance of ion gradients and electrical potential across membranes is also crucial for survival and underlies muscle contraction, nerve conduction and secretory processes (Chapter 8).

Both water deficiency and water excess impair function of organs and tissues. Water balance (daily intake and loss) and water distribution between cells and the surrounding fluid are subject to complex regulation.

Clinical relevance

Water and electrolyte disorders are common in clinical practice. Historically, textbooks of biochemistry (and many biochemistry courses) have treated water and electrolyte balance somewhat peripherally. And yet it is fundamental to maintaining metabolism, and underlies many essential treatment procedures.

Body water compartments

Approximately two-thirds of total body water is in the intracellular fluid (ICF), and one-third remains in the extracellular fluid (ECF). ECF consists of interstitial fluid and lymph (15% body weight), plasma (3% body weight), and the so-called transcellular fluids, which include gastrointestinal fluid, urine and cerebrospinal fluid (CSF) (Fig. 24.1

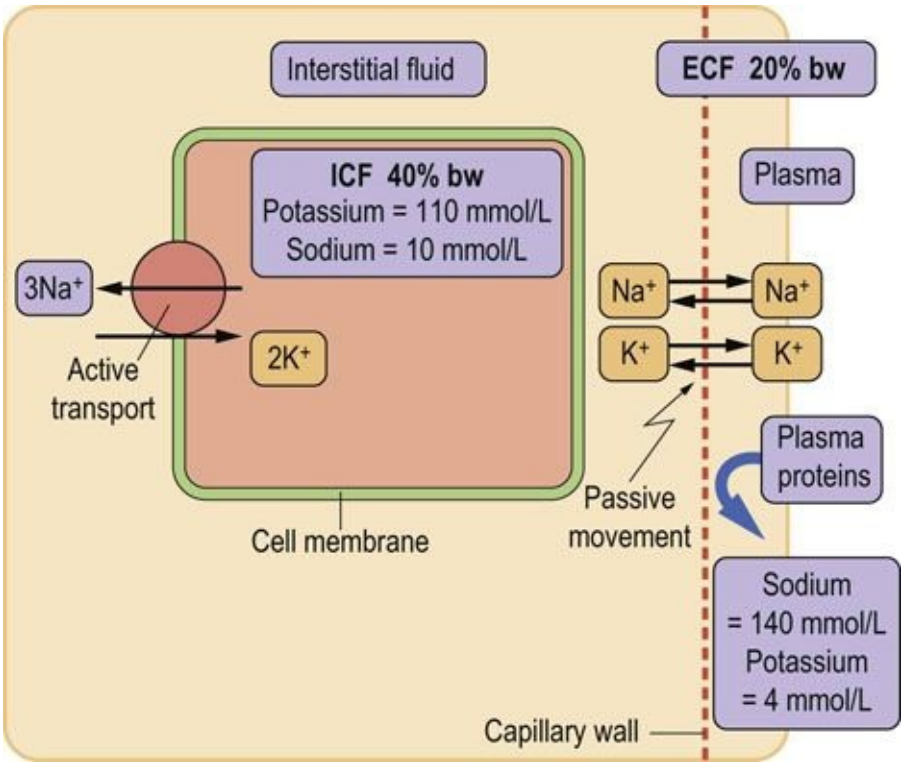


FIG. 24. 1 Distribution of body water, sodium and potassium.
The main body water compartments are the intracellular fluid (ICF) and the extracellular fluid (ECF). ECF includes interstitial fluid and plasma. The gradient of sodium and potassium concentrations is maintained across cell membranes by the Na^+/K^+-ATPase. Sodium is a major contributor to the osmolality of the ECF, and a determinant of the distribution of water between ECF and ICF. Distribution of water between plasma and interstitial fluid is determined by the oncotic pressure exerted by plasma proteins. bw, body weight.

Clinical box Concentration of ions

in plasma and serum

All physiologic phenomena occur in plasma – and therefore discussion of physiologic or pathologic conditions relates to plasma concentrations of ions.

However, in practice, the concentration of most ions is measured after the sampled blood has been allowed to clot (i.e using serum). Therefore, in the discussion of laboratory results we often mention serum values (Chapter 4).

). Two barriers are important for the understanding of exchanges taking place between different compartments: the cell membrane and the wall of the capillary vessel.

Capillary vessel wall separates plasma from the surrounding interstitial fluid

The capillary wall separates plasma from the interstitial fluid and is freely permeable to water and electrolytes but not to proteins. Ions and low-molecular-weight molecules are present in similar concentrations in the ECF and plasma but protein concentration is 4–5 times greater in plasma than it is in the interstitial fluid. The total plasma concentration of cations is about 150 mmol/L, of which sodium is approximately 140 mmol/L and potassium 4 mmol/L. The most abundant plasma anions are chloride and bicarbonate, with average concentrations of 100 mmol/L and 25 mmol/L, respectively (Fig. 24.2). In clinical practice the rest of the anions are, for the purposes of electrolyte balance, considered together, constituting the so-called anion gap (AG), which is calculated as follows.

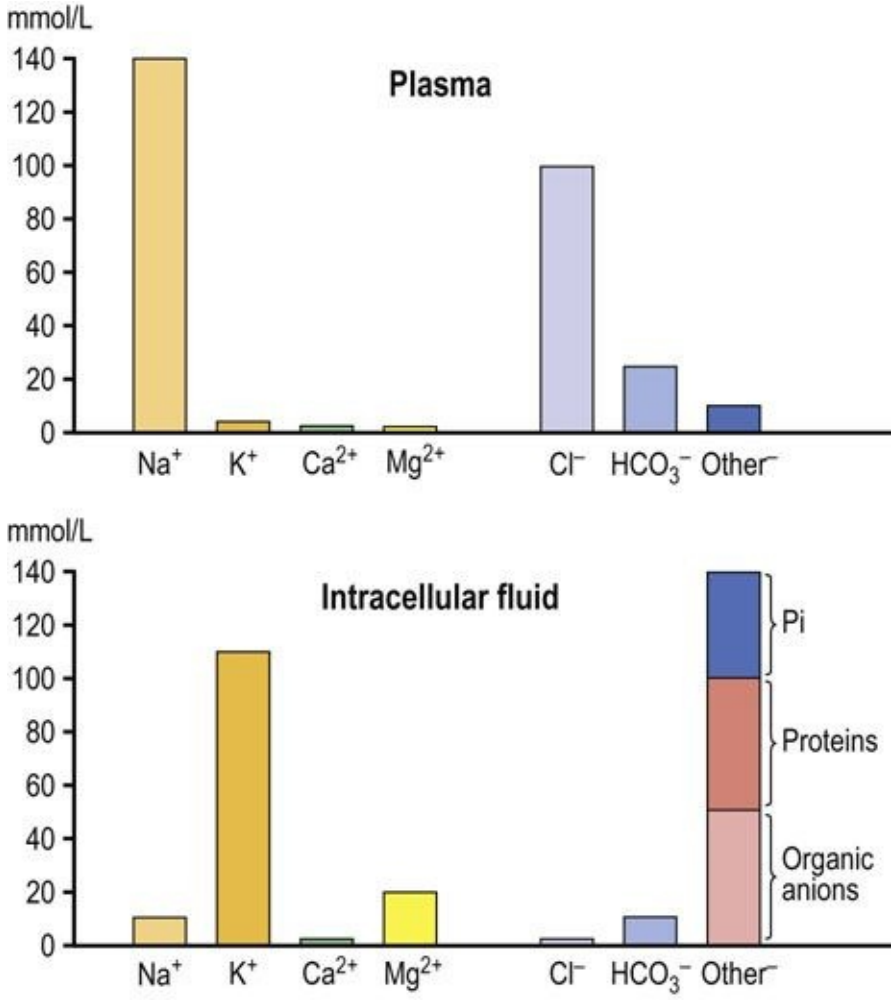


FIG. 24. 2 Ions present in the plasma and in the intracellular fluid.
The most important ions in plasma are sodium, potassium, calcium, chloride, phosphate, and bicarbonate. Sodium chloride, in a concentration close to 0.9% (thus 'physiologic saline'), is the main ionic component of the extracellular fluid. Potassium is the main intracellular cation. Glucose and urea also contribute to plasma osmolality. Their contribution is normally small, because they are present in plasma in relatively low molar concentrations (about 5 mmol/L each). However, when glucose concentration increases in diabetes, its contribution to osmolality becomes significant. Plasma urea increases in renal failure but it does not contribute to water movement between ECF and ICF because it freely crosses cell membranes. The main intracellular cation is potassium and the main anions are phosphates and proteins. There is also a substantial amount of magnesium in cells.

$$AG = (Na^+ + K^+) - (Cl^- + HCO_3^-)$$

The anion gap (in a healthy person, approximately 10 mmol/L) includes phosphate, sulfate, protein, and organic anions such as lactate, citrate, pyruvate, acetoacetate, and β-hydroxybutyrate. However, it may increase several-fold in conditions where inorganic and organic anions accumulate, *e.g.* in renal failure or diabetic ketoacidosis. For this reason it is clinically important.

Plasma membrane separates the intracellular and extracellular fluid

In the ICF, the main cation is potassium, present in a concentration of about 110 mmol/L. This is almost 30-fold greater than its concentration in the ECF and in plasma (4 mmol/L). The main anions in the ICF are proteins and phosphate. In the ECF, the situation is reversed: the main cation is sodium, present in a concentration of about 145 mmol/L. In the ICF the concentration of sodium (and chloride) is only 10 mmol/L.

Water diffuses freely across most cell membranes but the movement of ions and neutral molecules is restricted

Small molecules are transported across cell membranes by specific transport proteins, the ion pumps. The most important is the sodium-potassium ATPase (Na^+/K^+-ATPase), also referred to as the sodium–potassium pump.

Na^+/K^+-ATPase maintains the sodium and potassium gradients across the cell membrane

Na^+/K^+-ATPase can be considered either as an ion transporter (sodium pump) or as an enzyme (ATPase). It maintains chemical and electrical potential gradients (it is electrogenic) across the cell membrane (Fig. 24.3). It hydrolyzes one ATP molecule, and the released energy drives the transfer of three sodium ions from the cell to the outside, and two potassium ions from the outside into the cell (Fig. 24.4).

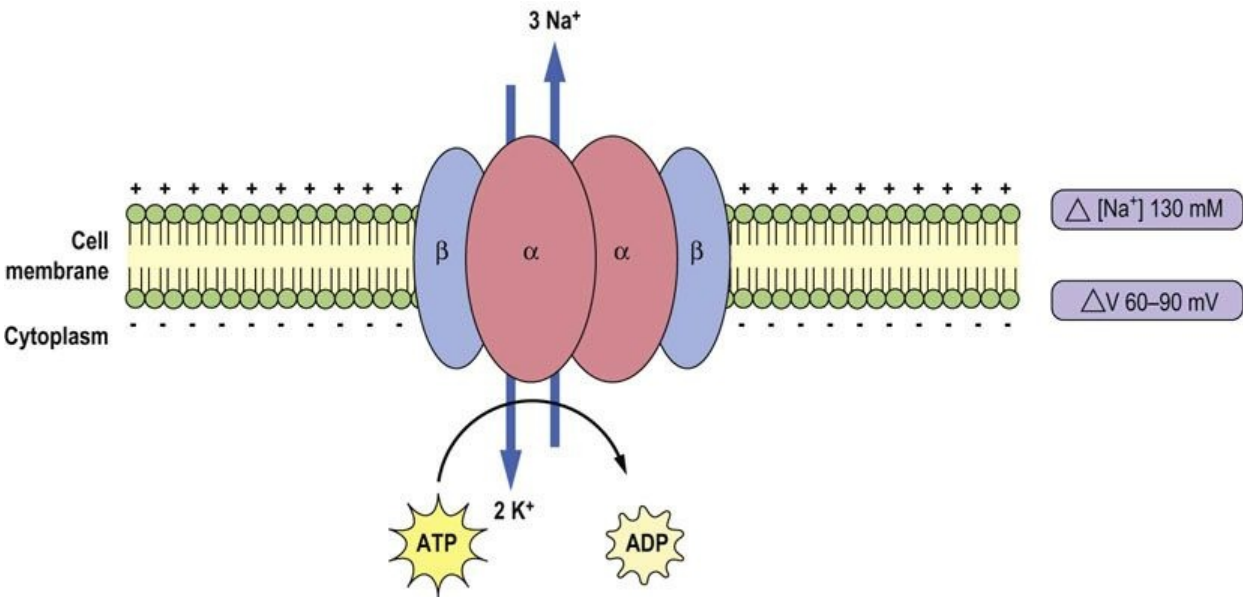


FIG. 24.3 Na^+/K^+-ATPase (the sodium–potassium pump) generates transmembrane potential and ion concentration gradients across the cell membrane.
The transmembrane difference in sodium concentration (ΔNa^+), and transmembrane voltage difference (ΔV) are shown on the right. For each molecule of hydrolyzed ATP, it moves two potassium ions into the cell and three sodium ions out of the cell. Na^+/K^+-ATPase consists of two main subunits – catalytic subunit (α) and structural subunit (β).

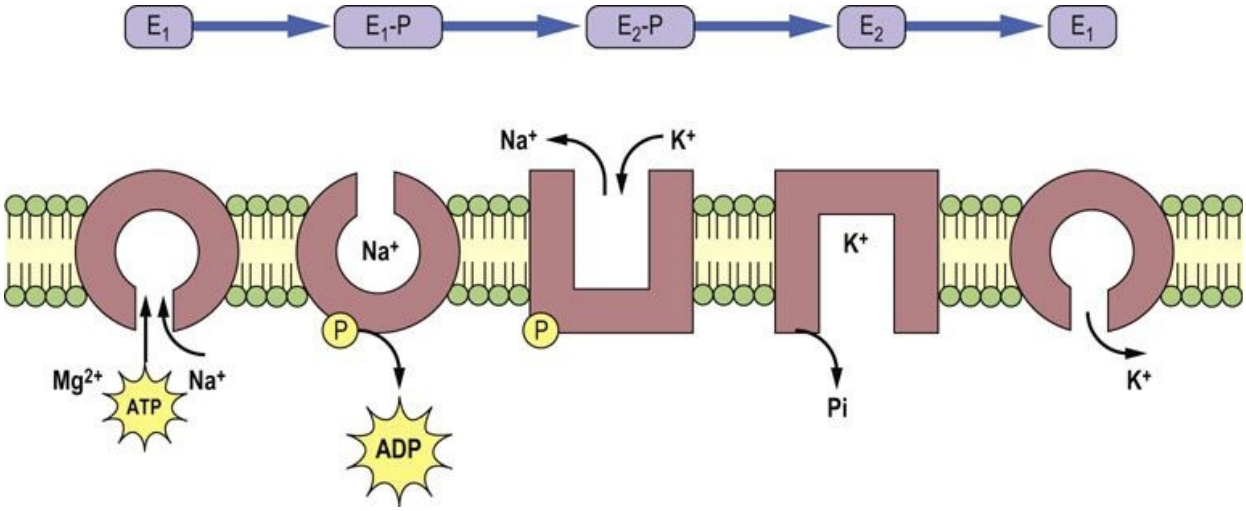


FIG. 24.4 Catalytic function of the Na^+/K^+-ATPase.
The catalytic subunit of the Na^+/K^+-ATPase can be either phosphorylated (E_1-P and E_2-P) or dephosphorylated (E_1 and E_2) and this changes its conformation and affinity towards substrates. The E_1 form exhibits high affinity towards ATP, magnesium and

sodium and low affinity towards potassium, whereas the E_2 form exhibits high affinity for potassium and low for sodium. After the release of ADP, there is conformational change from E_1-P to E_2-P. This promotes extracellular delivery of sodium and the binding of extracellular potassium. The latter process induces dephosphorylation of E_2-P and potassium release into the intracellular compartment.

The Na^+/K^+-ATPase is the major determinant of cytoplasmic sodium concentration (Chapter 8). It also has an important role in regulating cell volume, cytoplasmic pH and calcium levels through the Na^+/H^+ and Na^+/Ca^{2+}exchangers. One of the primary requirements for continuous sodium-pump-driven adaptation comes from changes in dietary sodium and potassium. Hormones that control the volume and ionic composition of the ECF often act directly on the sodium pump in the kidney and intestine.

Na^+/K^+-ATPase activity is subject to short-and long-term regulation by a number of hormones

The Na^+/K^+-ATPase is activated by sodium and ATP at cytoplasmic sites. The structures of the catalytic subunit of the enzyme and its phosphorylation sites are shown in Fig. 24.5. Half-maximal activation of the enzyme by intracellular sodium occurs at sodium concentration of 10–40 mM. which is often above the steady-state concentration. Accordingly, small changes in the cytoplasmic sodium concentration can have large effect on its activity. Some hormones appear to alter the Na^+/K^+-ATPase activity by changing its apparent affinity for sodium (for instance, angiotensin II and insulin increase the affinity).

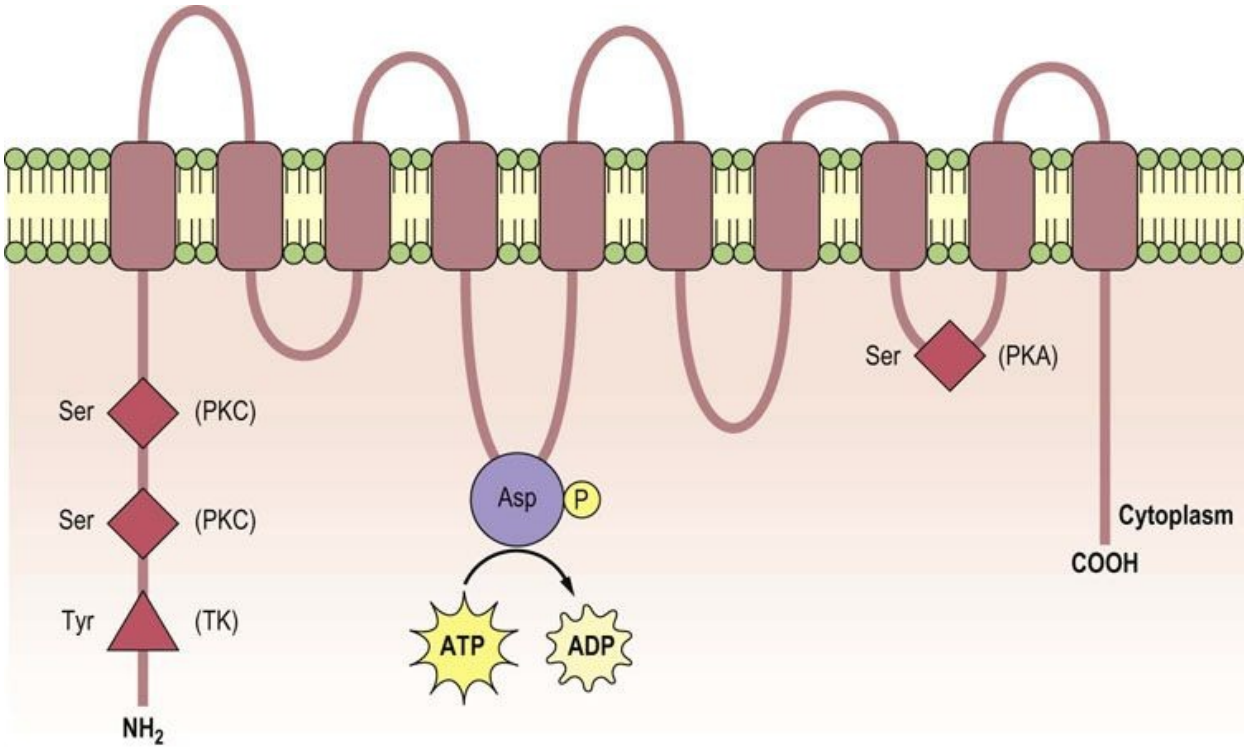


FIG. 24.5 Structure of the α-subunit of the Na^+/K^+-ATPase.
The α-subunit consists of 10 membrane-spanning domains (M1–M10) with intracellular amino-and carboxy-terminal domains. The ATP-binding domain and the phosphorylation site are located in the long M4–M5 cytoplasmic loop (where Asp residue undergoes phosphorylation). Other phosphorylation sites on serine and tyrosine residues are mediated by protein kinase A (PKA), protein kinase C (PKC) and tyrosine kinase (TK).

The Na^+/K^+-ATPase is subject to regulation by a number of hormones, including aldosterone. Short-term regulation involves either direct effects on the kinetic properties of the enzyme or its translocation between the plasma membrane and intracellular stores. Long-term regulatory mechanisms affect the enzyme's synthesis or degradation.

Peptide hormones such as vasopressin and PTH that act through G-protein-coupled receptors affect the activity of Na^+/K^+-ATPase. The G-proteins activate adenylyl cyclase, which generates cAMP. cAMP, in turn, activates protein kinase A (PKA). PTH, angiotensin II, norepinephrine and dopamine also trigger G-protein-mediated activation of phospholipase C, which activates protein kinase C (PKC). Both PKA and PKC affect Na^+/K^+-ATPase by serine phosphorylation of its α-subunit (Fig. 24.5). Insulin increases the apparent sodium affinity of Na^+/K^+-ATPase through the activation of the tyrosine kinase receptor and phosphorylation of the α-subunit of the enzyme.

Passive movement of electrolytes through ion channels is driven by the electrochemical gradient

For most cells, the membrane potential ranges from 50 to 90 mV, being negative inside the cell. The electrochemical gradient is a source of energy for transport of many substances such as the co-transport of the sodium ions with glucose (the SGLT transporter, Chapter 10), amino acids, and phosphate. Membrane depolarization also promotes an increase in intracellular calcium by activating voltage-dependent Ca^{2+}channels (Chapter 8). The role of ion gradients in nerve transmission is described in Chapter 41.

Because water and sodium transport on the luminal side of the epithelial cells (in the intestine and the kidneys) is linked to the ion gradient generated by the Na^{+}/K^{+}-ATPase, this enzyme is critical to water absorption in the intestine and its reabsorption in the kidneys. The impairment of the sodium pump function in the kidney and small intestine is linked to pathophysiology of hypertension and chronic diarrhea, respectively.

Osmolality: osmotic and oncotic pressures

Molecules dissolved in body water contribute to the osmotic pressure

Osmolality depends on the concentration of molecules in water and osmotic pressure is proportional to the molal concentration of a solution. One millimole of a substance dissolved in 1 kg H_2O at 38°C exerts an osmotic pressure of approximately 19 mmHg. Under physiologic conditions, the average concentration of all osmotically active substances in the ECF is 290 mmol/kg H_2O. Normally, the ICF osmolality is identical.

Differences in osmolality cause movement of water between intracellular and extracellular fluid

A change in the concentration of osmotically active ions in either ECF or ICF creates a gradient of osmotic pressure and, consequently, causes the movement of water. **Water always diffuses from lower osmolality to higher to equalize osmotic pressures** (Fig. 24.6).

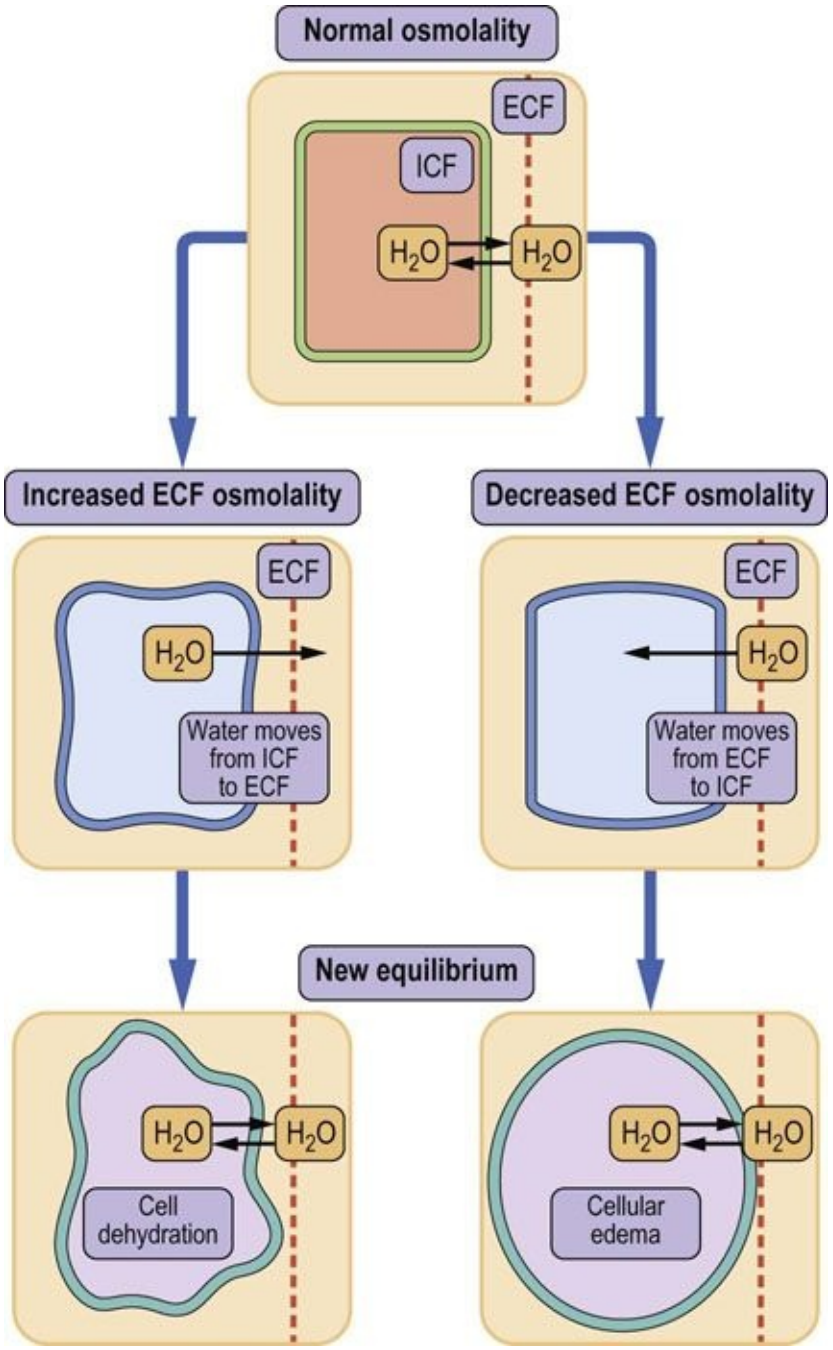


FIG. 24.6 Water redistribution caused by changes in osmolality.
Osmotic pressure controls the movement of water between compartments. An increase in ECF osmolality draws water from the cells, and leads to cellular dehydration. On the other hand, when ECF osmolality decreases, water moves into the cells and this may cause cell edema. The arrows indicate direction of water movement. ECF, extracellular fluid; ICF, intracellular fluid.

Because sodium is the most abundant ion in the ECF, it is also the most important determinant of its osmolality. Glucose at its normal plasma

concentration (5 mmol/L; 90 mg/dL) does not contribute significantly to osmolality but it becomes its major determinant when its concentration increases in diabetes (remember the 'osmotic symptoms' in poorly controlled diabetes (Chapter 21).

Balance between the oncotic and hydrostatic pressure changes across the vascular bed and is fundamental for the circulation of substrates and nutrients

The movement of water between the plasma and interstitial fluid depends on the plasma protein concentration. Proteins, particularly albumin, exert osmotic pressure in the plasma (about 3.32 kPa; 25 mmHg). This is known as **the oncotic pressure** and it retains water in the vascular bed. It is balanced by **the hydrostatic pressure**, which forces fluid in the opposite direction, *i.e.* out of the capillaries. In the arterial part of the capillaries, the hydrostatic pressure prevails over the oncotic pressure, and water and low-molecular-weight compounds filter out into the extravascular space. In contrast, in the venous part of capillaries, oncotic pressure prevails over hydrostatic pressure, and fluid is drawn into the vascular lumen (Fig. 24.7). A reduction in plasma oncotic pressure, which occurs, for instance, as a consequence of a decrease in the plasma albumin concentration, results in the movement of fluid into the extravascular space and in the development of edema.

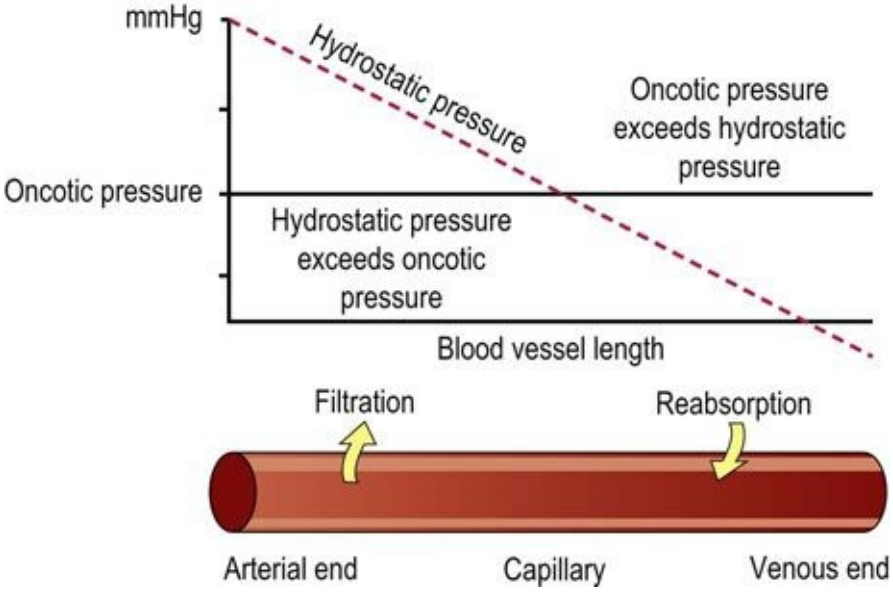


FIG. 24.7 Oncotic and hydrostatic pressures determine movements of fluid between plasma and interstitial fluid.

Clinical box Edema results from a loss of protein

An 8-year-old girl was referred to a nephrologist after it had been noticed that her face was puffy and her ankles became swollen over a period of about 2 weeks. Dipstick test for urine protein yielded a strongly positive (++++) result and measurement in a 24-hour collection showed protein excretion of 7 g/day. The reference value for urinary protein excretion is less than 0.15 g/day.

Comment.

The cause of the proteinuria was the damage to the renal filtration barrier. Renal biopsy showed the so-called minimal change disease, with the urinary protein loss causing, in turn, hypoalbuminemia and a decrease in the plasma oncotic pressure. This led to edema. The condition went into remission after treatment with a glucocorticoid.

Cells protect themselves against changes of osmolality and volume

An increase in the intracellular concentration of sodium stimulates the Na^+/K^+-ATPase, which extrudes sodium from the cell. This is followed by the egress of water and protects the cell from volume changes. Another protective mechanism is the intracellular generation of osmotically active substances. For instance, the brain cells adapt to increased ECF osmolality by increasing their amino acid concentration, and cells in the renal medulla exposed to a hyperosmotic environment produce an osmotically active alcohol, sorbitol, and increase the concentration of the amino acid taurine.

The body constantly exchanges water with the environment

In a steady state, the intake of water equals its loss. The main source of water is oral intake and the main source of its loss is urine excretion. Water is also lost through the lungs, sweat and feces: this is called the ‘insensible’ loss and in normal circumstances amounts to approximately 500 mL daily (Fig. 24.8). Insensible loss can increase substantially in high temperatures, during intensive exercise, and also as a result of fever. Checking the patient's fluid balance is one of the essential daily routines on medical and surgical wards.

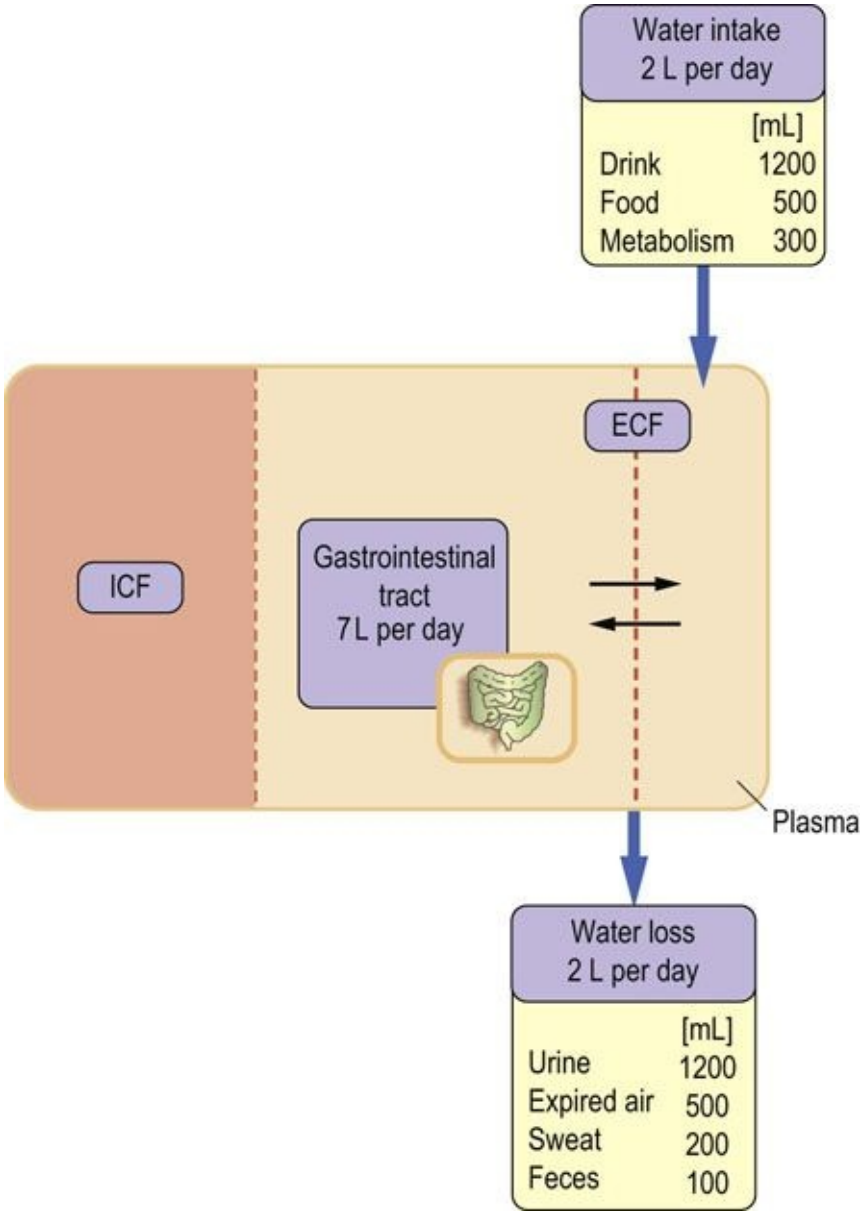


FIG. 24.8 Daily water balance in an adult person.

Water is obtained from the diet and from oxidative metabolism, and is lost through the kidneys, skin, lungs, and the intestine. Note how much water enters and leaves the gastrointestinal tract daily; this explains why severe diarrhea quickly leads to dehydration. See also Chapter 10.

Clinical box Body fluids differ in ionic composition

Clinical abnormalities that may develop after fluid loss depend on the composition of what is lost. For instance, **sweat** contains less sodium than extracellular fluid: therefore, excessive sweating leads to a predominant loss of water and 'concentrates' sodium in the extracellular fluid, causing hypernatremia. On the other hand, sodium content of the **intestinal fluid** is similar to that of plasma but contains considerable amounts of potassium. Thus, its loss (for instance, in severe diarrhea) would result in dehydration and hypokalemia, but may not change plasma sodium concentration (Table 24.1).

Table 24.1

Electrolyte content of the body fluids

	Sodium (mmol/L)	Potassium (mmol/L)	Bicarbonate (mmol/L)	Chloride (mmol/L)
Plasma	140	4	25	100
Gastric juice				
Small intestinal fluid	140	10	Variable	70
Feces in diarrhea	50–140	30–70	20–80	Variable
Bile, pleural, and peritoneal fluids	140	5	40	100
Sweat	12	10	–	12

Loss of fluid that has an electrolyte content similar to that of plasma leads to

dehydration with normal serum electrolyte concentrations. On the other hand, when the sodium content of the lost fluid is less than that of plasma (e.g. sweat), dehydration may be accompanied by hypernatremia. Overhydration is usually accompanied by hyponatremia.

Adapted with permission from Dominiczak MH, editor: *Seminars in Clinical Biochemistry*, ed. 2, Glasgow, 1997, Glasgow University.

Potassium

Monitoring serum potassium concentration is fundamentally important

Normal serum concentration of potassium is 3.5–5 mmol/L. Because its intracellular concentration is much higher than concentration in plasma, a relatively minor shift of potassium between the ECF and ICF may result in major changes in its serum concentration. Both high and low concentrations of potassium (hyperkalemia and hypokalemia, respectively) affect the cardiac muscle and can be life-threatening.

On the EKG, hyperkalemia can lead to the loss of P-wave, characteristic tall peaked T-waves, and widened QRS complexes. Hypokalemia, on the other hand, may prolong PR interval, cause peaked P-wave, flatten the T-wave and cause prominent U-waves.

Serum potassium concentration below 2.5 mmol/L or above 6.0 mmol/L is dangerous (Fig. 24.9). The most common cause of severe hyperkalemia is renal failure: in this condition potassium cannot be adequately excreted in the urine. On the other hand, low serum potassium usually results from excessive losses, either in urine or through the gastrointestinal tract. Kidneys account for more than 90% of the body potassium loss, and diuretic therapy can induce both hypo- and hyperkalemia. Changes in serum potassium concentration are also associated with acid–base disorders (Chapter 25).

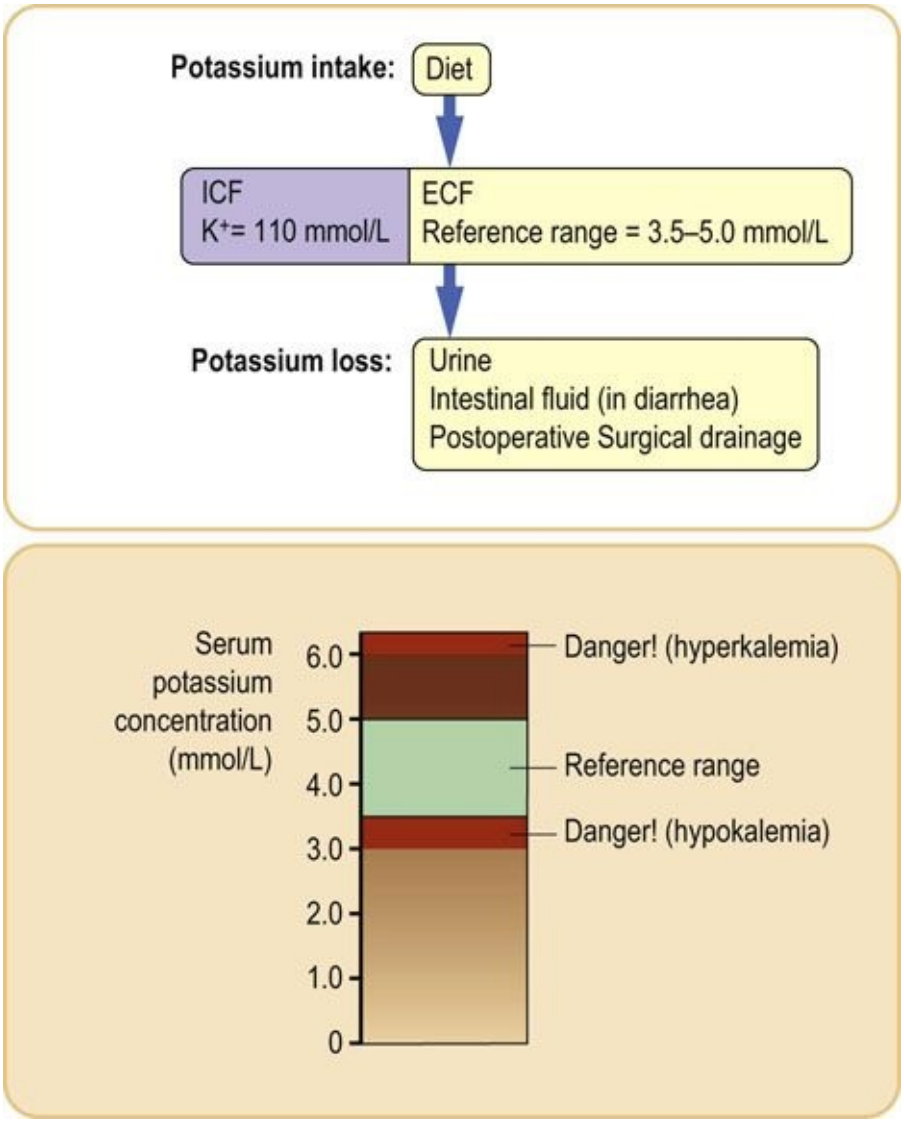


FIG. 24.9 Potassium balance.
Serum potassium concentration is maintained within narrow limits. Both low (hypokalemia) and high (hyperkalemia) concentrations are be dangerous, as potassium affects the contractility of heart muscle. Generally, serum potassium concentrations above 6.0 mmol/L and below 2.5 mmol/L are regarded as emergencies. The upper panel shows main sources of potassium loss.

Renin–angiotensin system

Renin–angiotensin system controls blood pressure and the vascular tone

Renin is an enzyme produced principally in the juxtaglomerular apparatus of the kidney; it is stored in the secretory granules and released in response to a decreased renal perfusion pressure (decreased delivery of Na^+ to the macula densa) and increased sympathetic tone. Renin is a protease that uses circulating angiotensinogen as its substrate. Renin secretion is regulated by pathways involving G-protein-coupled receptors and adenylate cyclase – PKA pathway, which activates the cAMP-responsive binding protein (CREB). CREB is a transcription factor which subsequently recruits its co-activators and binds to the cAMP-responsive element in the renin gene promoter, initiating transcription. Renin secretion is also stimulated by norepinephrine and prostaglandin E_2 (PGE_2).

Another signaling pathway involving the G-proteins and an increase in cytosolic calcium decreases the activity of the adenylate cyclase and inhibits renin secretion (this pathway is activated by angiotensin II, and endothelins I and II).

Angiotensinogen is a glycoprotein comprising more than 400 amino acids synthesized in the liver, and its different forms have variable structures and molecular mass. Renin cleaves a 10-amino acid peptide, angiotensin I, from the angiotensinogen. Angiotensin I then becomes a substrate for peptidyl-dipeptidase A (angiotensin-converting enzyme; ACE). ACE removes two amino acids from the angiotensin I, producing angiotensin II. This reaction can also be catalyzed by enzymes such as chymase and cathepsin. Another form of angiotensin, angiotensin 1–9, is formed by an isoform of ACE (ACE2) and is subsequently degraded to angiotensin 1–7. The latter can also be formed from angiotensin II by endopeptidases. Substantial amounts of angiotensin II are formed in the kidney. Juxtaglomerular cells contain ACE, angiotensin I and angiotensin II. Angiotensin II is also synthesized in the glomerular and tubular cells, and is secreted into the tubular fluid and the interstitial space. Angiotensin II receptors are present on the tubular and renal vascular cells: therefore, locally produced angiotensin II probably influences tubular reabsorption and renal

vascular tone through autocrine and paracrine action. The renin–angiotensin system is illustrated in Figure 24.10.

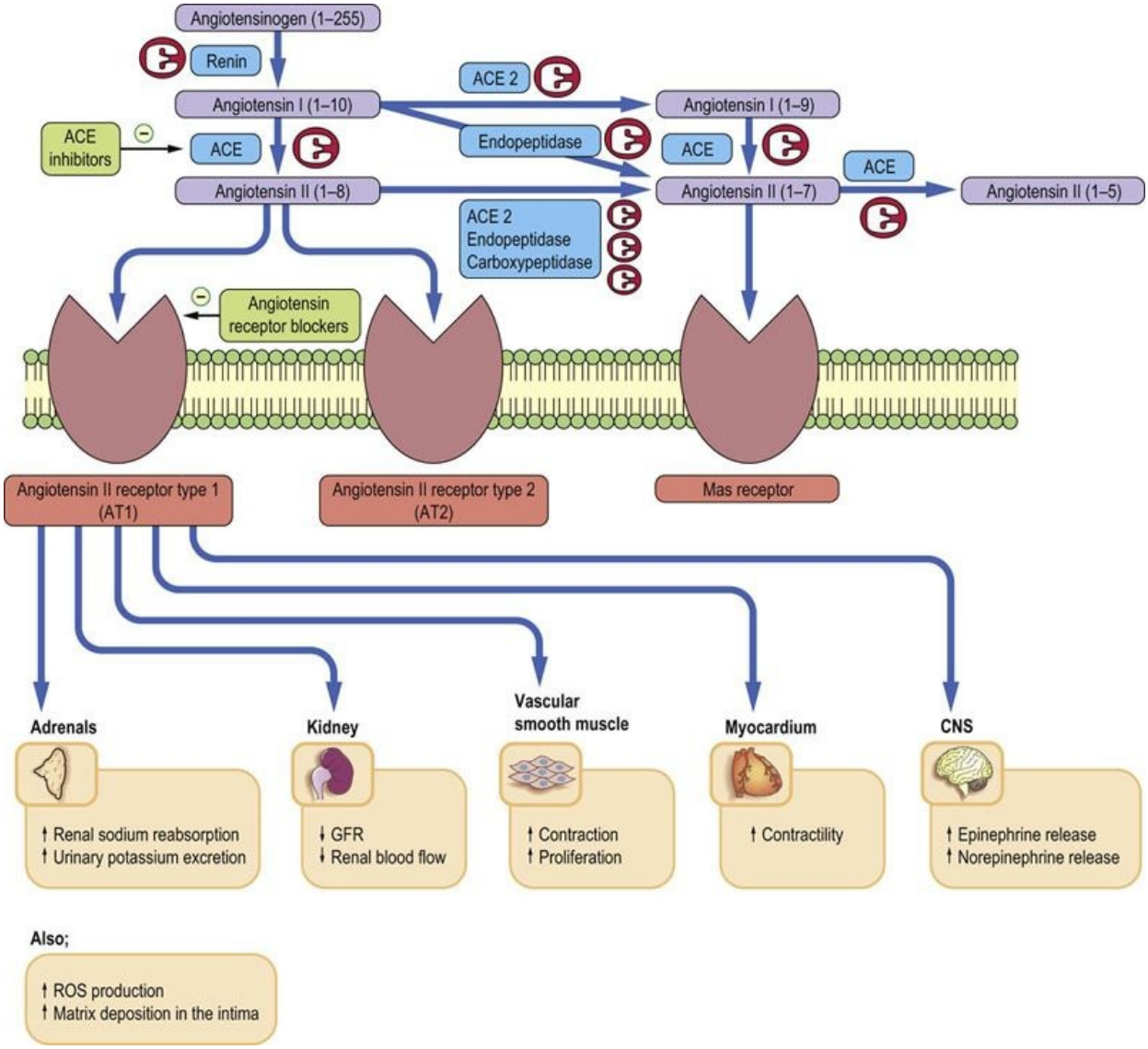


FIG. 24.10 Renin–angiotensin system.
Renin converts angiotensinogen into angiotensin I. Angiotensin I is further converted into angiotensin II by the angiotensin-converting enzyme (ACE). It also yields other angiotensin peptides. Cellular actions of angiotensins are mediated by angiotensin receptors type 1 (AT1), type 2 (AT2) and MAS receptors that bind angiotensin (1–7). The renin–angiotensin system is a target for two major classes of hypotensive drugs: ACE blockers (e.g. ramipril, enalapril) and AT1 receptor antagonists (e.g. losartan). ACE blockers are also extensively used in the treatment of heart failure. VSMC, vascular smooth muscle cells; CNS, central nervous system; ROS, reactive oxygen species; *AT1 receptor blocked by, e.g., losartan;* *AT2 receptor blocked by saralasin.

Angiotensin receptors are important in the pathogenesis of cardiovascular disease

Angiotensin II constricts vascular smooth muscle, thereby increasing blood pressure and reducing renal blood flow and glomerular filtration rate. It also promotes aldosterone release and vascular smooth muscle proliferation through the activation of AT1 receptors which signal through G-proteins and phospholipase C (Fig. 24.11). Generally, AT1 receptor activation has effects that promote cardiovascular disease: the stimulation of inflammatory phenomena, extracellular matrix deposition, and generation of reactive oxygen species (ROS). It is also pro-thrombotic. These actions are counteracted by the stimulation of the AT2 receptors, which causes vasodilatation through stimulation of NO production, promotes sodium loss and inhibits vascular smooth muscle cell proliferation. The actions of angiotensin (1–7), which acts through the so-called MAS receptor (it may also bind to AT1 and AT2), also seem to be cardioprotective. Drugs that inhibit ACE are now extensively used in the treatment of hypertension and heart failure (see Fig. 24.11 and Box on p. 61).

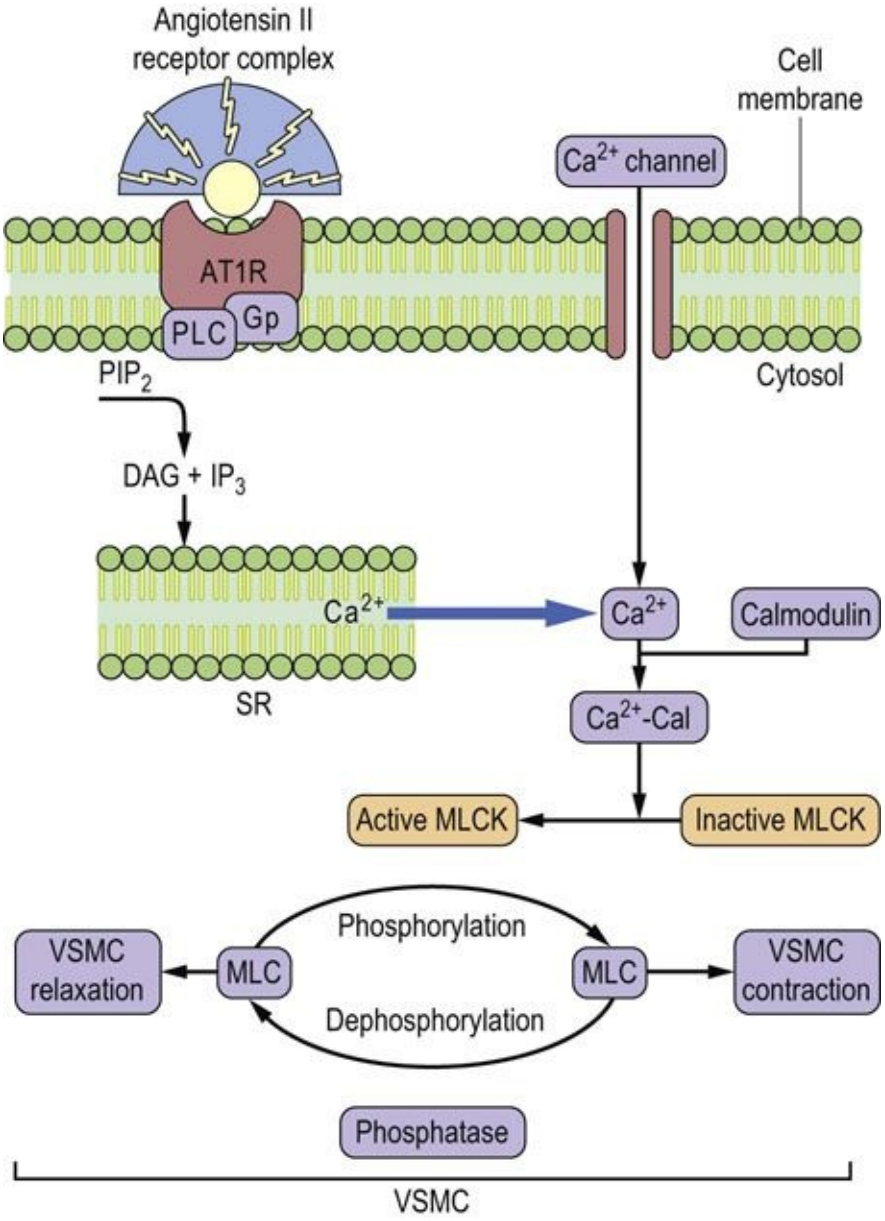


FIG. 24.11 Angiotensin II-induced vasoconstriction.
The angiotensin type 1 receptor (AT1R) is coupled to G-proteins. The binding of angiotensin II leads to phospholipase C-mediated formation of inositol 1,4,5-trisphosphate (IP_3) and 1,2-diacylglycerol. This mobilizes Ca^{2+} from the sarcoplasmic reticulum (SR), and causes entry of the extracellular Ca^{2+} through activated calcium channels. An increase in the cytosolic Ca^{2+} initiates contractile response of the vascular smooth muscle cells. Ca^{2+} subsequently binds to calmodulin (Cal). The Ca^{2+}–calmodulin complex activates myosin light-chain kinase. The kinase phosphorylates myosin light chains and elicits muscular tension. This effect is terminated by the dephosphorylation of myosin (Chapter 20). AT1 receptor antagonists, such as losartan, inhibit vasoconstrictor effects of angiotensin II and are used in the treatment of hypertension. AT1R, angiotensin type 1 receptor; DAG, 1,2,diacylglycerol; VSMC, vascular smooth muscle

cell; MLCK, myosin light-chain kinase; Cal, calmodulin.

Clinical box Renin–angiotensin-aldosterone system and cardiac failure

A 65-year-old man with a previous anterior myocardial infarction presented with increasing fatigue, shortness of breath and ankle edema. Physical examination showed mild tachycardia, and a raised jugular venous pressure. An echocardiogram showed that the function of the left ventricle during systole was poor. The patient's serum measurements revealed sodium 140 mmol/L, potassium 3.5 mmol/L, protein 34 (normal 35–45) g/dL, creatinine 80 µmol/L (0.90 mg/dL), and urea 7.5 mmol/L (45 mg/dL).

Comment.

This man presents with symptoms and signs of cardiac failure. The impaired function of the heart leads to a decreased blood flow through the kidney, activation of the renin–angiotensin system and stimulation of aldosterone secretion. Aldosterone causes an increased renal reabsorption of sodium and water retention, thereby increasing extracellular fluid volume and edema.

Clinical box Arterial hypertension is a common disease

Hypertension is inappropriately increased arterial blood pressure. The desirable level of systolic blood pressure is below 140 mmHg and diastolic pressure 90 mmHg (optimal values are still lower, below 120/80 mmHg). According to the World Health Organization, up to 20% of the population of the developed world may suffer from the condition. Arterial hypertension has been classified as ‘essential’ (primary) or ‘secondary’. A cause of essential hypertension has not yet been identified, although it is

known to involve multiple genetic and environmental factors including neural, endocrine, and metabolic components. A sodium-rich diet is a recognized factor in the development of hypertension.

Hypertension is associated with an increased risk of stroke and myocardial infarction and causes one in every eight deaths worldwide. A range of drugs is used in the modern treatment of hypertension. These include diuretics such as bendrofluazide, drugs blocking adrenoreceptors, inhibitors of the angiotensin-converting enzyme and the antagonists of angiotensin receptors type 1 (see Box on p. 61 and Box on p. 315; pheochromocytoma is described on p. 558)

Aldosterone

Aldosterone regulates sodium and potassium homeostasis

Aldosterone is a major mineralocorticosteroid hormone in man (Chapter 17), and is produced in the adrenal cortex. It regulates extracellular volume and vascular tone, and controls renal sodium and potassium transport. It binds to the cytosolic mineralocorticoid receptor in the epithelial cells, principally in the renal collecting ducts. The receptor moves to the nucleus and binds to specific domains on targeted genes, altering their expression.

Aldosterone regulates the Na^+/K^+-ATPase in both the long and short term, and also regulates transporters such as the Na^+/H^+ exchanger type 3 in the proximal tubule, the Na^+/Cl^- co-transporter in the distal tubule, and the epithelial sodium channel in the renal collecting duct. The overall result is an increased sodium reabsorption, and increased potassium and hydrogen ion secretion.

Hyperaldosteronism is a common finding in hypertension

Primary hyperaldosteronism occurs as a result of abnormal adrenal activity and is rare. It may be a result of a single adrenal tumor, an adenoma **(Conn's syndrome)**. The more common secondary hyperaldosteronism is due to an increased secretion of renin. **Pheochromocytomas** are catecholamine-secreting tumors that cause hypertension in about 0.1% of hypertensive patients. It is

important to correctly diagnose pheochromocytoma, because it can be surgically removed (Chapter 43 and Box on p. 558).

Natriuretic peptides

Natriuretic peptides are important markers of heart failure

A family of peptides known as the natriuretic peptides are involved in the regulation of fluid volume. The two main ones are the **atrial natriuretic peptide** (ANP) and the **brain natriuretic peptide** (BNP). ANP is synthesized predominantly in the cardiac atria as a 126-amino acid propeptide (pro-ANP). It is then cleaved into a smaller 98-amino acid *N*-terminal peptide, and the biologically active 28-amino acid ANP. BNP is synthesized in the cardiac ventricles as a 108-amino acid propeptide, and is cleaved into a 76-amino acid *N*-terminal peptide and a biologically active 32-amino acid BNP. BNP 32 and another peptide, CNP (23 amino acids long), were isolated from the porcine brain, thus the name. All natriuretic peptides possess a ring-type structure due to the presence of a disulfide bond.

Natriuretic peptides promote sodium excretion and decrease the blood pressure. ANP and BNP are secreted in response to atrial stretch and to ventricular volume overload. They bind to G-protein-linked receptors: the A-type receptors are located predominantly in the endothelial cells and the B-type receptors in the brain. There is cross-reactivity between different natriuretic receptors with regard to these peptides.

The signaling pathway includes the membrane guanyl cyclase and the soluble guanyl cyclase, the latter being stimulated by NO. The generated cGMP acts on protein kinase C and phosphodiesterase 2 and 3, thus regulating cAMP synthesis.

Clinical test box Diagnostic use of the brain natriuretic peptide (BNP) propeptides

Instead of measuring the active forms of natriuretic peptides in plasma, it is more convenient to measure the propeptides which are present in plasma in equimolar amounts to the active species. Thus, proBNP (1–76) reaches higher levels in cardiac failure than BNP

32. Similarly, proANP (1–98) has a longer half-life in plasma than biologically active 1–28 ANP and therefore is present in the circulation in higher concentrations.

Clinical test box ANP and BNP as markers of heart failure

Importantly, the levels of ANP and BNP are increased in heart failure and therefore their measurements are used as early biochemical markers of this condition. These measurements are particularly useful for excluding heart failure in patients who present with nonspecific symptoms such as shortness of breath.

Vasopressin and aquaporins

Reabsorption of water in the collecting ducts of the kidney is controlled by the posterior pituitary hormone vasopressin through its control of membrane water channels, the aquaporins.

Vasopressin determines the final volume and concentration of the urine

Vasopressin (also known as the **antidiuretic hormone**, ADH) controls water reabsorption in the collecting ducts of the kidney. Vasopressin is synthesized in the supraoptic and paraventricular nuclei of the hypothalamus and is transported along axons to the posterior pituitary. It is stored there before being further processed and released. It binds to a receptor located on the membranes of tubular cells in the collecting ducts (Fig. 24.12). The receptor is coupled to G-proteins and activates PKA. The PKA phosphorylates aquaporin 2 (AQP2). This stimulates AQP2's translocation to the cell membrane, increasing water reabsorption in the collecting duct. On the other hand, Vasopressin secretion needs to be suppressed to allow urine dilution. Failure to maximally suppress

vasopressin results in the inability to dilute urine below the osmolality of plasma.

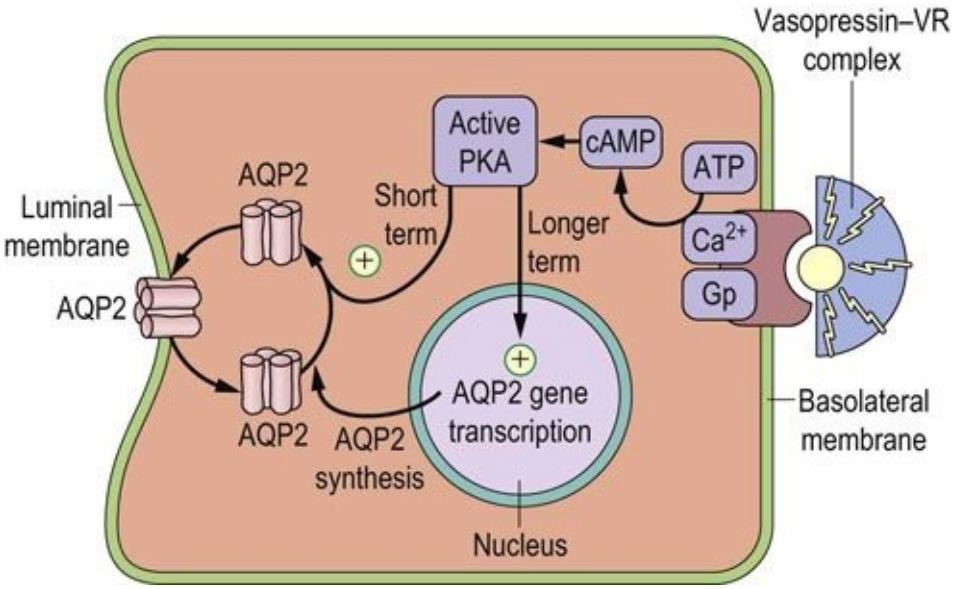


FIG. 24.12 Vasopressin regulates water reabsorption in the collecting duct. Vasopressin controls the aquaporin 2 (AQP2) water channel. Vasopressin binds to its receptor (VR) and, through G-proteins, (Gp) stimulates production of cAMP, which, in turn, activates protein kinase A (PKA). PKA phosphorylates cytoplasmic AQP2 and induces its translocation to the cell membrane, increasing capacity for water transport. Vasopressin also regulates expression of the AQP2 gene.

Glucocorticoids stimulate vasopressin secretion primarily through their hemodynamic effects, which decrease arterial pressure. Incidentally, vasopressin is also stimulated by nicotine.

Aquaporins are membrane channel proteins which transport water

The aquaporin water channel is illustrated in Figure 24.13. AQP 2 and 3 are present in the collecting duct and are regulated by vasopressin. AQP1 is expressed on the apical and basolateral membranes of the proximal tubules and in the descending loop of Henle, and is not under vasopressin control. It is also present in erythrocytes, renal proximal tubular cells, and in the capillary endothelium.

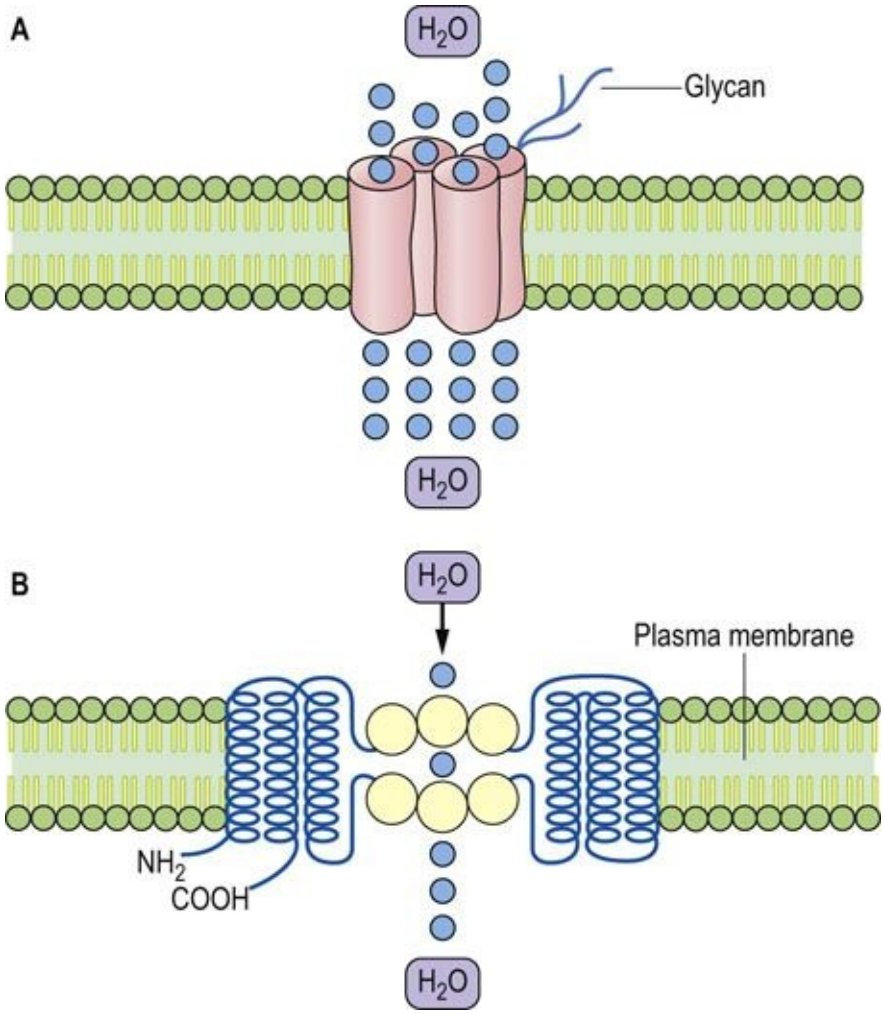


FIG. 24.13 Aquaporin water channel.
(A) Aquaporin 1 is a multisubunit water channel with a glycan unit attached to one of the subunits. **(B)** Each of the two monomers has two tandem repeat structures, each consisting of three membrane-spanning regions and connecting loops embedded in the membrane.

Defects in vasopressin secretion and mutations of genes coding for aquaporins cause clinical conditions

Vasopressin deficiency causes the condition known as **diabetes insipidus**, in which large amounts of dilute urine are excreted. On the other hand, an excessive secretion of vasopressin may occur following major trauma or surgery. This is known as the **syndrome of inappropriate antidiuretic hormone secretion (SIADH)**, and leads to water retention. Mutations in the vasopressin receptor gene and also in the AQP2 gene lead to different types of **nephrogenic**

diabetes insipidus, a condition associated with passing large amounts of urine and with dehydration.

Integration of water and sodium homeostasis

Aldosterone and vasopressin together control the handling of sodium and water

Normally, despite variations in fluid intake, plasma osmolality is maintained within narrow limits (280–295 mmol/kg H_2O). Vasopressin contributes to the control of plasma osmolality by regulating water metabolism. It responds to both osmotic and volume signals: on the one hand, its secretion, and thirst, are stimulated by signals from osmoreceptors which respond to very small (approximately 1%) increases in plasma osmolality; on the other hand, vasopressin release is stimulated by a decrease (more than 10%) in the circulating volume.

Water excess increases plasma volume, renal blood flow, and GFR

When there is water excess, the production of renin is suppressed. Low concentration of aldosterone allows the urinary sodium loss. Because excess water ‘dilutes’ the plasma, plasma osmolality decreases. This decrease in osmolality, sensed by the hypothalamic osmoreceptors, suppresses both thirst and the secretion of vasopressin. Suppression of vasopressin leads to the urinary loss of water. Thus, the overall response to water excess is increased loss of sodium and water in urine.

Water deficit (dehydration) decreases plasma volume, renal blood flow and GFR

When there is water deficit (dehydration), the decrease in renal blood flow stimulates the renin–angiotensin–aldosterone system. Aldosterone inhibits urinary sodium excretion. In parallel, water deficit causes an increase in the plasma osmolality. This stimulates vasopressin secretion, with a consequent decrease in the urine volume. Thus, the overall response to water deficit is sodium and water retention (Fig. 24.14).

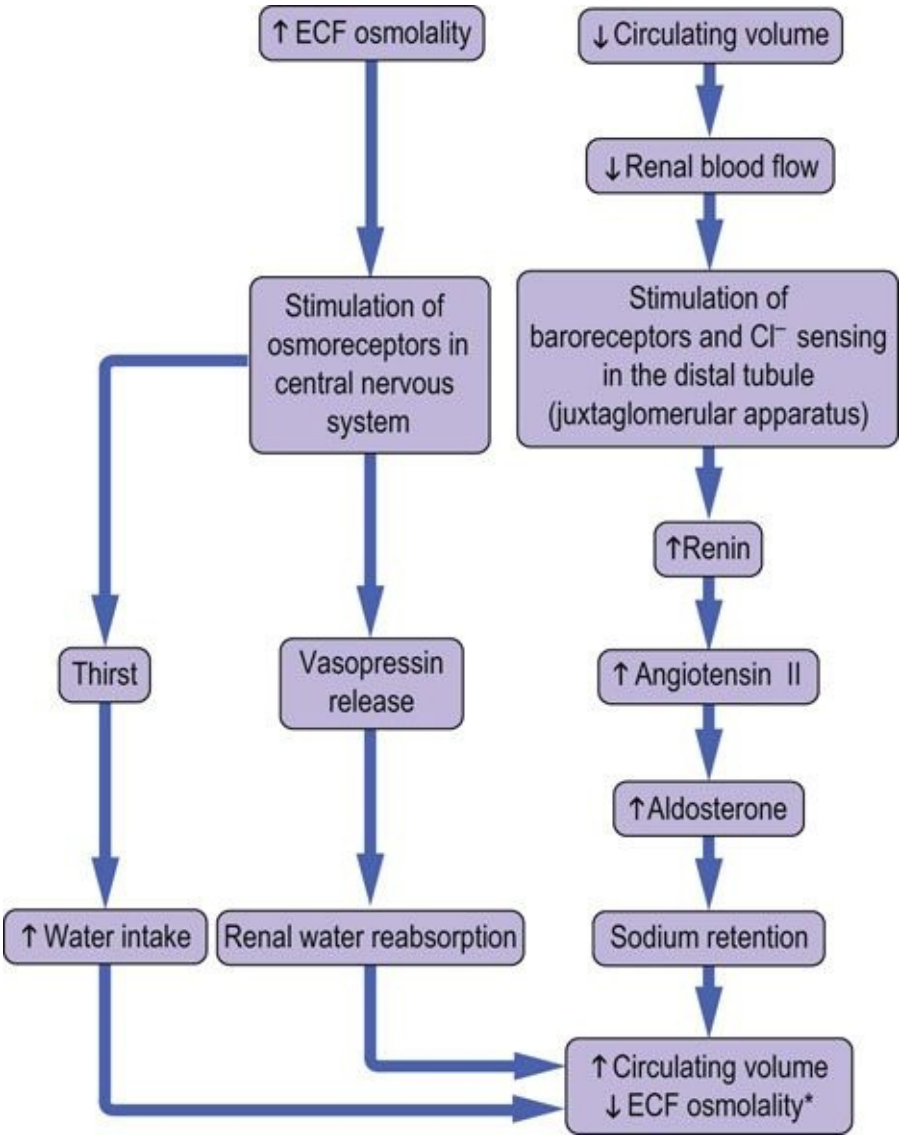


FIG. 24.14 Links between water and sodium metabolism.
Water and sodium metabolism are closely interrelated. **An increase in ECF osmolality** stimulates secretion of vasopressin and leads to increased renal water reabsorption. This 'dilutes' the ECF and the osmolality decreases. This response is reinforced by the stimulation of thirst. A **decrease in the plasma volume** also stimulates water retention through stimulation of the pressure-sensitive receptors (baroreceptors) in the juxtaglomerular apparatus. *Osmolality will decrease if the degree of water retention is relatively greater than that of sodium retention.

Clinical box Poor fluid intake leads to dehydration

An 80-year-old man had been admitted to hospital after lying for a

prolonged period on the floor at home after suffering acute stroke. He had poor tissue turgor, dry mouth, tachycardia and he was hypotensive. Serum measurements revealed: sodium 150 mmol/L, potassium 5.2 mmol/L, bicarbonate 35 mmol/L, urea 19 mmol/L (90.3 mg/dL) and creatinine 110 µmol/L (1.13 mg/dL).

Reference values are:

Sodium: 135–145 mmol/L
Potassium: 3.5–5.0 mmol/L
Bicarbonate: 20–25 mmol/L
Urea: 2.5–6.5 mmol/L (16.2–39 mg/dL)
Creatinine: 20–80 µmol/L (0.28–0.90 mg/dL)

Comment.

This patient presents with dehydration, indicated by the high sodium and urea values, and mildly elevated creatinine. He was treated with intravenous fluids, predominantly in the form of 5% dextrose, to replace the water deficit.

Serum sodium concentration is a marker of fluid and electrolyte disorders

Water and electrolyte disturbances result from an imbalance between the intake of fluids and electrolytes and their loss, and from the movement of water and electrolytes between body compartments. A decreased sodium concentration **(hyponatremia)** usually indicates that the extracellular fluid is being ‘diluted’ (due to an excess of water), whereas an increased sodium concentration (hypernatremia) means that the extracellular fluid is being ‘concentrated’ (due to water loss). Hyponatremia may also result from loss of sodium but this is rare.

Assessment of water and electrolyte status is an important part of clinical practice

The assessment of water and electrolyte balance is an important part of clinical examination. In addition to the physical examination and medical history, the following measurements are required:

- **Serum electrolyte concentrations:** the profile commonly requested by a

physician includes sodium, potassium, chloride and bicarbonate concentrations.

- **Serum urea (blood urea nitrogen) and creatinine.**
- **Urine volume, osmolality and sodium concentration.**
- **Serum osmolality.**
- **Fluid balance chart:** patients who have or who are at risk of developing abnormalities of water or electrolyte balance need a daily record of fluid intake and loss.

Summary

■ Both deficit of body water (dehydration) and its excess (overhydration) cause potentially serious clinical problems. Therefore, the assessment of water and electrolyte balance is an important part of clinical examination.
■ Body water balance is closely linked to the balance of dissolved ions (electrolytes), the most important of which are sodium and potassium.
■ Movement of water between ECF and ICF is controlled by osmotic gradients.
■ Movement of water between the lumen of a blood vessel and the interstitial fluid is controlled by the osmotic and hydrostatic pressures.
■ The main regulators of water and electrolyte balance are vasopressin (water) and aldosterone (sodium and potassium).
■ The renin–angiotensin–aldosterone system is the principal regulator of the blood pressure and vascular tone.
■ Measurements of natriuretic peptides help to diagnose the cardiac failure.

Active learning

1. Comment on the role of Na^+/K^+-ATPase in maintaining the ion gradients across cell membrane.
2. Explain the role of the renin–angiotensin system in the maintenance of blood pressure.
3. Describe water movements between ECF and ICF which take place in water deprivation.
4. Why does the low concentration of albumin in plasma lead to edema?

Further reading

Androgue, HJ, Madias, NE. Mechanisms of disease: sodium and potassium in the pathogenesis of hypertension. *N Engl J Med*. 2007; 356:1966–1978.

Bekheirnia, R, Schrier, RW. Pathophysiology of water and sodium retention: edematous states with normal kidney function. *Curr Opin Pharmacol*. 2006; 6:202–207.

Chobanian, A, Bakris, GL, Black, HR, *et al.* The Seventh Report of the Joint National Committee on Prevention, Detection, Evaluation and Treatment of High Blood Pressure. *JAMA*. 2003; 289:2560–2572.

Ellison, DH, Berl, T. The syndrome of inappropriate antidiuresis. *N Engl J Med*. 2007; 356:2064–2072.

Ernst, ME, Moser, M. Use of diuretics in patients with hypertension. *N Engl J Med*. 2009; 361:2153–2164.

Goetze, JP. B-type natriuretic peptide: from posttranslational processing to clinical measurement. *Clin Chem*. 2012; 58:83–91.

Richards, AM, Troughton, RW. Use of natriuretic peptides to guide and monitor heart failure therapy. *Clin Chem*. 2012; 58:62–71.

Schrier, RW. Body water homeostasis: clinical disorders of urinary dilution and concentration. *J Am Soc Nephrol*. 2006; 17:1820–1832.

Verkman, AS. Aquaporins in clinical medicine. *Ann Rev Med*. 2012; 63:303–316.

Websites and downloads

Biesalski, HK, Bischoff, SC, Boehles, HJ, et al, Water, electrolytes, vitamins and trace elements – Guidelines on Parenteral Nutrition. *Ger Med Sci Ger Med Sci* 2009. www.ncbi.nlm.nih.gov/pmc/articles/PMC2795367

Medline Plus. Water and electrolyte balance. www.nlm.nih.gov/medlineplus/fluidandelectrolytebalance.html.

Powell-Tuck, J, Peter Gosling, P, Dileep, N, *et al*. British Consensus Guidelines on Intravenous Fluid Therapy for Adult Surgical Patients. www.bapen.org.uk/pdfs/bapen_pubs/giftasup.pdf, 2011.

CHAPTER 25

Regulation of Hydrogen Ion Concentration (Acid–Base Balance)

Marek H. Dominiczak and Mirosława Szczepańska-Konkel

Learning objectives

After reading this chapter you should be able to:

- Explain the nature of the bicarbonate buffer.
- Describe the gas exchange in the lungs.
- Describe the respiratory and metabolic components of the acid–base balance.
- Define and classify acidosis and alkalosis.
- Comment on clinical conditions associated with disturbances of the acid–base balance.

Introduction

Acids are produced in the course of metabolism

Metabolism generates carbon dioxide within cells. Carbon dioxide dissolves in water, forming carbonic acid, which in turn dissociates releasing hydrogen ion. The acids derived from sources other than CO_2 are known as nonvolatile; by definition, they cannot be removed through the lungs, and must be excreted via the kidney. The net production of nonvolatile acids is in the order of 50 mmol/24 h.

Lactic acid is produced during anaerobic glycolysis and its concentration in plasma is the hallmark of hypoxia. Ketoacids (acetoacetic and β-hydroxybutyric acid) are important in diabetes (Chapter 21). The metabolism of sulfur-containing amino acids and phosphorus-containing compounds also generates inorganic acids.

In spite of the amount of hydrogen ion produced, its blood concentration (or its negative logarithm, the pH), is remarkably constant: it remains between 35 and 45 nmol/L (pH 7.35–7.45). Maintenance of stable pH is essential because it affects the ionization of proteins (Chapter 2) and, consequently, the activity of many enzymes and other biologically active molecules such as ion channels. Changes in pH together with the partial pressure of carbon dioxide (pCO_2) affect the shape of the hemoglobin saturation curve, and thus tissue oxygenation (Chapter 5). Also, a decrease in pH increases sympathetic tone and may lead to cardiac dysrhythmias.

Maintaining the acid–base balance involves lungs, erythrocytes and kidneys

The acid–base balance involves the lungs, the erythrocytes, and the kidneys (Fig. 25.1). The lungs control the exchange of carbon dioxide and oxygen between the blood and the atmosphere; the erythrocytes transport gases between lungs and tissues; and the kidneys control plasma bicarbonate synthesis and the excretion of the hydrogen ion.

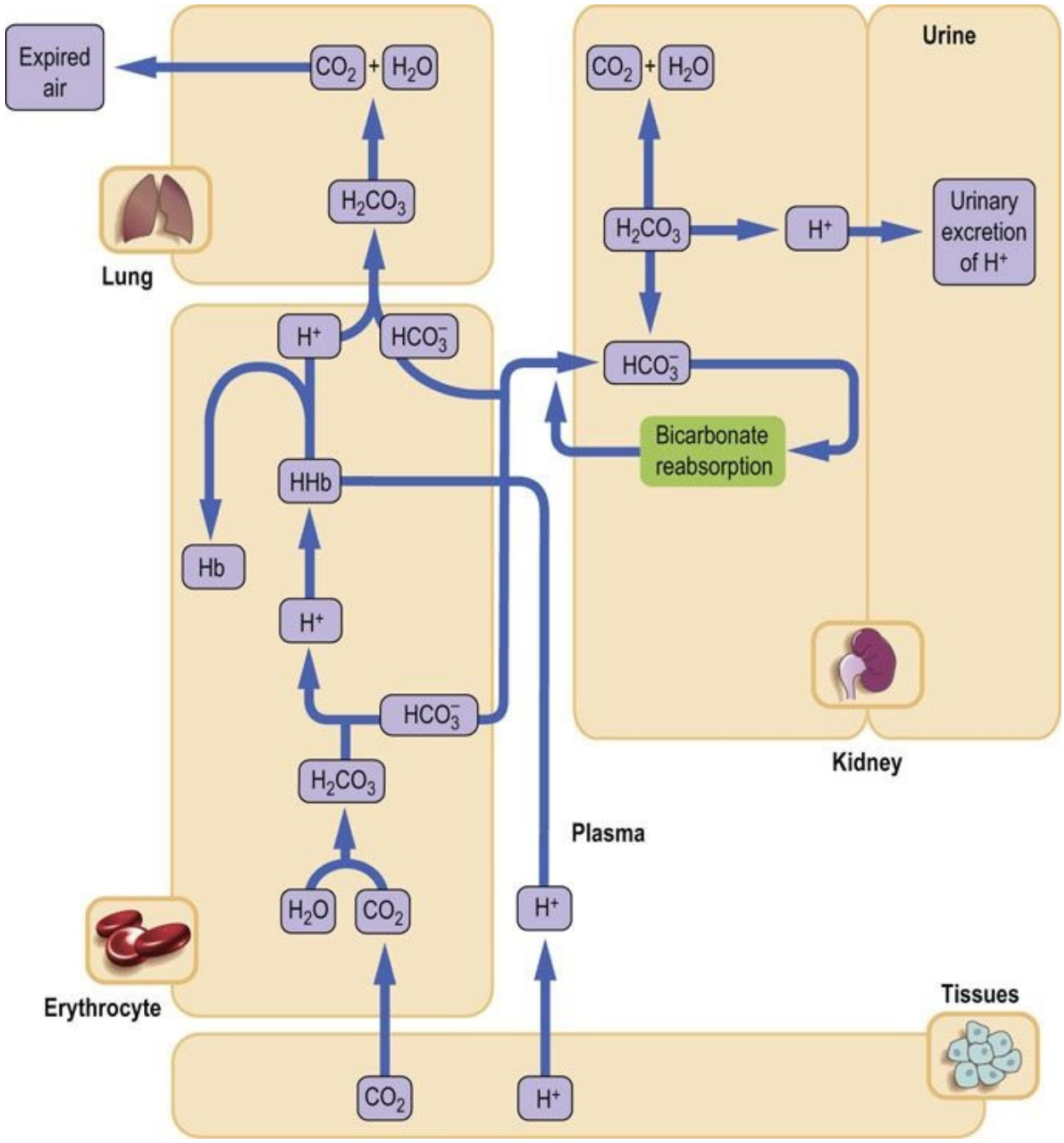


FIG. 25.1 Acid–base balance.
Lungs, kidneys, and erythrocytes contribute to the maintenance of the acid–base balance. The lungs control the gas exchange with the atmospheric air. Carbon dioxide generated in tissues is transported in plasma as bicarbonate; the erythrocyte hemoglobin contributes to CO_2 transport. Hemoglobin also buffers the hydrogen ion derived from carbonic acid. The kidneys reabsorb filtered bicarbonate in the proximal tubules and generate new bicarbonate in the distal tubules, where there is a net secretion of hydrogen ion. Hb, hemoglobin.

Clinical relevance

Clinically, understanding of the acid–base balance is important in many sub-specialties of medicine and surgery, and particularly in anesthesiology and critical care medicine.

Body Buffer systems: respiratory and metabolic components of the acid–base balance

Blood and tissues contain buffer systems that minimize changes in hydrogen ion concentration

The main buffer that neutralizes hydrogen ions released from cells is the **bicarbonate buffer**. Another important buffer is **hemoglobin**, which contributes to buffering of hydrogen ion generated from the carbonic anhydrase reaction. Within cells, the hydrogen ion is neutralized by intracellular buffers, mainly **proteins** and **phosphates** (Table 25.1 and Chapter 2).

Table 25.1

The main buffers in the human body

Buffer	Acid	Conjugate base	Site of main buffering action
Hemoglobin	HHb	Hb^-	Erythrocytes
Proteins	HProt	$Prot^-$	Intracellular fluid
Phosphate buffer	$H_2PO_4^-$	HPO_4^{2-}	Intracellular fluid
Bicarbonate	$CO_2 \rightarrow H_2CO_3$	HCO_3^-	Extracellular fluid

See Chapter 2 for the principles of buffering action. The Brønsted–Lowry definition of an acid is ‘a molecular species that has a tendency to lose a hydrogen ion, forming a conjugate base’.

Bicarbonate buffer is an open system which remains at equilibrium with atmospheric air

Buffering capacity of the bicarbonate buffer exceeds all the ‘closed’ buffer systems. The metabolically produced CO_2 diffuses through cell membranes and

dissolves in plasma. The plasma solubility coefficient of CO_2 is 0.23 if pCO_2 is measured in kPa (0.03 if pCO_2 is measured in mmHg; 1 kPa = 7.5 mmHg or 1 mmHg = 0.133 kPa)). Thus, at the normal pCO_2 of 5.3 kPa (40 mmHg), the concentration of dissolved CO_2 (dCO_2) is:

$$dCO_2\,(\text{mmol / L}) = 5.3\,\text{kPa} \times 0.23 = 1.2\,\text{mmol / L}$$

CO_2 equilibrates with H_2CO_3 in plasma in the course of a slow, nonenzymatic reaction. Normally plasma H_2CO_3 concentration is very low, about 0.0017 mmol/L. However, because of the equilibrium between H_2CO_3 and dissolved CO_2 (theoretically all dissolved CO_2 could eventually convert into H_2CO_3), this component of the bicarbonate buffer can be taken as equal to the sum of the H_2CO_3 and the dissolved CO_2. The key equation describing the behavior of the bicarbonate buffer is the **Henderson–Hasselbalch equation** (Chapter 2). It expresses the relationship between pH and the components of the buffer:

$$\text{pH} = \text{p}K + \log\left([\text{bicarbonate}] \,/\, \text{pCO}_2 \times 0.23\right)$$

The equation demonstrates that blood pH is determined by the ratio between the concentration of plasma bicarbonate (the **'base' component** of the buffer) and the concentration of dissolved CO_2 (the **'acid' component**). Normally, at pCO_2 of 5.3 kPa and dCO_2 concentration 1.2 mmol/L (see above) the plasma bicarbonate concentration is about 24 mmol/L. The p*K* of the bicarbonate buffer is 6.1. Let's insert the actual concentrations of buffer components into the preceding equation:

$$\text{pH} = 6.1 + \log(24 \,/\, 1.2) = 7.40$$

Thus the normal concentration of bicarbonate and normal partial pressure of CO_2 correspond to pH 7.40 (hydrogen ion concentration 40 nmol/L). The bicarbonate buffer minimizes changes in hydrogen ion concentration when acid

is added to blood.

When the H^+ concentration in the system increases, the bicarbonate component of the buffer accepts (H^+), forming carbonic acid, which is subsequently converted into CO_2 and H_2O in the reaction catalyzed by **carbonic anhydrase**:

$$H^+ + HCO_3^- \rightleftharpoons H_2CO_3 \rightleftharpoons CO_2 + H_2O$$

The CO_2 is eliminated through the lungs. The excess hydrogen ion has been neutralized and the [bicarbonate]/pCO_2 ratio brought back towards normal.

On the other hand, when the H^+ concentration decreases, the carbonic acid component of the buffer will dissociate to supply H^+ for a reaction that will yield water and bicarbonate ion (at this stage the increase in plasma bicarbonate is minimal):

$$H_2CO_3 \rightarrow H^+ + HCO_3^-$$

Subsequently, the ventilation rate will decrease, retaining CO_2 and attempting to normalize the [bicarbonate]/pCO_2 ratio:

$$CO_2 + H_2O \rightarrow H_2CO_3$$

Thus the denominator in the Henderson–Hasselbalch equation (pCO_2) is controlled by the lungs. For this reason it is called **'the respiratory component of the acid–base balance'**. On the other hand, plasma bicarbonate concentration is controlled by the kidneys and erythrocytes and, consequently, it is called **'the metabolic component of the acid–base balance'** (Fig. 25.2).

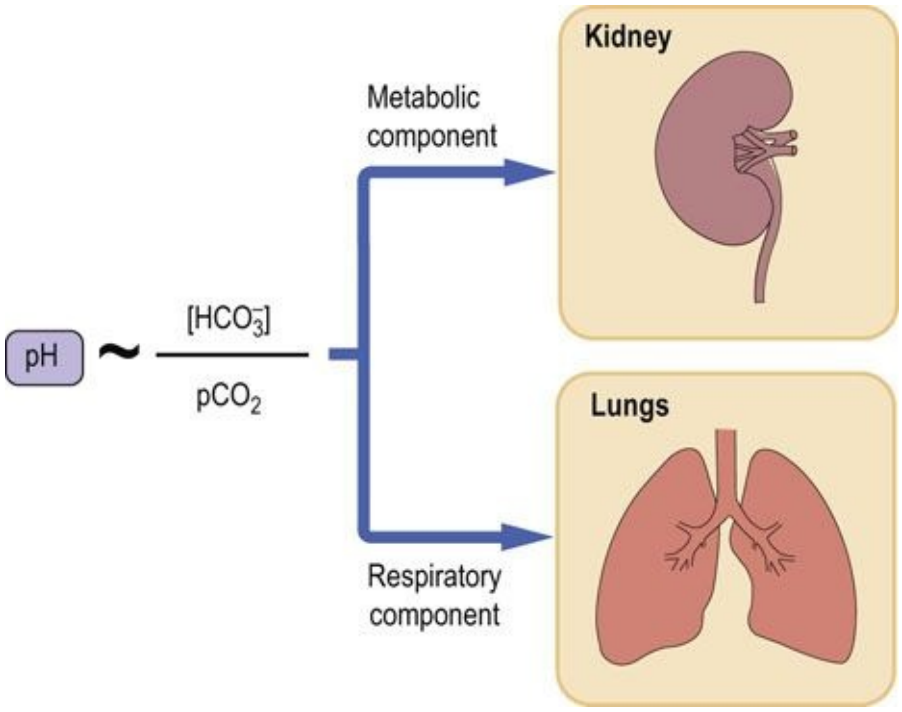


FIG. 25.2 The bicarbonate buffer.
Blood pH is proportional to the ratio of plasma bicarbonate to the partial pressure of carbon dioxide (pCO_2). The components of the bicarbonate buffer are thus the carbon dioxide and the bicarbonate. The pCO_2 is the respiratory component of acid–base balance, and bicarbonate is the metabolic component.

Carbonic anhydrase converts the dissolved CO_2 into carbonic acid

Erythrocytes and renal tubular cells contain a zinc-containing enzyme, carbonic anhydrase (CA), which converts dissolved CO_2 into carbonic acid. Carbonic acid dissociates, yielding hydrogen and bicarbonate ions:

$$CO_2 + H_2O \overset{CA}{\rightleftharpoons} H_2CO_3 \rightleftharpoons H^+ + HCO_3^-$$

This is how renal tubular cells and erythrocytes produce bicarbonate. The kidneys regulate bicarbonate reabsorption and synthesis, and the erythrocytes adjust its concentration in response to changes in pCO_2.

Respiratory and metabolic components of the acid–base

balance are interlinked

The respiratory and metabolic components of the acid–base balance are closely inter-dependent: one tends to compensate for the untoward changes in the other. When the primary disorder is respiratory (such as, for instance, a severe chronic obstructive airway disease, COAD) and causes accumulation of CO_2, a compensatory increase in bicarbonate reabsorption by the kidney takes place. Conversely, the decrease in pCO_2 (as happens during hyperventilation in an asthmatic attack) would cause a kidney response, leading to decreased bicarbonate reabsorption.

Conversely, when the primary problem is metabolic (for instance, diabetic ketoacidosis), a decrease in bicarbonate concentration and the resulting decrease in pH stimulate the respiratory center to increase the ventilation rate. The CO_2 is blown off and plasma pCO_2 decreases. This is why patients with metabolic acidosis hyperventilate. On the other hand, an increase in plasma bicarbonate (causing an increase in pH) leads to a decrease in the ventilation rate, and to CO_2 retention. Thus, the compensatory change always tends to normalize the [bicarbonate]/pCO_2 ratio, helping to bring the pH towards normal (Fig. 25.3).

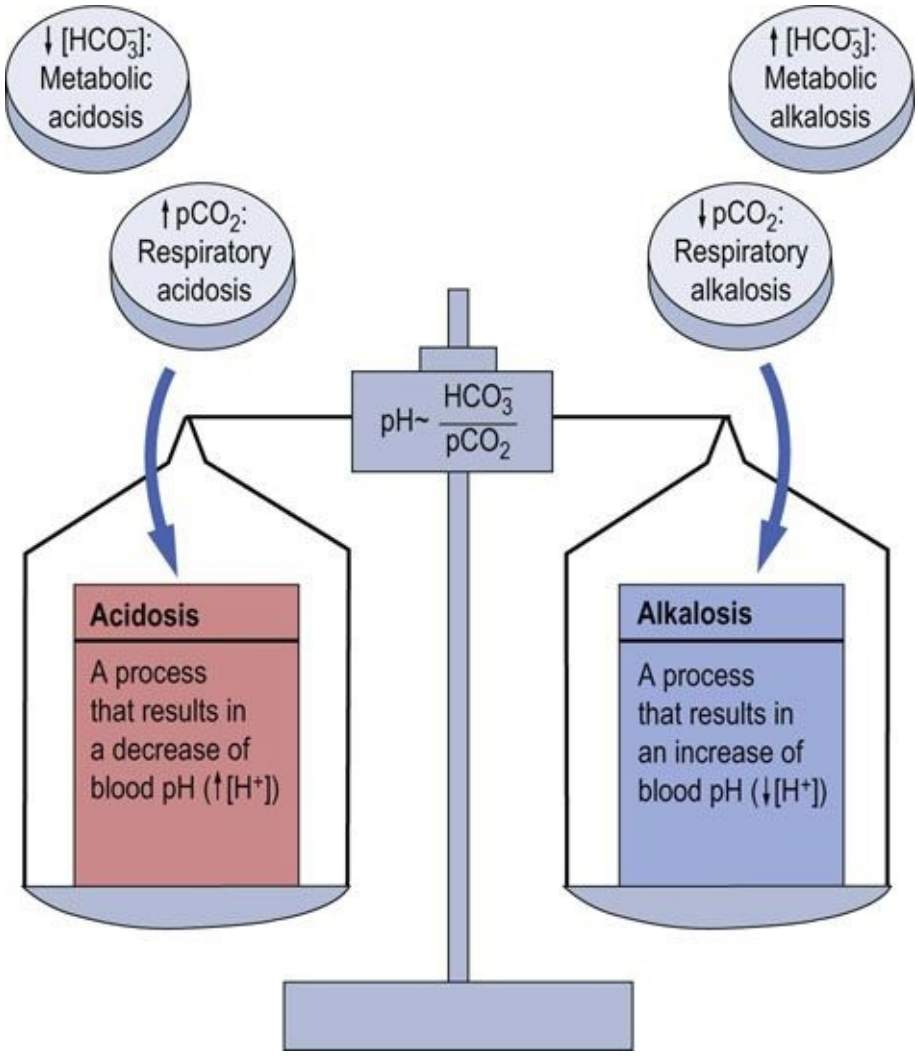


FIG. 25.3 Disorders of the acid–base balance.
A primary increase in pCO_2, or a decrease in plasma bicarbonate concentration, lead to acidosis. A decrease in pCO_2, or an increase in plasma bicarbonate, lead to alkalosis. If the primary change is in pCO_2, the disorder is called respiratory, and if the primary change is in plasma bicarbonate, it is called metabolic.

Intracellular buffering

Intracellular buffers are proteins and phosphates

The two main intracellular buffers are proteins and phosphates, and the buffering is governed by the $HPO_4^{2-}/H_2PO_4^-$ and [protein]/protein-H } ratios. Hemoglobin is an important extracellular protein buffer.

Note that when the hydrogen ion is present in plasma in excess, it enters cells in exchange for potassium and this may result in an increase in plasma

potassium concentration. Conversely, a decrease in plasma hydrogen ion, or bicarbonate excess, would be buffered by cell-derived hydrogen ion. Hydrogen ion would enter plasma in exchange for potassium, decreasing plasma potassium concentration. Thus acidemia, (low blood pH) may be associated with hyperkalemia, and alkalemia (high blood pH) is associated with hypokalemia (Fig. 25.4).

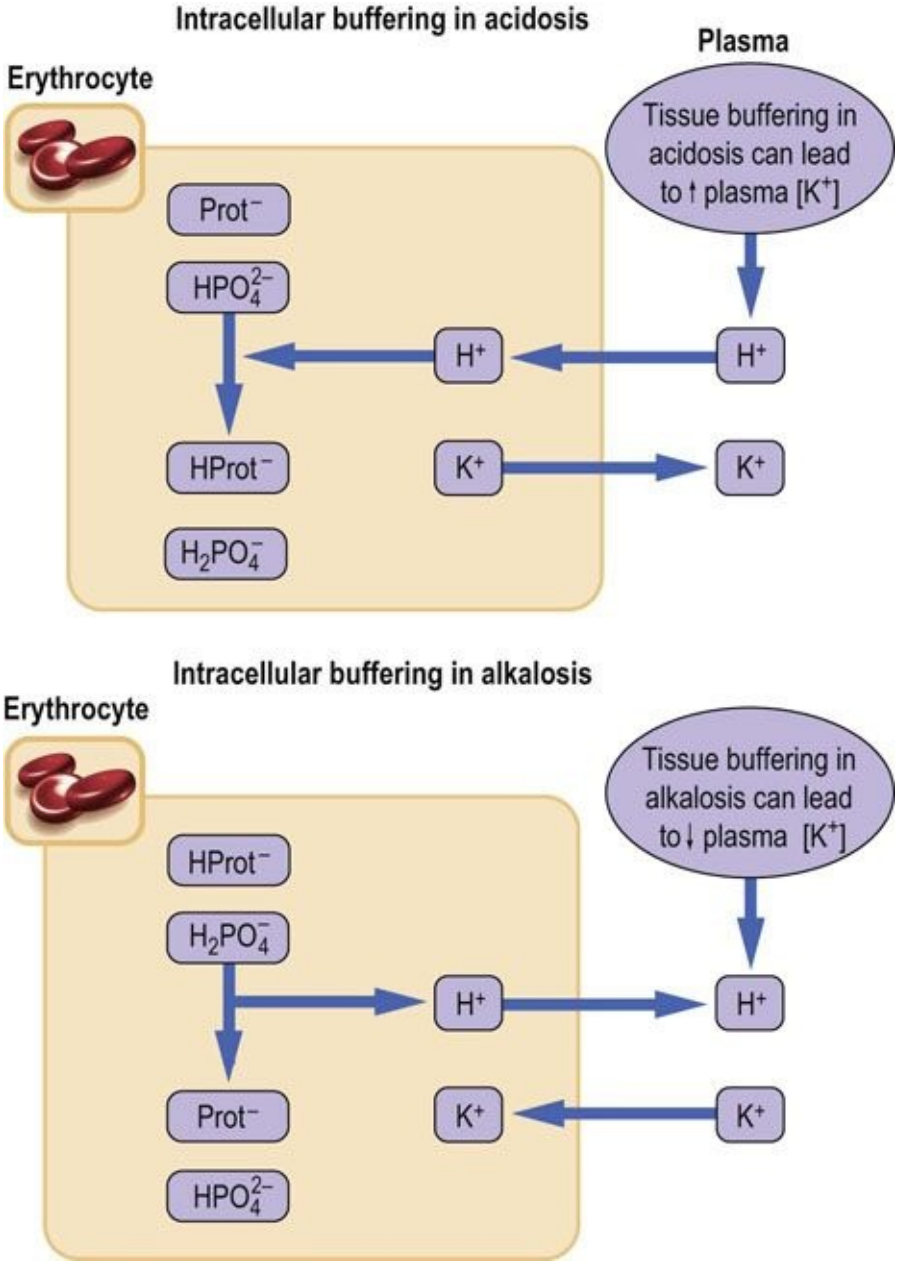


FIG. 25.4 Intracellular buffers: proteins, phosphates, and the potassium–hydrogen ion

exchange.
Intracellular buffers are primarily proteins and phosphates. However, the hydrogen ion enters cells in exchange for potassium. Therefore, an accumulation of the hydrogen ion in the plasma (acidemia) and the consequent entry of excess of hydrogen ion into cells increase plasma potassium concentration. Conversely, a deficit of hydrogen ion in plasma (alkalemia) may lead to a low plasma potassium concentration. Prot, protein.

The measurement of blood gases

The ‘blood gas measurement’ is an important first-line laboratory investigation. In respiratory failure, it is also essential to guide oxygen therapy and assisted ventilation.

The measurements are performed on a sample of arterial blood, taken usually from the radial artery in the forearm. The jargon term ‘blood gases’ means the measurements of **pO_2, pCO_2, and pH** (or hydrogen ion concentration) from which the concentration of **bicarbonate** is calculated using the Henderson–Hasselbalch equation. Several other indices are also computed: they include the total amount of buffers in the blood (the **buffer base**) and the difference between the desired (normal) amount of buffers in the blood and the actual amount (**base excess**). The reference values for pH, pCO_2, and O_2 are given in Table 25.2.

Table 25.2

Reference ranges for blood gas results

A. Reference ranges*

	Arterial	Venous
$[H^+]$	35–45 mmol/L	
pH	7.35–7.45	
pCO_2	4.6–6.0 kPa (35–45 mm Hg)	4.8–6.7 kPa (36–50 mm Hg)
pO_2	10.5–13.5 kPa (79–101 mm Hg)	4.0–6.7 kPa (30–50 mm Hg)
Bicarbonate	23–30 mmol/L	22–29 mmol/L

B. Comparison of conventional and SI units of hydrogen ion concentration

Conventional units: pH	SI units: $[H^+]$ nmol/L
6.8	160
7.1	80
7.4	40
7.7	20

*The measured values in 'blood gases' are pH, pCO_2 and pO_2; the bicarbonate concentration is calculated from pH and pCO_2 values; pH below 7.0 or above 7.7 is life threatening. (Adapted with permission from Hutchinson AS. In Dominiczak MH, editor. *Seminars in clinical biochemistry*, Glasgow, 1997, Glasgow University Press.)

Lungs: the gas exchange

The lungs supply oxygen necessary for tissue metabolism and remove the generated CO_2

Approximately 10,000 L of air pass through the lungs of an average person each day. Lungs lie in the thoracic cavity surrounded with the pleural sac, a thin ‘bag’ of tissue that lines the thoracic cage at one end, and attaches to the external surface of the lungs at the other. When the thoracic cage expands during inspiration, negative pressure created in the expanding pleural sac inflates the lung.

The airways are ‘tubes’ of progressively decreasing diameter. They consist of the trachea, large and small bronchi, and even smaller bronchioles (Fig. 25.5). At the end of the bronchioles, there are pulmonary alveoli – structures lined with endothelium and covered with a film of surfactant – the main component of which is dipalmitoylphosphatidylcholine (Chapter 26). Surfactant decreases the surface tension of the alveoli. The gas exchange takes place in the alveoli.

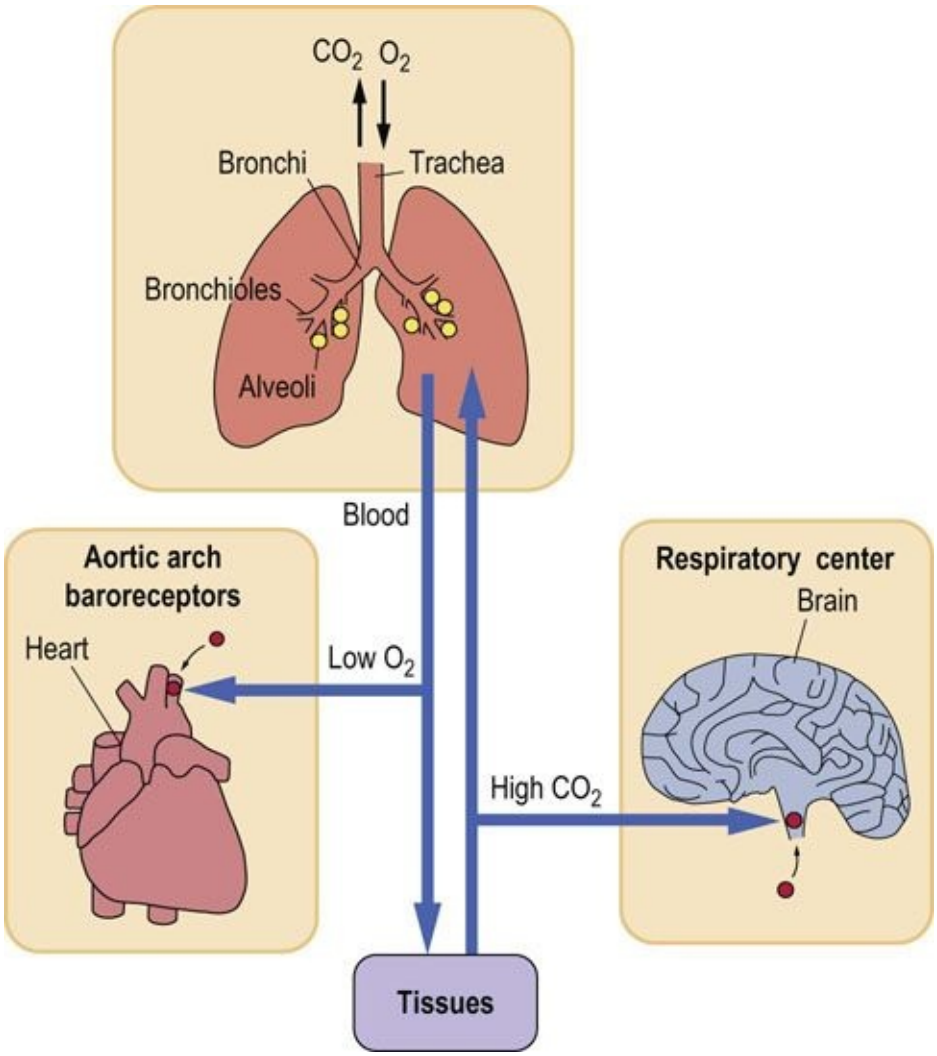


FIG. 25.5 Control of the respiratory rate by pCO_2 and pO_2.
Lung ventilation and perfusion are main factors controlling gas exchange. The pCO_2 regulates the ventilation rate through central chemoreceptors in the brainstem. When pO_2 decreases, this control switches to pO_2-sensitive peripheral receptors in the carotid bodies and in the aortic arch.

Clinical box Respiratory acidosis occurs in chronic lung disease

A 56-year-old woman was admitted to a general ward with increasing breathlessness. She had smoked 20 cigarettes a day for the previous 25 years and reported frequent attacks of 'winter bronchitis'. Blood gas measurements revealed a pO_2 of 6 kPa (45 mmHg), pCO_2 of 8.4 kPa (53 mmHg), and pH 7.35 (hydrogen

ion concentration 51 nmol/L); bicarbonate concentration was 35 mmol/L (for reference ranges, refer to Table 25.2).

Comment.

This patient presented with an exacerbation of **chronic obstructive pulmonary disease** (COPD) and a respiratory acidosis. Her pCO_2 was high, and therefore her ventilation was dependent on hypoxic drive. Her bicarbonate concentration was also increased, as a result of metabolic compensation of respiratory acidosis. One must be careful when treating such patients with high concentrations of oxygen, because the increased pO_2 may remove hypoxic drive and cause respiratory depression. Monitoring of arterial pO_2 and pCO_2 on oxygen treatment is mandatory. This patient was successfully treated with 28% oxygen.

Respiration rate is controlled by the respiratory center located in the brainstem

Both the partial pressures of oxygen (pO_2) and carbon dioxide (pCO_2) affect the ventilation rate: the respiratory center has chemoreceptors sensitive to pCO_2 and to pH. Under normal circumstances it is not the pO_2 that stimulates ventilation, but the increase in pCO_2 or the decrease in pH. However, when the pO_2 falls and hypoxia develops, it begins to control ventilation through a set of receptors located in the carotid bodies in the aortic arch. When arterial pO_2 decreases to less than 8 kPa (60 mmHg), this '**hypoxic drive'** becomes the main controller of the ventilation rate. Persons who suffer from hypoxia due to chronic lung disease depend on hypoxic drive to maintain their ventilation rate (see Clinical Box on this page).

Ventilation and lung perfusion together determine gas exchange

Blood supply to the pulmonary alveoli is provided by the pulmonary arteries that carry deoxygenated blood from the periphery and through the right ventricle. After its oxygenation in the lungs, the blood flows through pulmonary veins to

the left atrium. In the alveolar capillaries of the lungs, it accepts oxygen, which diffuses through the alveolar wall from the inspired air; at the same time the CO_2 diffuses from the blood into the alveoli (Fig. 25.5) and is removed with the expired air.

The rate of diffusion of gases in and out of the blood is determined by the difference in partial pressures between alveolar air and blood. Table 25.3 shows the pO_2 and pCO_2 in the lungs. Compared with the atmospheric air, pCO_2 in the alveolar air is slightly higher and pO_2 slightly lower (this is due to the water vapor pressure). Carbon dioxide is much more soluble in water than oxygen, and equilibrates with blood more rapidly. Therefore, when problems develop, one first notices a decrease in blood pO_2 (hypoxia). An increase in pCO_2 (hypercapnia) occurs later and usually indicates a more severe disease. The other major factor determining gas exchange is the rate at which the blood flows through the lungs (the perfusion rate). Normally the alveolar ventilation rate is approximately 4 L/min and the perfusion 5 L/min (the ratio of ventilation to perfusion (Va/Q) is 0.8).

Table 25.3

Partial pressures of oxygen and carbon dioxide in atmospheric air, lung alveoli, and the blood [kPa(mm Hg)]

	Dry air	Alveoli	Systemic arteries	Tissue
pO_2	21.2(39)	13.7(98)	12(90)	5.3(40)
pCO_2	<0.13(0.1)	5.3(40)	5.3(40)	6.0(45)
Water vapor		6.3(47)		

Partial pressure gradients determine diffusion of gases through the alveolar/blood barrier (1 kPa = 7.5 mmHg).

Different combinations of disturbed ventilation and perfusion may occur

In pathologic conditions, some parts of the lung may be well-perfused, but poorly ventilated. This occurs when some alveoli collapse and are unable to

exchange gases. As a result, the blood pO_2 decreases, because there is no diffusion of oxygen from the alveolar air. The presence of oxygen-poor blood in the arterial circulation is known as **the 'shunt' condition**. On the other hand, when ventilation is adequate but perfusion poor, gas exchange cannot take place: in such cases, part of the lung behaves as if it had no alveoli at all, forming **the 'physiologic dead space'.** Examples of conditions related to poor ventilation, poor perfusion, or the combination of both, are given below:

- Rib-cage deformities impair ventilation by limiting lung movement.
- Chest trauma may decrease ventilation as a result of lung collapse; impairing ventilation.
- Alveoli may be actually destroyed in pulmonary emphysema.
- Inadequate synthesis of surfactant leads to the collapse of alveoli, impairment of ventilation and to the respiratory distress syndrome.
- Obstruction or narrowing of the bronchial tree (a mechanical obstruction by inhaled objects, or narrowing by a growing tumor): this impairs ventilation.
- Constriction of the bronchi occurs in asthma impairing ventilation.
- The efficiency of ventilation may be reduced by impaired elasticity of the lung or dysfunction of relevant muscles (the diaphragm and intercostal muscles of the chest wall).
- Fluid present in the alveoli (pulmonary edema) impairs ventilation by affecting diffusion of gases.
- Defects in the neural control impair ventilation through affecting lung movement.
- Lung perfusion is compromised in circulatory problems such as shock and heart failure.

Table 25.4 lists pathologic conditions related to gas exchange.

Table 25.4

Blood partial pressures of oxygen and carbon dioxide depend on lung perfusion and ventilation

	Alveolar pO_2	Alveolar pCO_2	Arterial pO_2	Arterial pCO_2	Comment
Poor ventilation, adequate perfusion	↓	↑	↓	Normal	Physiologic shunt
Adequate ventilation, poor perfusion	↑	↓	↓*	↑*	Physiologic dead space

*Depending on a degree of shunt.

*Depending on a degree of shunt.

Handling of carbon dioxide by erythrocytes

Erythrocytes transport CO_2 to the lungs in a 'fixed' form – as bicarbonate

Human metabolism produces CO_2 at a rate of 200–800 mL/min. The CO_2 dissolves in water and generates carbonic acid, which in turn dissociates into hydrogen and bicarbonate ions. Thus, CO_2 generates large numbers of hydrogen ions:

$$CO_2 + H_2O \rightleftharpoons H_2CO_3 \rightleftharpoons H^+ + HCO_3^-$$

In plasma, the above reaction is nonenzymatic and proceeds slowly, generating only minute amounts of carbonic acid, which remains in equilibrium with a large amount of dissolved CO_2. However, the same reaction in the erythrocytes is catalyzed by carbonic anhydrase, which 'fixes' CO_2 as bicarbonate. The generated hydrogen ion is buffered by hemoglobin.

The bicarbonate ion produced by erythrocyte carbonic anhydrase reaction moves to plasma in exchange for chloride ion (the 'chloride shift') (Fig. 25.6). As much as 70% of all CO_2 produced in tissues becomes bicarbonate; approximately 20% is carried 'fixed' to hemoglobin as carbamino groups, and only 10% remains dissolved in plasma.

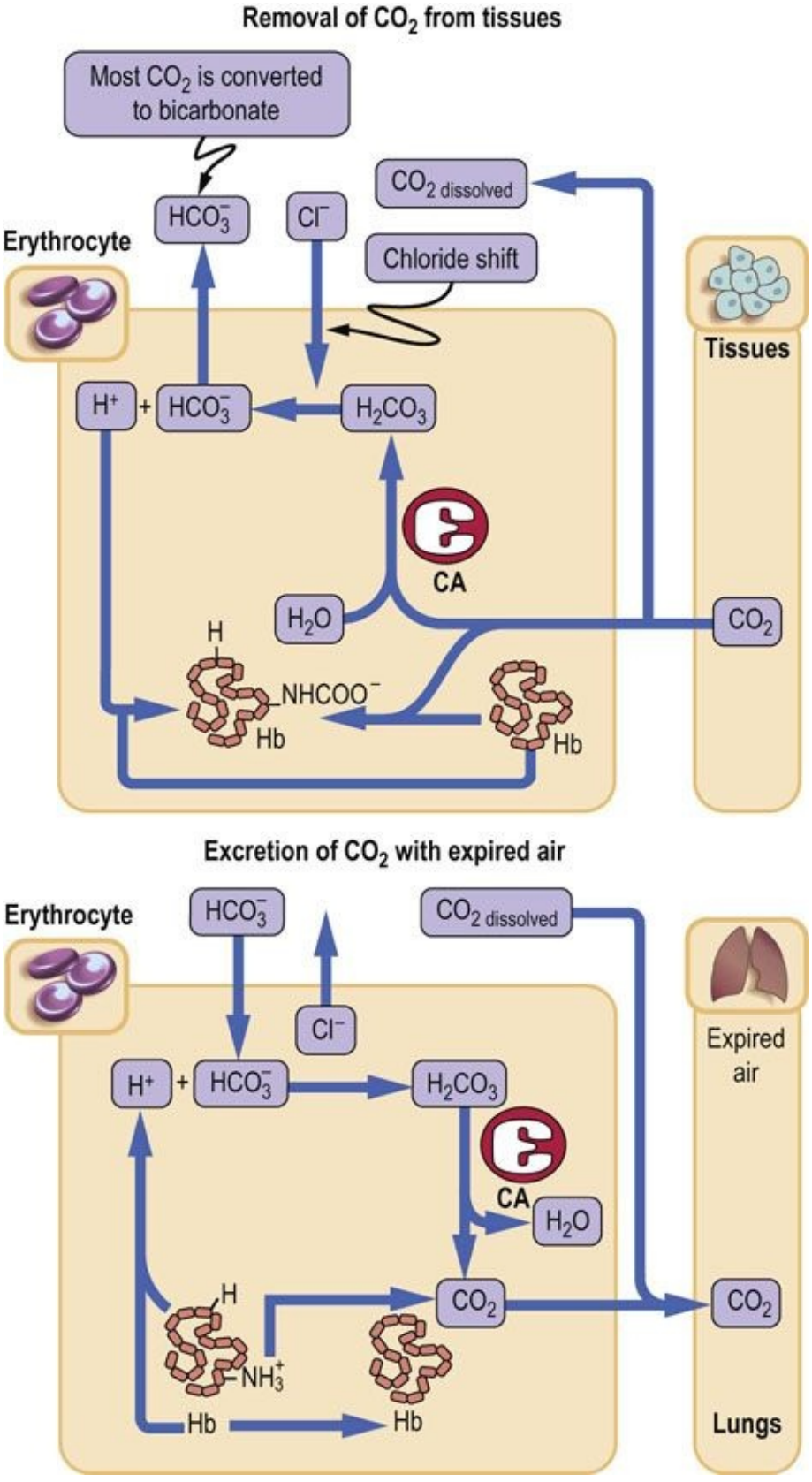

Removal of CO_2 from tissues
Most CO_2 is converted to bicarbonate
HCO_3^-
Cl^-
CO_2 dissolved
Erythrocyte
Chloride shift
Tissues
H^+ + HCO_3^-
H_2CO_3
CA
H_2O
CO_2
H
$NHCOO^-$
Hb
Hb
Excretion of CO_2 with expired air
Erythrocyte
HCO_3^-
CO_2 dissolved
Cl^-
Expired air
H^+ + HCO_3^-
H_2CO_3
CA
H_2O
H
CO_2
CO_2
NH_3^+
Hb
Hb
Lungs

FIG. 25.6 CO_2 transport by the erythrocytes.
Erythrocyte carbonic anhydrase converts approximately 70% of the CO_2 produced in tissues into bicarbonate for transport to the lungs: approximately 20% of the total amount is transported bound to hemoglobin, as carbamates ($-NHCOO^-$) and the rest as dissolved gas in plasma. CA, carbonic anhydrase.

In the lungs, higher pO_2 facilitates dissociation of CO_2 from hemoglobin. This is known as the **Haldane effect**. The hemoglobin releases its hydrogen ion, which reacts with bicarbonate, and forms carbonic acid, which, in turn, releases CO_2.

Handling of bicarbonate by the kidneys

The kidneys control plasma bicarbonate concentration and the removal of the hydrogen ion. In common with erythrocytes, renal tubular cells (proximal and distal) contain carbonic anhydrase.

Proximal tubules reabsorb bicarbonate in the process aided by carbonic anhydrase

Normally, bicarbonate is reabsorbed in the proximal tubule and the urine is almost bicarbonate-free. The surfaces of the renal tubular cells facing the lumen are impermeable to bicarbonate. The filtered bicarbonate combines with hydrogen ion secreted by the cells, and forms carbonic acid, which is converted into CO_2 by carbonic anhydrase located on the luminal membrane. The CO_2 diffuses into cells, where intracellular carbonic anhydrase converts it back into carbonic acid, dissociating into hydrogen and bicarbonate ions. Bicarbonate is returned to the plasma, and the hydrogen ion is secreted into the lumen of the tubule to trap more filtered bicarbonate. Note that, in this process, hydrogen ion is used exclusively to aid bicarbonate reabsorption, and there is no net excretion (Fig. 25.7).

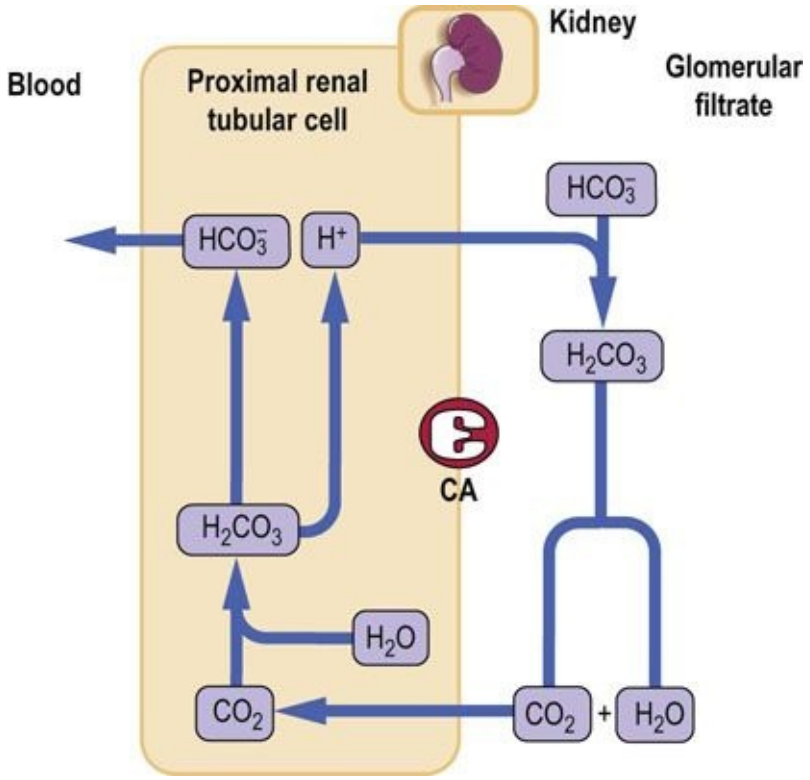


FIG. 25.7 Bicarbonate reabsorption in the kidney.
Bicarbonate reabsorption takes place in the proximal tubule. There is no net excretion of hydrogen ion. CA, carbonic anhydrase.

Distal tubules generate new bicarbonate and excrete hydrogen

The situation is different in the distal tubule where bicarbonate is generated. The mechanism is identical to that of bicarbonate reabsorption, but this time there is both a net loss of hydrogen ions from the body and a net gain of bicarbonate. The CO_2 diffuses into cells. The distal tubule carbonic anhydrase converts it into carbonic acid, which dissociates into hydrogen ion and bicarbonate. Bicarbonate is transported to the plasma, and hydrogen ion is secreted into the tubule lumen. However, no bicarbonate is present in the lumen of the distal tubule (all has been reabsorbed earlier) and the hydrogen ion is buffered (trapped) by phosphate ions present in the filtrate, and by ammonia synthesized by the proximal tubules. It is subsequently excreted in the urine (Fig. 25.8).

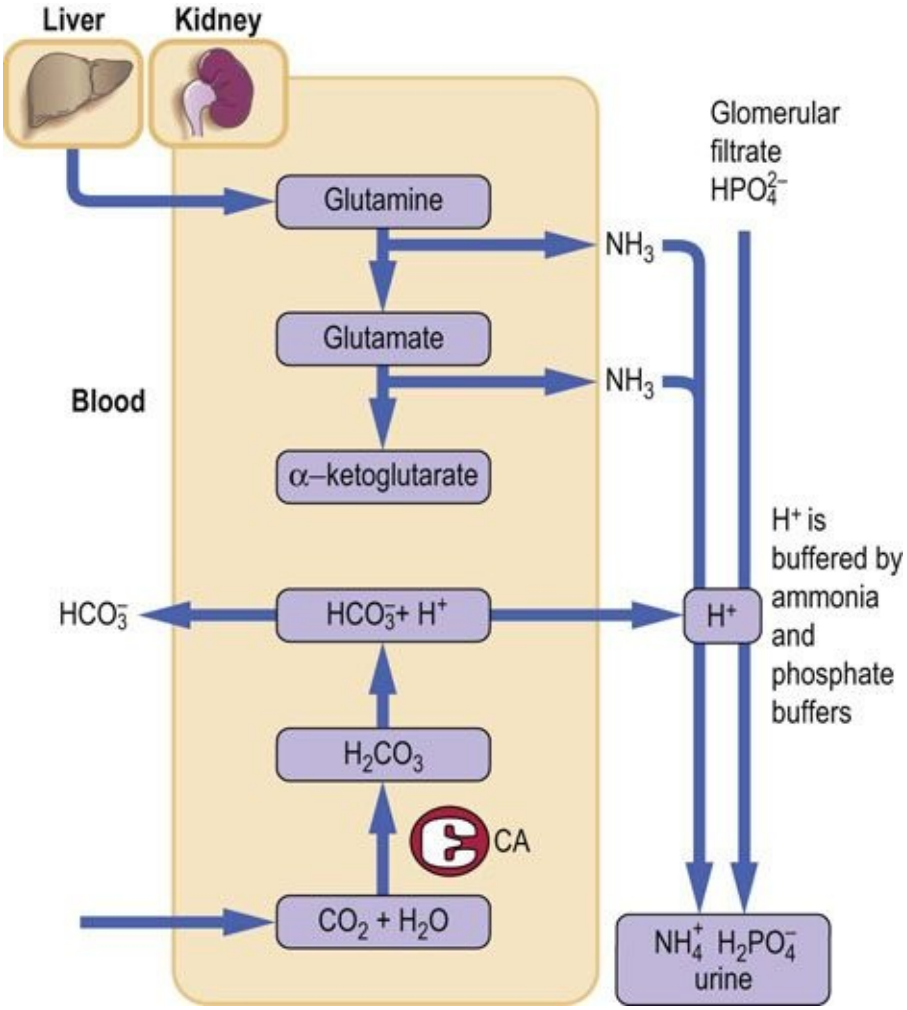


FIG. 25.8 Hydrogen ion excretion by the kidney.
Excretion of the hydrogen ion takes place in the distal tubules. Hydrogen ion reacts with ammonia, forming the ammonium ion. Hydrogen ion is also buffered in the tubule lumen by phosphate. Approximately 50 mmol of hydrogen ion is excreted daily. CA, carbonic anhydrase.

Ammonia generated by glutaminase reaction participates in the excretion of hydrogen ion

Ammonia is generated during the transformation of glutamine into glutamic acid catalyzed by glutaminase. Ammonia diffuses across the luminal membrane, allowing hydrogen ion to be trapped inside the tubule as the ammonium ion **(NH_4^+)**, to which the membrane is impermeable.

Disorders of the acid–base balance

Classification of the acid–base disorders

The concept of respiratory and metabolic components of the acid–base balance forms a basis for the classification of acid–base balance disorders (Fig. 25.3). They are divided into acidosis or alkalosis. Acidosis is a process that leads to the accumulation of hydrogen ion. Alkalosis causes a decrease in hydrogen ion concentration (acidemia and alkalemia are terms that simply describe blood pH). Thus, acidosis and alkalosis result in academia and alkalemia, respectively (see the Clinical Box on this page).

Clinical box Essential definitions

An acid, according to the Brønsted–Lowry definition, is ‘a molecular species that has a tendency to lose a hydrogen ion, forming a conjugate base’.
Acidemia is an excess of hydrogen ion in blood.
Alkalemia is decreased concentration of hydrogen ion in blood.
Acidosis is a process that leads to accumulation of hydrogen ion.
Alkalosis is the process that decreases the amount of hydrogen ion.

There are four main disorders of acid–base balance

The key to further classification is the ‘location’ of the primary cause within the respiratory and metabolic components of the system. If the primary cause is the change in pCO_2, the acidosis or alkalosis is called **respiratory**, and if it is bicarbonate, the acidosis or alkalosis is called **metabolic**. Thus there are four main disorders of acid–base balance: **respiratory acidosis**, **metabolic acidosis**, **respiratory alkalosis**, and **metabolic alkalosis** (Fig. 25.3). However, mixed disorders can also develop; we consider them later.

Lung and the kidney work in a concerted way to minimize changes in plasma pH: this is known as the compensation

of acid–base disorders

Acidosis is accompanied by a decreased ratio of plasma bicarbonate to pCO_2, and alkalosis by an increased ratio. Whenever a problem occurs, the compensating mechanisms are triggered to bring the hydrogen ion concentration back towards normal. This translates into normalizing the [bicarbonate]/pCO_2 ratio in the Henderson–Hasselbalch equation.

Consequently, when the respiratory acidosis causes an increase in pCO_2, the kidney will increase the generation of bicarbonate, increasing its plasma concentration and normalizing the ratio. Conversely, when diabetic ketoacidosis causes depletion of plasma bicarbonate, ventilation rate increases, pCO_2 decreases and the ratio of bicarbonate/pCO_2 changes towards normal. While respiratory compensation can occur within minutes, metabolic compensation takes hours to days to develop fully (Table 25.5).

Table 25.5

Respiratory and metabolic compensation in the acid–base disorders

Acid–base disorder	Primary change	Compensatory change	Timescale of compensatory change
Metabolic acidosis	↓ plasma bicarbonate	↓ pCO_2 (hyperventilation)	Minutes/hours
Metabolic alkalosis	↑ plasma bicarbonate	↑ pCO_2 (hypoventilation)	Minutes/hours
Respiratory acidosis	↑ pCO_2	↑ renal bicarbonate generation: ↑ plasma bicarbonate	Days
Respiratory alkalosis	↓ pCO_2	↓ renal bicarbonate reabsorption: ↓ plasma bicarbonate	Days

Respiratory and metabolic compensation in the acid–base disorders minimizes changes in the blood pH. A change in the respiratory component leads to metabolic compensation, and a change in the metabolic component stimulates respiratory compensation.

Acidosis

Respiratory acidosis occurs most often in lung disease and results from decreased ventilation

The most common cause is the chronic obstructive airways disease (COAD). Severe asthmatic attack can result in respiratory acidosis because of bronchial constriction. Respiratory acidosis often accompanies hypoxia (respiratory failure); in such a case, an increase in pCO_2 often parallels the decrease in pO_2

(Table 25.6, see Clinical Box on page 336).

Table 25.6

Causes of acid–base disorders

Metabolic acidosis	Respiratory acidosis	Metabolic alkalosis	Respiratory alkalosis
Diabetes mellitus (ketoacidosis)	Chronic obstructive airways disease	Vomiting (loss of hydrogen ion)	Hyperventilation (anxiety, fever)
Lactic acidosis (lactic acid)	Severe asthma	Nasogastric suction (loss of hydrogen ion)	Lung diseases associated with hyperventilation
Renal failure (inorganic acids)	Cardiac arrest	Hypokalemia	Anemia
Severe diarrhea (loss of bicarbonate)	Depression of respiratory center (drugs, e.g. opiates)	Intravenous administration of bicarbonate (e.g. after cardiac arrest)	Salicylate poisoning
Surgical drainage of intestine (loss of bicarbonate)	Failure of respiratory muscles (e.g. poliomyelitis, multiple sclerosis)		
Renal loss of bicarbonate (renal tubular acidosis type 2 – rare)	Chest deformities		
Renal tubular acidosis – rare)	Airway obstruction		

Respiratory acidosis is common and is caused primarily by diseases of the lung that affect gas exchange. **Respiratory alkalosis** is rarer and is caused by hyperventilation, which decreases pCO_2. **Metabolic acidosis** is common and results from either overproduction or retention of nonvolatile acids in the circulation. **Metabolic alkalosis** is rarer: its most common causes are vomiting and gastric suction, both causing loss of hydrogen ion from the stomach.

Clinical box Respiratory alkalosis is caused by hyperventilation

A 25-year-old man was admitted to hospital with an asthmatic attack. Peak expiratory flow rate was 75% of his best. His blood gas values were pO_2 9.3 kPa (70 mmHg) and pCO_2 4.0 kPa (30 mmHg), with pH 7.50 (hydrogen ion concentration = 42 nmol/L). He was treated with nebulized salbutamol, a β_2-adrenergic stimulant (Chapter 39), which is a bronchodilator, and made a good recovery.

Comment.

This man's blood gases show a mild degree of respiratory alkalosis caused by hyperventilation and 'blowing off' the CO_2. Respiratory alkalosis causes reduction in serum levels of ionized calcium, leading to neuromuscular irritability. Ventilatory impairment that leads to CO_2 retention and respiratory acidosis is characteristic of a severe asthma. Reference ranges are given in Table 25.2.

Metabolic acidosis results from excessive production, or inefficient metabolism or excretion, of nonvolatile acids

A classic example of metabolic acidosis is the diabetic ketoacidosis, when ketoacids, acetoacetic acid, and β-hydroxybutyric acid accumulate in the plasma (Chapter 21). Acidosis may also occur during extreme physical exertion, when there is accumulation of lactic acid generated from muscle metabolism; in normal circumstances, lactate would be quickly metabolized on cessation of exercise. However, when large amounts of lactate are generated as a consequence of hypoxia, lactic acidosis may become life-threatening, as happens, for instance, in shock (Table 25.6).

Excretion of nonvolatile acids is also impaired in renal failure, and this also leads to metabolic acidosis. Renal failure develops when the perfusion of the kidneys is inadequate (e.g. in trauma, shock, or dehydration) or if there is an intrinsic kidney disease such as glomerulonephritis (inflammatory reaction in the renal tubular tissue).

Excessive loss of bicarbonate can also be a cause of metabolic acidosis. This is common when bicarbonate present in the intestinal fluid is lost as a result of severe diarrhea or surgical drainage after bowel surgery.

Impaired bicarbonate reabsorption and hydrogen ion secretion causes rare renal tubular acidoses

Defects in renal handling of bicarbonate and hydrogen ion lead to a group of relatively rare disorders known as renal tubular acidoses (RTA). The proximal RTA is caused by impaired reabsorption of bicarbonate, and the distal RTA by impaired hydrogen ion excretion. Proximal RTA is usually accompanied by other defects in proximal transport mechanisms (this is known as the Fanconi syndrome).

Generally, acidosis is a much more common condition than alkalosis.

Alkalosis

Alkalosis is rarer than acidosis

A mild respiratory alkalosis may be a consequence of hyperventilation during exercise, anxiety attack, or fever. It also occurs in pregnancy. Metabolic alkalosis is often associated with abnormally low serum potassium

concentration, as a result of cellular buffering. Cellular entry or exit of potassium ion is associated with the movement of hydrogen ion in an opposite direction. Thus, **alkalosis can cause hypokalemia, and hypokalemia** (**Chapter 22**) **may lead to alkalosis.** Severe metabolic alkalosis may also occur as a result of the massive loss of hydrogen ion from the stomach during vomiting (see Clinical Box on this page), or as a result of nasogastric suction after surgery. Lastly, it may occur when too much bicarbonate is given intravenously: for instance, during resuscitation from cardiac arrest (Table 25.6).

Clinical box Vomiting can lead to metabolic alkalosis

A 47-year-old man came to the outpatient clinic with a history of intermittent profuse vomiting and loss of weight. He had tachycardia, reduced tissue turgor, and hypotension. His blood pH was 7.55 (hydrogen ion concentration 28 nmol/L) and pCO_2 was 6.4 kPa (48 mmHg). His bicarbonate concentration was 35 mmol/L and there was also hyponatremia and hypokalemia.

Comment.

This patient presents with metabolic alkalosis caused by the loss of hydrogen ion through vomiting. Investigations showed gastric outlet obstruction due to scarring from chronic peptic ulceration. He subsequently underwent surgery for pyloric stenosis, with a good outcome. Note the increased pCO_2 as a result of respiratory compensation of metabolic alkalosis.

Mixed acid–base disorders

More than one acid–base disorder can exist in one patient. The result of this is a mixed acid–base disorder, sometimes quite difficult to diagnose (see Clinical Box below and Table 25.7).

Table 25.7

Comparison of simple and mixed disorders of the acid–base

balance

A. Mixed metabolic and respiratory acidosis

Condition	pH	pCO_2	Plasma bicarbonate
Metabolic acidosis	↓	↓ (respiratory compensation)	↓ (primary change)
Respiratory acidosis	↓	↑ (primary change)	↑ (metabolic compensation)
Mixed respiratory and metabolic acidosis	↓↓	↑ (respiratory acidosis)	↓ (metabolic acidosis)

B. Mixed metabolic and respiratory alkalosis (rare)

Disorder	pH	pCO_2	Plasma bicarbonate
Metabolic alkalosis	↑	↑ (respiratory compensation)	↑ (primary change)
Respiratory alkalosis	↑	↓ (primary change)	↓ (metabolic compensation)
Mixed respiratory and metabolic alkalosis	↑↑	↓ (respiratory alkalosis)	↑ (metabolic acidosis)

Mixed acid–base disorders result in a greater change in blood pH than simple disorders; they may pose diagnostic difficulties.

Clinical box Respiratory and metabolic disorders of acid–base balance can occur together

CARDIAC ARREST

During resuscitation of a 60-year-old man from a cardiorespiratory arrest, blood gas analysis revealed pH 7.00 (hydrogen ion concentration 100 nmol/L) and pCO_2 7.5 kPa (52 mmHg). His

bicarbonate concentration was 11 mmol/L. pO_2 was 12.1 kPa (91 mmHg) during treatment with 48% oxygen.

Comment.

This patient presents with a mixed disorder: a respiratory acidosis caused by lack of ventilation, and metabolic acidosis caused by the hypoxia that had occurred before oxygen treatment was instituted. The acidosis was caused by an accumulation of lactic acid: the measured lactate concentration was 7 mmol/L (reference range is 0.7–1.8 mmol/L (6–16 mg/dL)). The terms acidosis and alkalosis do not just describe blood pH changes: they relate to the processes that result in these changes. Therefore, in some instances, two independent processes may occur: for example, a patient may be admitted to hospital with diabetic ketoacidosis and coexisting emphysema causing respiratory acidosis. The final result could be a more severe change in pH than would have resulted from a simple disorder (Table 25.5). Any combination of disorders can occur; the skills of an experienced physician are usually required to diagnose this.

Summary

■ Maintenance of the hydrogen ion concentration within a narrow range is vital for survival.
■ Acid–base balance is regulated by the concerted action of lungs and kidneys. The erythrocytes play a key role in the transport of carbon dioxide in blood.
■ Main buffers in blood are hemoglobin and bicarbonate, whereas in the cells they include proteins and phosphate. The bicarbonate buffer system communicates with atmospheric air.
■ Acid–base disorders are acidosis and alkalosis and each of them can be either metabolic or respiratory.
■ Measurement determination of pH, pCO_2 and bicarbonate, and pO_2, is a first-line investigation and is frequently required in emergencies.

Active learning

1. Describe how the bicarbonate buffer copes with an addition of an acid to the system.
2. Compare bicarbonate handling by the proximal and distal tubules of the kidney.
3. Outline the role of ventilation in acid–base disorders.
4. Which disorders of the acid–base balance may be associated with gastrointestinal surgery?
5. Discuss the association between acid–base disorders and plasma potassium concentration.

Further reading

Corey, HE. Bench-to-bedside review: Fundamental principles of acid–base balance. *Critical Care*. 2005; 9:184–192.

Kellum, JA. Disorders of acid–base balance. *Crit Care Med*. 2007; 35:2630–2636.

Websites and downloads

Appel, SJ, Downs, CA, Understanding acid–base balance. *Nursing: Fall* 2008; 38:9–1.

http://journals.lww.com/nursing/Fulltext/2008/09002/Understanding_acid_base_balance.3.aspx

McNamara, J, Worthley, LIG, Acid–Base Balance: Part I. Physiology. *Critical Care and Resuscitation* 2001; 3:181–18. http://cicm.org.au/journal/2001/september/ABa1.pdf

Neligan, PJ, Deutschman, CS. Acid–base balance in critical care medicine. www.ccmtutorials.com/renal/Acid%20Base%20Balance%20in%20Critical%20Care%20Medicine-NELIGAN.pdf.

Respiratory Acid/Base Balance, Gas Transport and Erythrocytes. www.sci.utah.edu/~macleod/bioen/be6000/notes/L14-acid–base.pdf

CHAPTER 26

Bone Metabolism and Calcium Homeostasis

Marek H. Dominiczak and William D. Fraser

Learning objectives

After reading this chapter you should be able to:

- Describe the chemical composition of bone, and the process of bone mineralization.
- Recognize major cells in bone and their interactions in the bone remodeling cycle.
- Understand the role of major factors contributing to the regulation of serum calcium concentration.
- Explain the pivotal role of the parathyroid hormone-related protein in hypercalcemia of malignancy.
- Understand the role of vitamin D and its metabolism in health and disease.
- Define osteoporosis and its causes.

Introduction

Many cell functions depend on tight control of extracellular calcium concentration.

These include neural transmission, cellular secretion, contraction of muscle cells, cell proliferation, the permeability of cell membranes, blood clotting, and the mineralization of bone. Bone serves as a reservoir of calcium when deficiency exists, and as its store when the body is calcium replete. The skeleton contains 99% of the calcium present in the body in the form of hydroxyapatite; the remainder is distributed in the soft tissues, teeth, and the extracellular fluid (ECF).

Bone structure and bone remodeling

Bone is a specialized connective tissue that, along with cartilage, forms the skeletal system

In addition to serving a supportive and protective role, bone is the site of substantial metabolic activity. There are two types of bone: the thick, densely calcified external bone (cortical or compact bone) and a thinner, honeycomb network of calcified tissue (trabecular bone).

Collagen and hydroxyapatite are the main components of the bone matrix

Within the bone matrix, the major protein (90%) is type 1 collagen (see Box on p. 12). The calcium-rich crystals of hydroxyapatite ($Ca_{10}[PO_4]_6[OH]_2$) are found on, within, and between the collagen fibers. The attachment of hydroxyapatite to collagen and the calcification of bone are, in part, controlled by the presence of glycoproteins and proteoglycans with a high ion-binding capacity. Collagen fibers orientate so that they have the greatest density per unit volume and are packed in layers, giving bone the lamellar structure observed on microscopy. Post-translational modifications of collagen result in the formation of intra-and intermolecular pyridinoline and pyrrole crosslinks. This microarchitecture allows bone to function as the major reservoir of calcium for the body.

The noncalcified organic matrix within bone, known as osteoid, becomes mineralized through two mechanisms. Within the extracellular space of the bone, plasma membrane-derived matrix vesicles act as a focus for deposition of calcium phosphate. Crystallization eventually obliterates the vesicle membrane, leaving a collection of clustered hydroxyapatite crystals. Within this environment, the bone-forming cells **(osteoblasts)** secrete packets of matrix proteins that rapidly mineralize, and these combine with matrix vesicle-derived crystals. Pyrophosphate present in the matrix inhibits this process. The alkaline phosphatase secreted by the osteoblasts destroys pyrophosphate, allowing mineralization to proceed. Mineralization is highly dependent on an adequate supply of calcium and phosphate. When mineral deprivation exists, there is an increase in the percentage of the nonmineralized organic matrix (osteoid) within bone, resulting in the clinical condition of **osteomalacia**.

Bone constantly changes its structure through remodeling

Small amounts of calcium are exchanged daily between bone and the ECF as a result of constant bone remodeling, *i.e.* coupled processes of resorption by the bone-resorbing cells **(osteoclasts)** and formation by the osteoblasts (Fig. 26.1). This exchange maintains a relative calcium balance between newly formed and older bone. Bone undergoes constant mechanical adaptation. Increased mechanical load stimulates bone formation and excess osteoclastic activity underpins several diseases, in particular **osteoporosis, rheumatoid arthritis,** and **metastatic cancers**.

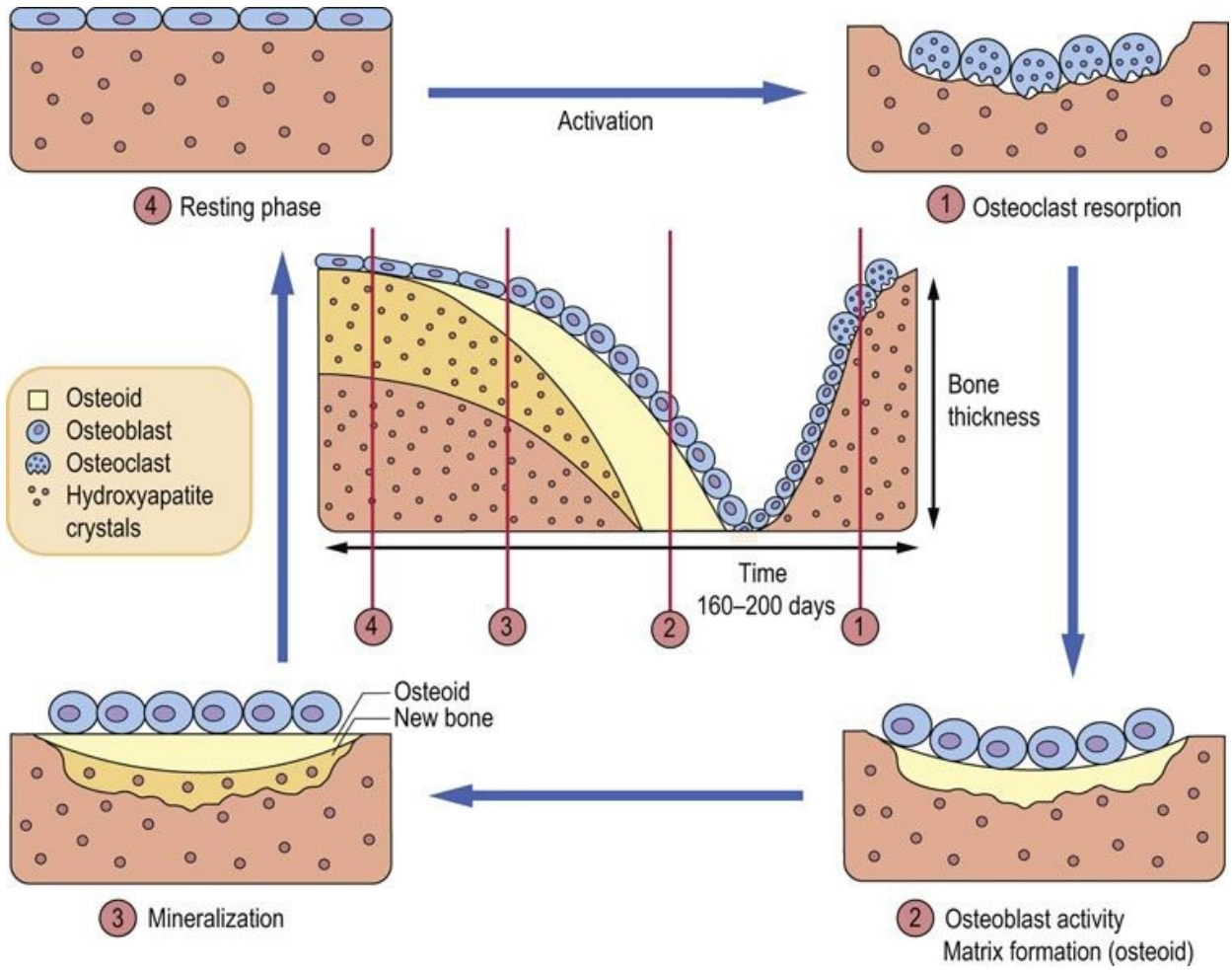


FIG. 26.1 Maintaining bone mass: the bone remodeling cycle.
Resorption and formation of bone by osteoclasts and osteoblasts is coupled. The processes numbered 1-4 are set in a timeline in the central panel.

Osteoclasts are bone-resorbing cells

Osteoclasts are multinucleated, giant tissue-specific macrophages. They are derived from pluripotent hematopoietic mononuclear cells in the bone marrow and remain in contact with a calcified surface.

RANK receptor and its ligand RANKL are essential for differentiation, maturation, and regulation of osteoclasts

The maturation of osteoclasts from their progenitor cells is directed by growth factors, particularly monocyte-colony stimulating factor (M-CSF). Another essential factor is the membrane receptor protein structurally related to the tumor necrosis factor receptor, called **receptor activator of nuclear factor NFκB** (RANK). It binds the TNF-related cytokine called **RANK ligand** (RANKL). The binding of RANKL to RANK can be decreased when it binds instead to **osteoprotegerin** (OPG), a protein which also belongs to the TNF receptor superfamily. RANKL and OPG together control differentiation and activation of osteoclasts. RANKL stimulates and OPG inhibits bone resorption. Importantly, estrogens induce OPG synthesis.

RANK controls osteoclasts through intracellular signaling cascades and transcription factors

Essentially, RANK prepares the osteoclast to resorb bone. It stimulates intracellular signaling cascades (Chapter 40), which in turn activate transcription factors that control genes. It also cooperates with other immunoglobulin-like membrane receptors. Intracellular signaling involves, among others, adaptor molecules known as TNF receptor-associated cytoplasmic factors (TRAFs). TRAFs assemble further signaling proteins and activate pathways involving NFκB and activator protein-1 (AP-1). Other pathways involve c-Jun terminal kinase, p38 stress-activated protein kinase, extracellular signal-regulated kinase (ERK) and the src pathway, which involves phosphatidylinositol-3-kinase (PI3K) and Akt kinase (Chapter 40). The effect of this is activation of a transcription factor known as the nuclear factor of activated T cells-2 (NFAT2). The end result is the induction of genes coding for tartrate-resistant acid phosphatase, cathepsin K, calcitonin and the β2 integrin, which directly control bone resorption. Osteoclast resorption of bone releases a range of molecules: collagen peptides, pyridinoline crosslink fragments, and calcium from the bone matrix (through the action of lysosomal enzymes, collagenases and cathepsins).

Collagen breakdown products (hydroxyproline) in serum and urine, and collagen fragments (**amino-terminal and carboxy-terminal telopeptides** NTX and CTX, respectively) can be measured in clinical laboratories.

Parathyroid hormone contributes to osteoclast activation

Parathyroid hormone (PTH) activates osteoclasts indirectly via osteoblasts and calcitonin. Local factors such as the cytokines interleukin-1 (IL-1), tumor necrosis factor (TNF), transforming growth factor-β (TGF-β) and interferon-α (INF-α) are also important regulators of osteoclasts and act through RANKL and OPG.

Osteoblasts are bone-forming cells

The osteoblast is derived from mesenchyme. Mature osteoblasts synthesize type 1 collagen, osteocalcin, cell attachment proteins (thrombospondin, fibronectin, bone sialoprotein, osteopontin), proteoglycans, and growth-related proteins. They control bone mineralization.

The function of the osteoblasts is regulated by several hormones and growth factors

PTH binds to a specific receptor and stimulates production of cyclic adenosine monophosphate (cAMP), ion and amino acid transport, and collagen synthesis. Calcitriol (1,25-dihydroxycholecalciferol; $1,25(OH)_2D_3$) stimulates synthesis of alkaline phosphatase, matrix, and bone-specific proteins, and can decrease osteocalcin secretion. Growth factors such as osteoblast stimulating factor 1 (OSF-1), TGF-β, insulin-like growth factors (IGF-1 and IGF-2), and platelet-derived growth factor (PDGF) serve as autocrine regulators of osteoblast function. There are also the bone morphogenetic proteins (BMP) which belong to the TGF-β superfamily. Serum biochemical markers reflecting osteoblast function are **bone-specific alkaline phosphatase**, **osteocalcin**, and markers of collagen formation: **carboxy-terminal procollagen extension peptide** (ICTP) **and amino or carboxy-terminal procollagen extension peptides** (PINP, P1CP).

A protein that is a member of the LDL-receptor family plays an important role in osteoblast differentiation

Two signaling pathways are important in skeletal development: the Wnt/β-catenin pathway and the TGF-β/BMP pathway (Wnt is a widely present signalling glycoprotein involved particularly in embryo development). The LDL-receptor-related protein 5 (LRP5, Chapter 18), together with another receptor, activates the Wnt pathway. Mutation of the gene encoding for LRP5 was shown to increase bone mass and the formation of dense bone. On the other hand, the loss-of-function mutation caused osteoporosis.

Serum calcium

The total plasma calcium concentration is maintained between 2.2 and 2.60 mmol/L (8.8–10.4 mg/dL). Calcium exists in the circulation in three forms. The **ionized Ca^{2+}** is the most important, physiologically active form (50% of total calcium). The majority of the remaining calcium, is **protein bound**, mainly to negatively charged **albumin** (40%), and the rest is complexed to substances such as citrate and phosphate (10%).

If plasma protein concentration increases (as, for instance, in dehydration), protein-bound calcium and total serum calcium increase. In conditions of reduced plasma proteins (e.g. liver disease, nephrotic syndrome or malnutrition), the protein-bound calcium concentration is reduced, decreasing the total calcium, although ionized calcium is maintained within the reference range. In many acute and chronic illnesses the albumin concentration decreases. While this decreases the total calcium concentration, it does not change the concentration of the ionized fraction. Therefore, in clinical laboratories, the **'adjusted calcium'** concept is used: the measured value is extrapolated to albumin concentration of 40 g/L (4 g/dL).

$$\text{Adjusted } Ca^{2+} = \text{measured } Ca^{2+}\,(\text{mmol / L}) + 0.02(40 - \text{albumin [g / L]})$$

$$\text{Adjusted } Ca^{2+} = \text{measured } Ca^{2+}\,(\text{mg / dL}) + 0.8(4.0 - \text{albumin [g / dL]})$$

Calcium homeostasis

Parathyroid hormone (PTH) responds to changes in ionized calcium and phosphate

PTH is an 84-amino acid, single-chain peptide hormone secreted by the chief cells of the parathyroid glands. A decrease in extracellular ionized calcium or an increase in serum phosphate concentration stimulates its secretion. Chronic severe magnesium deficiency can inhibit its release from secretory vesicles, and low concentrations of calcitriol interfere with its synthesis. $PTH_{(1–84)}$ is mainly metabolized into a biologically active $PTH_{(1–34)}$ amino-terminal fragment and an inactive carboxy-terminal fragment, $PTH_{(35–84)}$ (Fig. 26.2). Most of the cellular actions of PTH are mediated by G-protein and cAMP signaling.

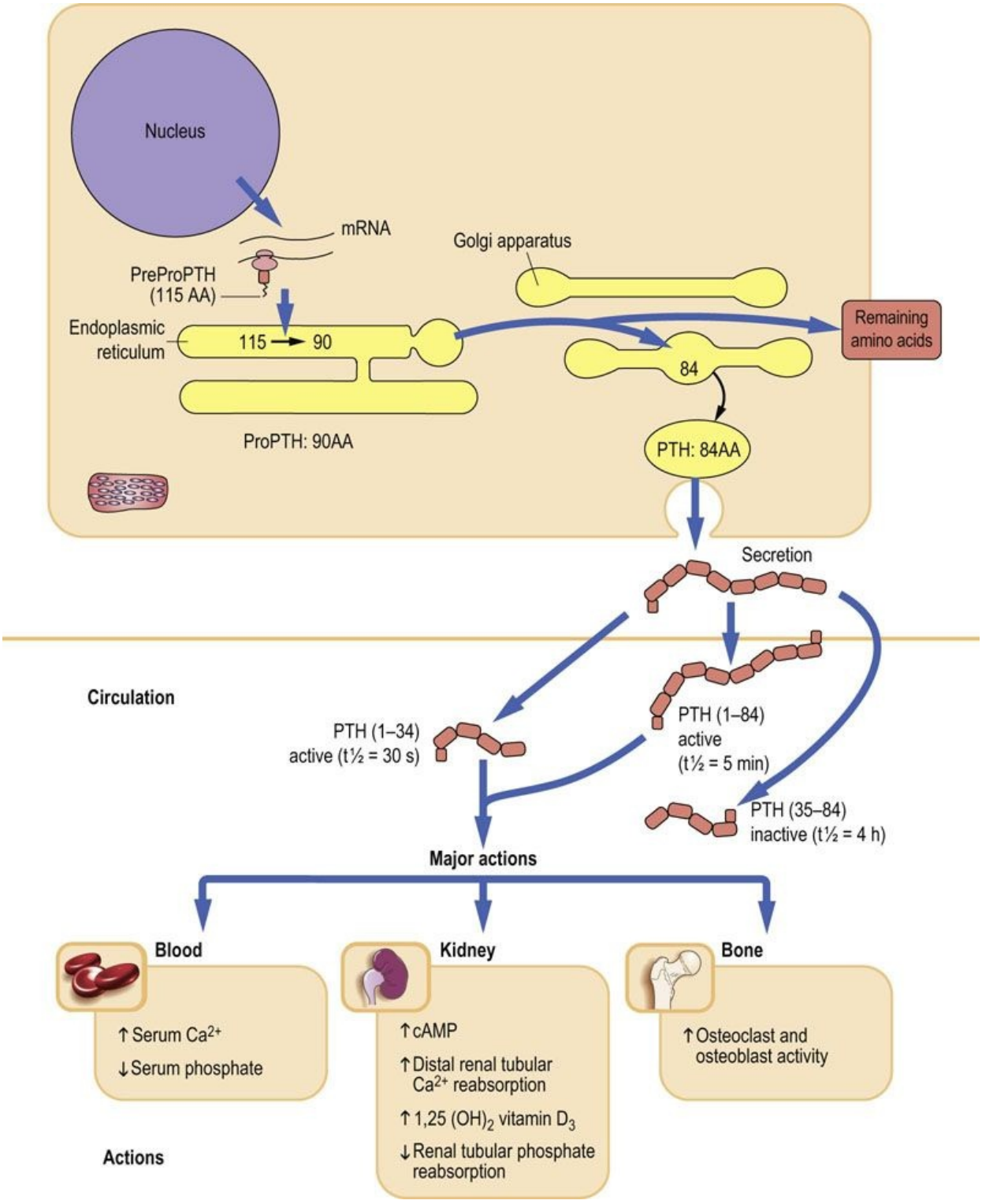


FIG. 26.2 Synthesis and major actions of the parathyroid hormone (PTH).
PTH mobilizes calcium from all available sources and decreases its renal excretion. AA, amino acids.

When plasma calcium decreases, PTH is released from the parathyroid glands, stimulating osteoclast-mediated bone resorption, renal reabsorption of calcium, and its absorption in the small intestine (mediated by calcitriol). PTH secretion is feedback-regulated by calcium: an increase in calcium decreases PTH secretion (Fig. 26.3).

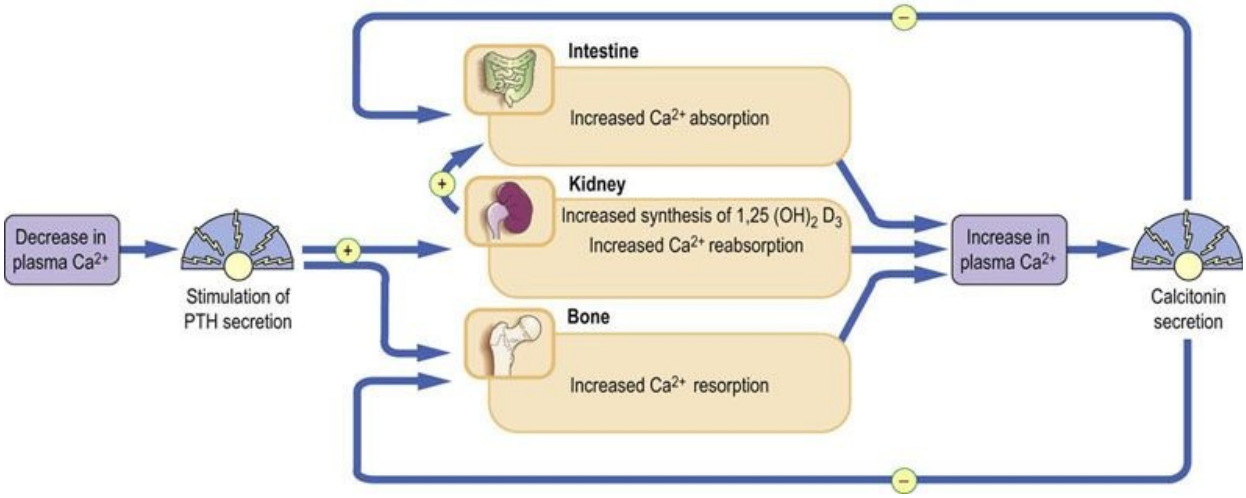


FIG. 26.3 Major hormones influencing calcium homeostasis.
A decrease in plasma ionized calcium stimulates release of PTH; this promotes Ca^{2+} reabsorption from the kidney, resorption from bone, and absorption by the gut via increased production of $1,25(OH)_2D_3$. As a result, plasma calcium increases. Conversely, an increase in plasma ionized calcium stimulates release of calcitonin, which inhibits reabsorption of calcium by the kidney and osteoclast-mediated bone resorption.

Clinical box A woman with severe right-sided flank pain: primary hyperparathyroidism

A 52-year-old woman presented to the accident and emergency department with severe right-sided flank pain. Further questioning revealed a history of recent depression, generalized weakness, recurrent indigestion, and aches in both hands. Blood was detected on stick testing of urine and radiography revealed the presence of kidney stones. The pain settled with opiate analgesia. Serum adjusted calcium was 3.20 mmol/L (12.8 mg/dL; normal range

2.2–2.6 mmol/L, 8.8–10.4 mg/dL), serum phosphate 0.65 mmol/L (2.0 mg/dL; normal range 0.7–1.4 mmol/L, 2.2–5.6 mg/dL), and PTH 16.9 pmol/L (169 pg/mL; normal range 1.1–6.9 pmol/L, 11–69 pg/mL).

Comment.

Most patients with primary hyperparathyroidism are now identified when asymptomatic hypercalcemia is discovered on routine biochemical testing. The primary hyperparathyroidism classically affects the skeleton, kidneys, and gastrointestinal tract, resulting in the well-recognized triad of complaints described as 'bones, stones, and abdominal groans'. Kidney stone disease is now the most common presenting complaint.

Calcium-sensing receptor is a cell surface G-protein coupled receptor

Ionized calcium is maintained within a narrow range through an extracellular calcium-sensing receptor (CaSR), which is a cell-surface G-protein-coupled receptor present in the chief cells of the parathyroid gland, the thyroidal C cells, and along the kidney tubules. Minute changes in ionized calcium modulate cellular function to maintain normocalcemia.

Vitamin D is synthesized in the skin by UV radiation

Vitamin D_2 **(ergocalciferol)** is synthesized in the skin by UV radiation of ergosterol, and vitamin D_3 (cholecalciferol) by UV irradiation of 7-dehydrocholesterol. Vitamin D_3 and its hydroxylated metabolites are transported in the plasma bound to a specific globulin, vitamin D-binding protein (DBP). Cholecalciferol is also found in the diet, where its absorption is associated with other fats, and it is transported to the liver in chylomicrons. It is released from chylomicrons in the liver by DBP and hydroxylated at the 25-position, forming **calcidiol** (25-hydroxycholecalciferol; $25(OH)D_3$). The metabolism of vitamin D is illustrated in Figure 26.4.

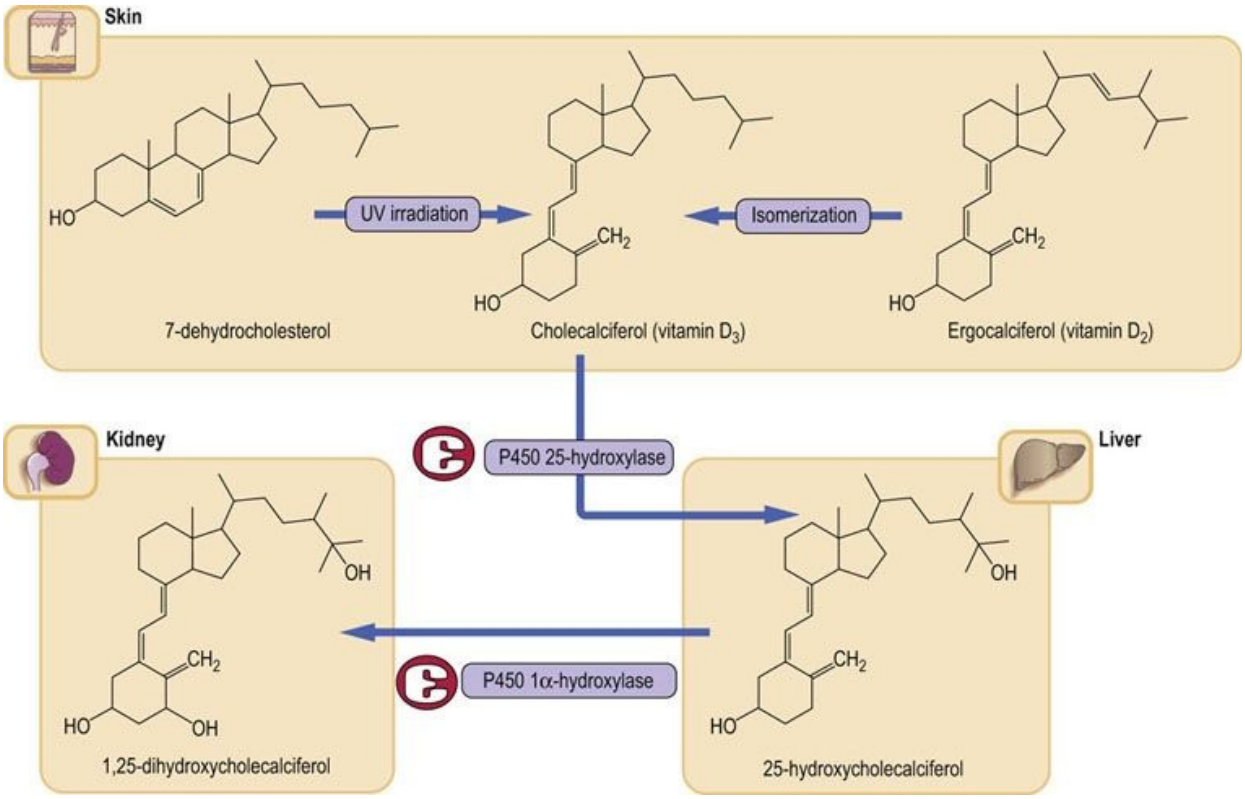


FIG. 26.4 Vitamin D metabolism.
Vitamin D is mainly synthesized in response to the action of sunlight on the skin; a smaller component comes from the diet. Normal liver and kidney function are essential to the formation of the active form $1,25(OH)_2D_3$. Plasma calcium concentration controls the level of $1,25(OH)_2D_3$ through the parathyroid hormone. Note that both hydroxylase enzymes belong to the cytochrome P450 superfamily. $1,25(OH)_2D_3$, 1,25-dihydroxycholecalciferol, calcitriol,$1,25(OH)_2D_3$; 25-hydroxycholecalciferol, calcidiol, $25(OH)D_3$

Calcidiol is the main liver storage form of vitamin D

The 25-hydroxylation step is carried out by a hepatic microsomal enzyme. The hepatic content of $25(OH)D_3$ regulates the rate of 25-hydroxylation. 25(OH)D3 is the major form of the vitamin found in both the liver and the circulation, in each case bound to DBP, Its levels in the circulation reflect hepatic stores of the vitamin. A significant proportion of $25(OH)D_3$ is subject to enterohepatic circulation, being excreted in the bile and reabsorbed in the small bowel. Disturbance in the enterohepatic circulation can lead to deficiency of this vitamin.

Calcitriol (1α,25-dihydroxycholecalciferol; $1,25(OH)_2D_3$)is the active metabolite of vitamin D

The main sites for further hydroxylation of the $25(OH)D_3$ at the 1-position are the renal tubules, although bone and the placenta can also carry out this reaction. The $25(OH)D_3$ 1α-hydroxylase is a mitochondrial enzyme. Its product, **calcitriol** ($1,25(OH)_2D_3$) is the most potent of the vitamin D metabolites and the only naturally occurring form of vitamin D that is active at physiologic concentrations. The 1α-hydroxylase activity is stimulated by PTH, low serum concentrations of phosphate or calcium, vitamin D deficiency, calcitonin, growth hormone, prolactin, and estrogen. Conversely, its activity is feedback inhibited by $1,25(OH)_2D_3$, hypercalcemia, high phosphate concentration and hypoparathyroidism.

$1,25(OH)_2D_3$ is transported in plasma also bound to DBP. Vitamin D may be described as a **hormone**. In the intestinal epithelial cells it binds to a cytoplasmic receptor like other steroid hormones (Chapters 17 and 35) and this ligand–protein complex is transported to the nucleus where it induces gene expression.

The renal tubules, cartilage, intestine, and placenta also contain a 24-hydroxylase, producing the inactive 24,25-dihydroxycholecalciferol ($24,25[OH]_2D_3$). The concentration of the $24,25[OH]_2D_3$ in the circulation is reciprocally related to the level of the $1,25(OH)_2D_3$.

$1,25(OH)_2D_3$ increases the absorption of calcium and phosphate from the gut via active transport by calcium-binding proteins

Together with PTH, $1,25(OH)_2D_3$ stimulates bone resorption by osteoclasts. These effects increase plasma calcium and phosphate concentrations. Low $1,25(OH)_2D_3$ causes abnormal mineralization of newly formed osteoid as a result of low calcium and phosphate availability and reduced osteoblast function. It leads to the development of **rickets** in children or **osteomalacia** in adults.

Calcitonin inhibits bone resorption

Calcitonin is a 32-amino acid peptide synthesized and secreted primarily by the parafollicular cells of the thyroid gland (C cells). Its secretion is regulated by serum calcium through the calcium-sensing receptor (CaSR): an increase in serum calcium results in a proportional increase in calcitonin, and a decrease

elicits a corresponding reduction in calcitonin. Chronic stimulation results in exhaustion of the secretory reserve of the C cells. The precise biological role of calcitonin is not known, but the main effect is inhibition of osteoclastic bone resorption (Fig. 26.3).

Calcium is absorbed in the small intestine and is excreted in urine and feces

Calcium is absorbed predominantly in the proximal small intestine. This is regulated through the quantity of calcium ingested in the diet, and two cellular calcium transport processes: the active, saturable transcellular absorption stimulated by $1,25(OH)_2D_3$, and the nonsaturable paracellular absorption controlled by the concentration of calcium in the intestinal lumen relative to the serum concentration.

In a normal adult taking a Western diet, the amount of calcium intake and its deposition in bone are matched by the excretion in urine and feces. During growth, a child is in positive calcium balance, whereas an elderly person may be in negative calcium balance. Changes in calcium absorption reflect alterations in dietary calcium intake, intestinal calcium solubility, and vitamin D metabolism.

Generally, as serum calcium increases, its excretion increases. When hypercalcemia is caused by **hyperparathyroidism,** PTH will act on the renal tubule, promoting reabsorption of filtered calcium and thus diminishing the effects of the increased filtration and inhibition of renal tubular reabsorption that are caused by increased serum calcium. Decreasing serum calcium is associated with a reduction in urinary excretion, mainly as a result of decreased amounts of filtered calcium. In **hypoparathyroid** patients, who lack PTH, renal tubular reabsorption of calcium is reduced.

Several hormones directly or indirectly affect calcium homeostasis

Thyroid hormone stimulates osteoclast-mediated resorption of bone. Adrenal and gonadal steroids, particularly estrogens in women and testosterone in men, stimulate osteoblast and inhibit osteoclast function. They also decrease renal calcium and phosphate excretion and increase intestinal calcium absorption.

Growth hormone has anabolic effects on bone, promoting the growth of the skeleton. Its effects are mediated by insulin-like growth factors (IGF-1 and IGF-2) acting on cells of the osteoblast lineage. Growth hormone increases the

urinary excretion of calcium and hydroxyproline, whilst decreasing the urinary excretion of phosphate.

Recent data suggest that the central nervous system may also be involved in bone homeostasis. Leptin, an adipokine, which regulates adipose tissue mass and controls appetite (Chapter 22), has been shown to have an inhibitory effect on bone formation. Leptin-deficient animals display high bone mass. However, mutations in the signaling pathway stimulated by leptin have no effect on the bone mass, which suggests that it is a central effect, probably mediated by the sympathetic nervous system.

Clinical box A 60-year-old woman with aches and pains in her bones: osteomalacia

A 60-year-old woman who had become increasingly infirm and housebound was referred to the metabolic outpatient clinic. She had experienced a gradual onset of diffuse aches and pains throughout her skeleton but especially around the hips. She was having difficulty walking, experienced generalized weakness, and recently developed sudden severe pain in her ribs and pelvis. Radiography detected fractured ribs. Adjusted serum calcium was 2.1 mmol/L (8.4 mg/dL; normal range 2.2–2.6 mmol/L, 8.8–10.4 mg/dL), serum phosphate 0.56 mmol/L (1.7 mg/dL; normal range 0.7–1.4 mmol/L, 2.2–4.3 mg/dL), alkaline phosphatase 300 IU/L (normal range 50–260 IU/L) and PTH 12.6 pmol/L (normal range 1.1–6.9 pmol/L, 11–69 pg/mL).

Comment.

In severe forms of osteomalacia, biochemical abnormalities are commonly seen, including low serum adjusted calcium, low serum phosphate, increased alkaline phosphatase and increased $PTH_{(1–84)}$. Patients may have diffuse bone pain or more specific pain related to a fracture, lateral bowing of the lower limbs, and a distinctive waddling gait. Ethnic groups with dark skin are particularly at risk in countries with low–average sunlight, as the majority of vitamin D in the body comes from synthesis by the action of UV light on 7-dehydrocholesterol. This may be exacerbated by traditional body-

covering dress and a diet that is high in phytates (unleavened bread) and low in calcium and vitamin D.

Disorders of calcium metabolism

Hypercalcemia is most commonly caused either by primary hyperparathyroidism or by malignancy

In practice, 90% of cases of hypercalcemia are due to either primary hyperparathyroidism or malignancy; a greater diagnostic challenge is presented when it becomes necessary to differentiate occult malignancy from the less common causes of hypercalcemia. There is a wide individual variation in the development of symptoms and signs of hypercalcemia (Fig. 26.5).

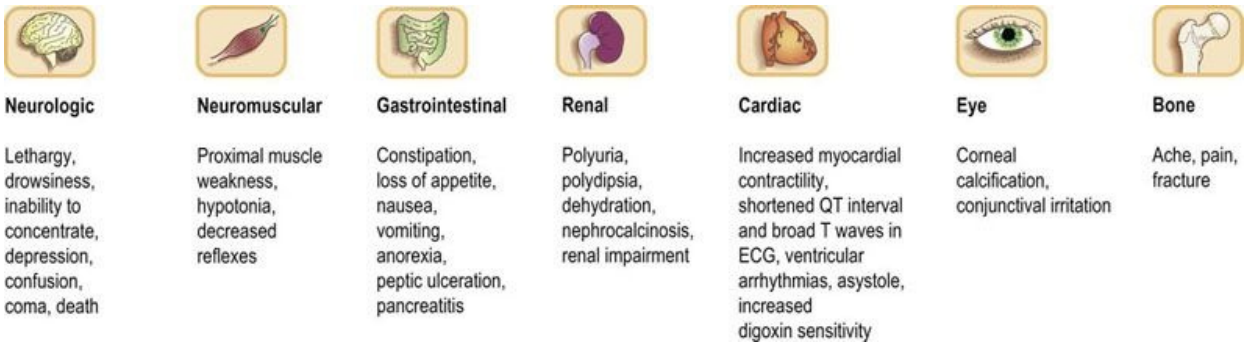


FIG. 26.5 Symptoms and signs of hypercalcemia.
Symptoms are more likely as the serum concentration of calcium increases.

The measurement of PTH has enabled the clinician to discriminate primary hyperparathyroidism from nonparathyroid causes of hypercalcemia (particularly malignancy): an increased or inappropriately detectable intact PTH in the presence of hypercalcemia is observed in primary hyperparathyroidism, whereas a PTH below the limit of detection of the assay is usually observed in nonparathyroid causes of hypercalcemia.

Vitamin D excess or overdose may be obvious from the history, but sometimes only becomes apparent after measurement of the concentrations of ergocalciferol, cholecalciferol and $1,25(OH)_2D_3$.

Primary hyperparathyroidism is relatively common

Primary hyperparathyroidism is a relatively common endocrine disease,

characterized by hypercalcemia associated with an increased or inappropriate concentration of PTH. It has an incidence ranging from 1 in 500 to 1 in 1000 of the population. In 80–85% of patients, a solitary parathyroid gland adenoma is present, and the condition is curable by successful removal of the adenoma. The signs and symptoms of parathyroidism are describe d in the box on p. 347.

Hypercalcemia tends to occur late in the course of malignant disease, and is usually a poor prognostic sign

Current evidence indicates that the most common cause of the hypercalcemia of malignancy (HCM) is the production, by tumors or their metastases, of the parathyroid hormone-related protein (PTHrP) that can circulate in blood and exert its effects on the skeleton and kidneys. Production of PTHrP is common in breast, lung, kidney or other solid tumors, but is more rare in hematologic, gastrointestinal, and head and neck malignancies. The amino-terminal portion of PTHrP possesses PTH-like activity that results in hypercalcemia, hypophosphatemia, phosphaturia, increased renal calcium reabsorption, and osteoclast activation.

The second type of hypercalcemia is the result of increased bone resorption by osteoclasts stimulated by factors produced by the primary tumor or, more usually, by metastases, that stimulate osteoclasts by alteration of the RANKL/OPG balance. Mediators such as PTHrP, cytokines, and growth factors (e.g. IL-1, TNF-α, lymphotoxin, and TGF-β) have all been shown to possess osteoclast-stimulating activity that results in significant bone resorption. Production of prostaglandins, especially PGE_2, has been demonstrated in several classes of tumors, particularly the breast cancer. Prostaglandins stimulate osteoclastic bone resorption.

Advanced concept box

Parathyroid hormone-related protein (PTHrP)

PTHrP is synthesized as three isoforms containing 139, 141, and 173 amino acids, as a result of alternative splicing of RNA. There is amino-terminal sequence homology with PTH: eight of the first 13 amino acids are identical in PTHrP and PTH, three are identical within residues 14–34, and a further three are identical within residues 35–84. Activation of the classic PTH receptor is by the

amino-terminal portion of both PTH and PTHrP, and there is a common α-helical secondary structure in the binding domain of both peptides. As a result of this structural similarity, PTHrP possesses many of the biological actions of PTH.

Excess of vitamin D is toxic

Increasing therapeutic use of potent vitamin D analogues, hydroxylated at position 1 or at positions 1 and 25, make vitamin D toxicity the third most common cause of hypercalcemia.

Hypocalcemia is common in clinical practice

Changes in ionized calcium can result from pH changes in plasma. Alkalemia (Chapter 25) increases protein binding of calcium, decreasing the ionized calcium concentration. The clinical signs of hypocalcemia are, in the main, due to neuromuscular irritability and are more obvious and severe when the onset of hypocalcemia is acute. In some cases, this irritability may be demonstrated by eliciting specific clinical signs. The **Chvostek's sign** is the presence of twitching of the muscles around the mouth (circumoral muscles) in response to tapping the facial nerve anterior to the ear, and the **Trousseau's sign** is the typical contraction of the hand in response to reduced blood flow in the arm induced by inflation of a blood pressure cuff. Numbness, tingling, cramps, tetany and even seizures may occur. Causes of hypocalcemia can be divided into those associated with low $PTH_{(1–84)}$, those in which the decreased serum calcium causes secondary hyperthyroidism, and rare cases in which there is PTH resistance. The most common cause of hypoparathyroidism is a complication of neck surgery.

Pseudohypoparathyroidism syndromes are characterized by hypocalcemia, hyperphosphatemia, and increased concentrations of $PTH_{(1–84)}$.The classic type of pseudohypoparathyroidism is due to end-organ resistance to PTH, caused by a genetic defect resulting in an abnormal regulatory subunit of the G-protein. Confirmation of the diagnosis is made by demonstrating a lack of increase in plasma or urinary cAMP in response to the infusion of PTH. The main causes of hypocalcemia are listed in Table 26.1.

Table 26.1

Causes of hypocalcemia

Hypoparathyroid	Nonparathyroid	PTH resistance
Postoperative	Vitamin D deficiency	Pseudohypoparathyroidism
Idiopathic	Malabsorption	Hypomagnesemia
Neck irradiation	Renal disease	
Anticonvulsant therapy	Vitamin D resistance	
	Hypophosphatemia	

Hypocalcemia may be a result of abnormalities in vitamin D metabolism

Deficiency, acquired or inherited disorders of its metabolism, and vitamin D resistance may occur. The most common causes of vitamin D deficiency are:
Reduced exposure to sunlight, which is common in institutionalized elderly persons and immigrants to Western Europe from the Middle East or Indian subcontinent who wear traditional dress.
Poor dietary intake, *i.e.* diets such as strict vegetarian, which have inadequate vitamin D content and, in the long term, may result in deficiency.
Malabsorption of vitamin D caused by celiac disease, Crohn's disease, pancreatic insufficiency, inadequate bile salt secretion, and nontropical sprue.
Liver disease (deficient 25-hydroxylation), and renal failure (1α-hydroxylase deficiency).

Metabolic bone disease

Osteoporosis is a common age-related disease of bone

Osteoporosis is defined as a significant reduction of bone mineral density compared with age-and sex-matched reference ranges, with an increased susceptibility to fractures. Seventy million people worldwide are at risk of osteoporosis. Osteoporosis is associated with aging. Bone density decreases from a peak achieved by the age of 30 years in men and women, and the rate of bone loss is accelerated in women after loss of estrogen secretion at the menopause. The progressive loss of bone is a result of uncoupling of bone turnover over a prolonged period of time, with a relative **increase of bone resorption** or **decrease in bone formation**. A number of factors have been recognized as contributing to an increased risk of osteoporosis (Fig. 26.6).

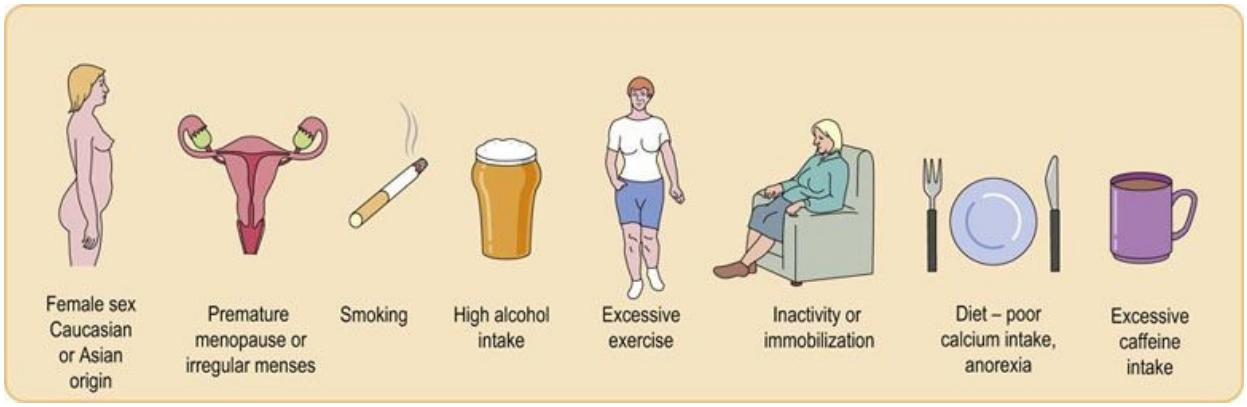


FIG. 26.6 Risk factors and secondary causes of osteoporosis.

Paget's disease of bone is characterized by areas of accelerated bone turnover

Osteoclasts in Paget's disease are large, numerous, and multinucleate; their activity is coupled to increased osteoblast number and activity. A common biochemical abnormality in this disease is increased serum alkaline phosphatase, indicating increased osteoblast activity. Increased collagen breakdown by osteoclasts results in a high serum and urine concentration of hydroxyproline,

pyridinolines, and collagen telopeptides. The bisphosphonates (see Fig. 26.7) have significant antiosteoclastic activity, and are the drugs of first choice for treating Paget's disease.

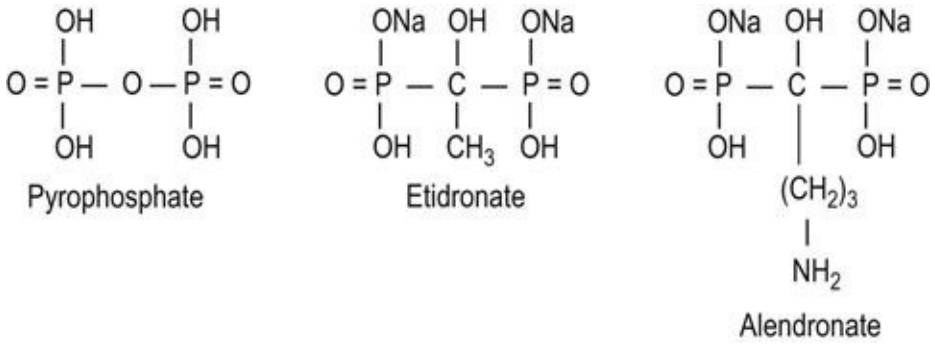


FIG. 26.7 Structural formulae of pyrophosphate and bisphosphonates.
The P-C-P bonds of the biphosphonates can resist enzymic cleavage, and the potency of these drugs is determined by the chemical sequence attached to the carbon molecule.

Clinical box A 62-year-old woman admitted after a fall: osteoporosis

A 62-year-old woman was admitted to hospital after a fall in her bathroom because of sudden onset of severe pain between the shoulder blades. Radiography detected a wedge fracture of two thoracic vertebrae with reduced bone density. A bone density assessment (the dual energy X-ray absorptiometry (DEXA) scan) showed severely reduced density in femur and spine. She had experienced the menopause after a hysterectomy at age 41 years but had been unable to tolerate hormone replacement therapy (HRT). Biochemical investigations were all within normal limits.

Comment.

Symptoms of osteoporosis develop at a late stage of the disease, and are often caused by the presence of fractures. Hip, vertebral, and wrist fractures are common in patients with osteoporosis.

Summary

■ Bone is a metabolically active tissue that undergoes constant remodeling.
■ The major cells involved in the remodeling process are osteoclasts and osteoblasts.
■ Bone metabolism is closely interrelated with the metabolism of calcium, which also involves the intestine and kidney.
■ Calcium balance is hormonally regulated by parathyroid hormone, vitamin D metabolites and calcitonin.
■ The measurement of calcium in serum is an important test in clinical laboratories, because both hypercalcemia and hypocalcemia lead to clinical symptoms.
■ Hypocalcemia is common in clinical practice.
■ The main causes of hypercalcemia are primary hyperparathyroidism, malignancy and vitamin D excess.
■ Osteoporosis, a decrease in bone density leading to bone fractures, is a major health problem.

Active learning

1. Describe the RANK–RANKL signaling system in osteoclasts.
2. Discuss factors that regulate osteoblast function.
3. Describe forms of calcium present in plasma. Which of them is biologically active?
4. Discuss the feedback mechanisms that maintain plasma calcium concentration.

Further reading

Boyle, WJ, Simonet, WS, Lacey, DL. Osteoclast differentiation and activation. *Nature*. 2003; 423:337–341.

Buckley, KA, Fraser, WD. Receptor activator for nuclear factor kappaB ligand and osteoprotegrin: regulators of bone physiology and immune responses/potential therapeutic agents and biochemical markers. *Ann Clin Biochem*. 2002; 39:551–556.

Datta, HK, Ng, WF, Walker, JA, *et al.* The cell biology of bone metabolism. *J Clin Pathol*. 2008; 61:577–587.

Fraser, WD. Paget's disease of bone. *CPD Rheumatology*. 2001; 2:8–13.

Fraser, WD. Hyperparathyroidism. *Lancet*. 2009; 374:145–158.

Harada, S, Rodan, GA. Control of osteoblast function and regulation of bone mass. *Nature*. 2003; 423:349–355.

Rachner, TD, Khosla, S, Hofbauer, L. Osteoporosis: now and the future. *Lancet*. 2011; 377:1276–1287.

Richards, JB, Rivadeneira, F, Pastinen, TM, *et al.* Bone mineral density, osteoporosis, and osteoporotic fractures: a genome-wide association study. *Lancet*. 2008; 371:1505–1512.

Wharton, B, Bishop, N. Rickets. *Lancet*. 2003; 362:1389–1400.

Websites and downloads

Peacock, M, Calcium metabolism in health and disease. *CJASN*. 2010;5(Suppl 1):523–53. http://cjasn.asnjournals.org/content/5/Supplement_1/S23.long

CHAPTER 27

Complex Carbohydrates

Glycoproteins

†Alan D. Elbein and Koichi Honke

Learning objectives

After reading this chapter you should be able to:

■ Describe the general structures of oligosaccharides (glycans) in various types of glycoproteins.

■ Outline the sequence of reactions involved in biosynthesis and processing of *N*-glycans to produce the various types of oligosaccharide chains.

■ Describe the role of *N*-glycans in protein folding, stability and cell–cell recognition.

■ Explain the importance of *O*-glycans in mucin function.

■ Describe how each of the monosaccharides involved in biosynthesis of *N*- and *O*-glycans is synthesized from glucose and activated for synthesis of glycoconjugates.

■ Distinguish lectins from other types of proteins and describe their role in physiology and pathology.

■ Describe several diseases that are associated with deficiencies in enzymes involved in synthesis, modification or degradation of complex carbohydrates.

Introduction

Glycoconjugates include glycoproteins, proteoglycans and glycolipids

Many mammalian proteins are glycoproteins, *i.e.* they contain sugars covalently linked to specific amino acids in their structure. There are two major types of sugar-containing proteins that occur in animal cells, generally referred to as glycoproteins and proteoglycans. Along with glycolipids, which are presented in the next chapter, all of these compounds are part of the group of sugar-containing macromolecules called glycoconjugates.

Glycoproteins (Fig. 27.1A) have short glycan chains; they can have up to 20 sugars but usually contain between three and 15 sugars. These oligosaccharides are highly branched, they do not have a repeating unit, and they usually contain amino sugars (*N*-acetylglucosamine or *N*-acetylgalactosamine), neutral sugars (D-galactose, D-mannose, L-fucose) and the acidic sugar sialic acid (*N*-acetylneuraminic acid). Glycoproteins generally do not contain uronic acids, acidic sugars that are a major part of the proteoglycans. Glycoproteins usually contain much smaller amounts of carbohydrate than of protein, typically from just a few percent carbohydrate to 10–15% sugar by weight.

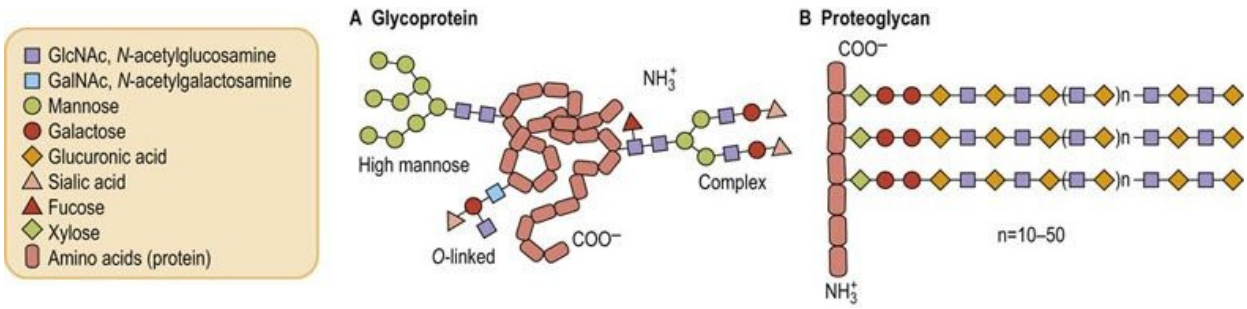


FIG. 27.1 Generalized model of the structure of glycoproteins and proteoglycans.

Proteoglycans (see Fig. 27.1B and Chapter 29) contain as much as 50–60% carbohydrate. In these molecules, the sugar chains are long, unbranched polymers that may contain hundreds of monosaccharides. These saccharide chains have a repeating disaccharide unit, generally made up of a uronic acid and

an amino sugar.

Most proteins in cell surface membranes that function as receptors for hormones or participate in other important membrane-associated processes such as cell–cell interactions are glycoproteins. Many of the membrane proteins of the endoplasmic reticulum or Golgi apparatus, as well as those proteins that are secreted from cells, including serum and mucous proteins, are also glycoproteins. In fact, glycosylation is the major enzymatic modification of proteins in the body. Addition of sugars to a protein can occur either at the same time, and location, as protein synthesis is occurring in the endoplasmic reticulum, *i.e.* co-translationally, or following the completion of protein synthesis and after the protein has been transported to the Golgi apparatus, *i.e.* posttranslationally. The functions of the carbohydrate chains of the resulting glycoproteins are diverse (Table 27.1).

Table 27.1

Function of carbohydrates on glycoproteins

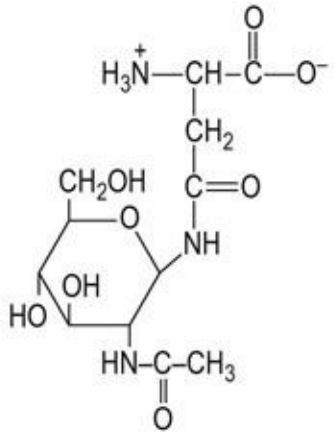
- Assist in protein folding to correct conformation
- Enhance protein solubility
- Stabilize the protein against denaturation
- Protect the protein from proteolytic degradation
- Target the protein to specific subcellular locations
- Serve as recognition signals for carbohydrate-binding proteins (lectins)

Structures and linkages

Sugars are attached to specific amino acids in proteins

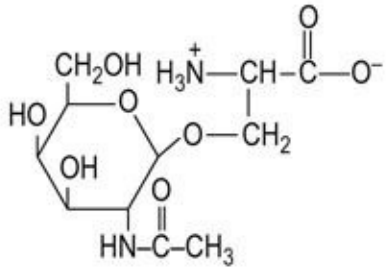
Sugars can be attached to protein in either **N-glycosidic** linkages or **O-glycosidic** linkages. *N*-linked oligosaccharides (*N*-glycans) are widespread in nature and are characteristic of membrane and secretory proteins. The attachment of these oligosaccharides to protein involves a glycosylamine linkage between an *N*-acetylglucosamine (GlcNAc) residue and the amide nitrogen of an asparagine residue (Fig. 27.2A). The asparagine that serves as an acceptor of this oligosaccharide must be in the **consensus sequence** Asn-X-Ser (Thr) in order to be recognized as the acceptor by the oligosaccharide transferring enzyme (see below). However, not all asparagine residues, even those in this consensus sequence, become glycosylated, indicating that other factors such as protein conformation or other properties of the protein may be involved.

A



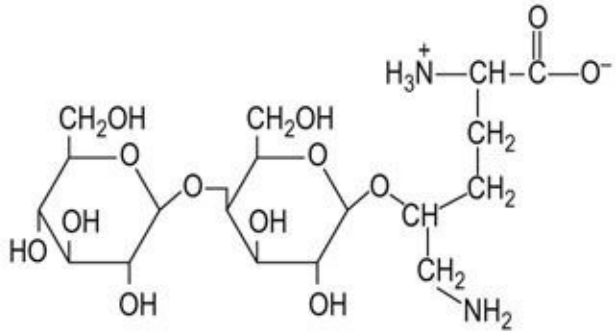
GlcNAc–asparagine
(*N*-linked oligosaccharides)

B



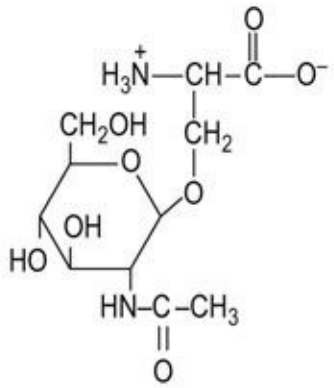
GalNAc–serine
(*O*-linked oligosaccharides)

C



Glc-Gal–hydroxylysine
(*O*-linked collagens)

D



GlcNAc–serine
(*O*-linked cytoplasmic proteins)

FIG. 27.2 Various linkages of sugars to amino acids in glycoproteins. GlcNAc, *N*-acetylglucosamine.

O-linked oligosaccharides (*O*-glycans) are most commonly found in proteins of mucous fluids, but also occur frequently on the same membrane and secretory proteins that contain the *N*-glycans. The *O*-glycans typically contain three or more sugars in linear or branched chains, attached to protein by a glycosidic linkage between an *N*-acetylgalactosamine (GalNAc) residue and the hydroxyl group of either a serine or threonine residue on the protein (Fig. 27.2B; see Fig. 27.4). There does not appear to be a consensus sequence of amino acids for *O*-linked glycosylation, so it is not known why some serines (or threonines) become glycosylated but others do not.

A glucosyl-galactose disaccharide is frequently linked to the hydroxyl group of hydroxylysine residues in the fibrous protein, collagen (Fig. 27.2C). **Hydroxylysine** is an uncommon amino acid, found only in collagens and proteins with collagenous domains. Hydroxylysine is not directly incorporated into protein as such, but is produced by posttranslational hydroxylation of lysine residues. Lysine hydroxylase requires vitamin C as a cofactor, which is why vitamin C is frequently used to expedite wound healing. Collagen is first synthesized in the cell as a precursor form called procollagen. Procollagens are usually synthesized as *N*-linked glycoproteins, but the *N*-glycan is removed as part of that peptide that is cleaved from procollagen during its maturation to collagen. Only the *O*-linked disaccharides remain on the mature collagen molecule. The less glycosylated collagens tend to form ordered, fibrous structures, such as occur in tendons, while the more heavily glycosylated collagens are found in meshwork structures, such as basement membranes in the vascular wall and renal glomerulus (see Chapter 29).

A single GlcNAc is attached to the hydroxyl group of serine or threonine residues on a number of cytoplasmic and nuclear proteins (Fig. 27.2D). This *O*-linked GlcNAc (*O*-GlcNAc) is linked to specific serine and threonine residues that become phosphorylated by protein kinases during hormonal stimulation or other signaling events. The enzyme that adds the GlcNAc is widespread, but how it is controlled is still not clear. There is a second enzyme that removes the GlcNAc from the serine and threonine residues, similar to the contrasting regulatory roles of protein kinases and phosphatases. The GlcNAc modification (**O-GlcNAcylation**) may represent a mechanism that allows the cells to block phosphorylation of specific serine and threonine residues on selected proteins,

while allowing others to still be phosphorylated. Then, the GlcNAc can be removed under appropriate conditions to allow phosphorylation. The donor substrate of *O*-GlcNAcylation is UDP-GlcNAc (see below), which is derived from glucose. *O*-GlcNAcylation is dependent on the intracellular UDP-GlcNAc level, which is correlated with extracellular glucose concentration. *O*-GlcNAcylation of proteins involved in the insulin signaling pathway brings about insulin resistance in skeletal muscle, adipose tissue and pancreatic β-cells, which causes type 2 diabetes. *O*-GlcNAcylation also regulates transcription factors as well as the proteasome, which is involved in protein turnover (Chapter 34).

A novel class of *O*-linked glycans in which mannose is linked to serine or threonine residues is found on the muscle and nerve-specific protein dystroglycan. A typical *O*-mannose glycan consists of GlcNAc, galactose and sialic acid, which are attached to the core mannose in this order. *O*-mannose glycans serve as linkers between the intracellular cytoskeleton and the extracellular matrix to maintain myocyte function. In fact, deficiency in the biosynthesis of *O*-mannose glycans causes muscular dystrophy.

One other amino acid that can serve as a site for glycosylation is tyrosine. The only example of this linkage is in the protein glycogenin, found at the core of glycogen (see Chapter 13). **Glycogenin** is a self-glucosylating protein that initially attaches a glucose to the hydroxyl group of one of its tyrosine residues. The protein then adds a number of other glucoses to the protein-linked glucose to make an oligosaccharide which serves as the acceptor for glycogen synthase.

N-glycans have either ‘high-mannose’ or ‘complex’ structures built on a common core

Although there are a large number of different carbohydrate structures produced by living cells, most of the oligosaccharides on glycoconjugates have many sugars and glycosidic linkages in common. All *N*-glycans have oligosaccharide chains that are branched structures, having a common core of three mannose residues and two GlcNAc residues (Fig. 27.3A and B), but differing considerably beyond the core region to give high-mannose and complex types of chains. The reason for this similarity in structure is that the high-mannose type of *N*-linked oligosaccharide is the biosynthetic precursor for all other *N*-glycans. As indicated below (see Fig. 27.11), the oligosaccharide is initially assembled on a carrier lipid in the endoplasmic reticulum as a high-mannose structure, then transferred to protein. It may remain as a high mannose structure, especially in

lower organisms, but in animals the oligosaccharide undergoes a number of processing steps in the endoplasmic reticulum and Golgi apparatus that involve removal of some mannoses and addition of other sugars. As a result, beyond the core region, the oligosaccharides give rise to a vast array of structures that are referred to as high-mannose (Fig. 27.3A), and complex chains (Fig. 27.3B).

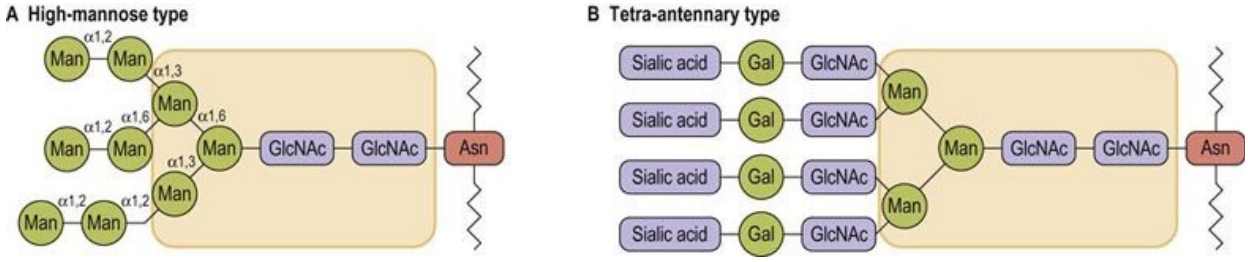


FIG. 27.3 Typical structures of high-mannose and complex, *N*-linked oligosaccharides. The core structure (shaded area) is common to both structures. Asn, asparagine; Gal, galactose; GlcNAc, N-acetylglucosamine Man, mannose.

Complex oligosaccharides are so named because of their more complex sugar composition, including galactose, sialic acid and L-fucose. Complex chains have terminal trisaccharide sequences composed of sialic acid-galactose-GlcNAc attached to each of the branched core mannoses (Fig. 27.3A). L-Fucose may also be found attached to the core GlcNAc (Fig. 27.1A), the terminal galactose (Fig. 28.12) or the penultimate GlcNAc (Fig. 27.9). In common with sialic acid, fucose is usually a terminal sugar on oligosaccharides; that is, no other sugars are attached to it. Some of the complex oligosaccharides have two trisaccharide sequences, one attached to each of the branched core mannoses, and are therefore called biantennary complex chains, whereas others have three (triantennary) or four (tetra-antennary) of the trisaccharide structures (Fig. 27.3B). More than 100 different complex oligosaccharide structures have now been identified on various cell surface proteins, providing great diversity (microheterogeneity) as mediators of cellular recognition and chemical signaling events.

General structures of glycoproteins

A glycoprotein may have a single *N*-glycan chain or it may have several of these types of oligosaccharides. Furthermore, the *N*-glycans may all have identical structures or they may be quite different in structure. For example, the influenza

virus coat glycoproteins, hemagglutinin and neuraminidase, are both glycoproteins which usually have seven *N*-glycan chains, of which five are biantennary complex chains and two are high-mannose structures. Thus, a range of related structures is commonly found in a single glycoprotein and, in fact, multiple different structures may also be found at a single site on a glycoprotein. **Microheterogeneity** of oligosaccharide structures results from incomplete processing on some chains during their biosynthesis (see Fig. 27.12). As a result, some of the oligosaccharides on a glycoprotein may be complete complex chains whereas others may be only partially processed. Such diversity in the processing of *N*-glycans is controlled by many factors. The protein structure near the *N*-glycosylation site can have some influence on the processing. In general, *N*-glycans exposed on the surface of a folded protein are vulnerable to the processing enzymes, whereas those shielded within the structure of the protein are inaccessible to the enzymes.

Many *N*-linked glycoproteins also contain *O*-glycans of the type shown in Figure 27.4. The number of *O*-glycans varies considerably depending on the protein and its function. For example, the low-density lipoprotein (LDL) receptor is present in plasma membranes of smooth muscle cells and fibroblasts, and functions to bind and endocytose circulating LDL, as a source of dietary cholesterol. The LDL receptor has two *N*-glycans that are located near the LDL-binding domain, and a cluster of *O*-glycans near the membrane-spanning region. As shown in Figure 27.5, this receptor has a membrane-spanning region of hydrophobic amino acids, an extended region of amino acids on the external side of the plasma membrane that contains the cluster of *O*-glycans, and a functional domain that is involved in binding LDL (see Chapter 18). Although the two *N*-glycans are near the functional domain, they do not play a role in binding LDL. Instead, their function appears to be in helping the protein to fold into the proper conformation in the endoplasmic reticulum so that it can be translocated to the Golgi apparatus. The negatively charged *O*-glycans, each having a sialic acid, are believed to function to keep the protein in an extended state, and to prevent it from folding back on itself.

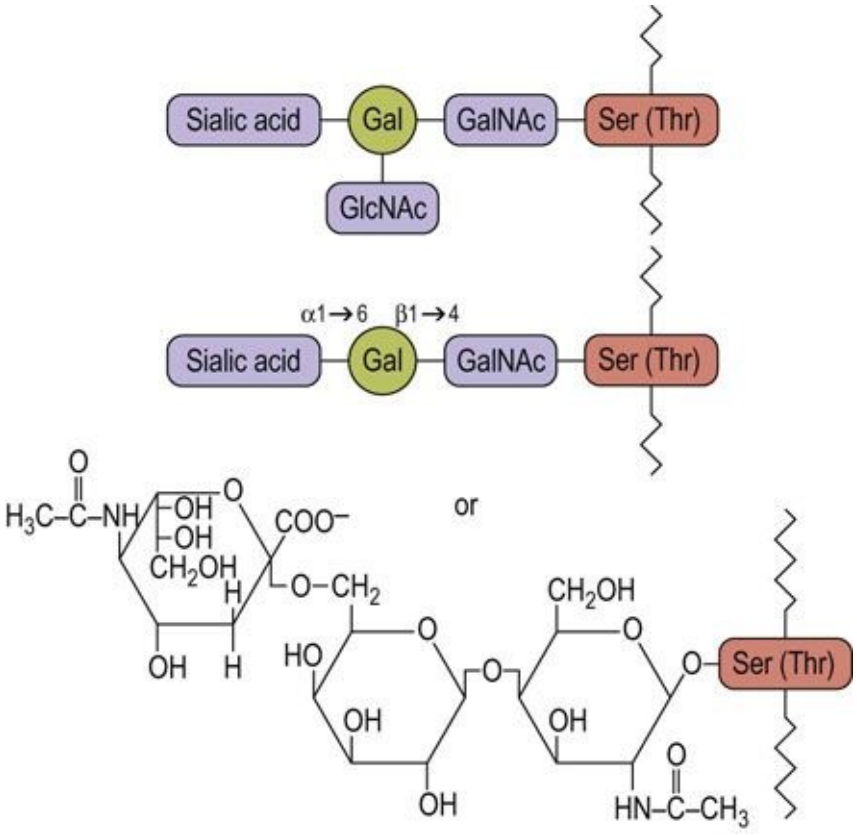


FIG. 27.4 Typical structures of *O*-linked oligosaccharides.

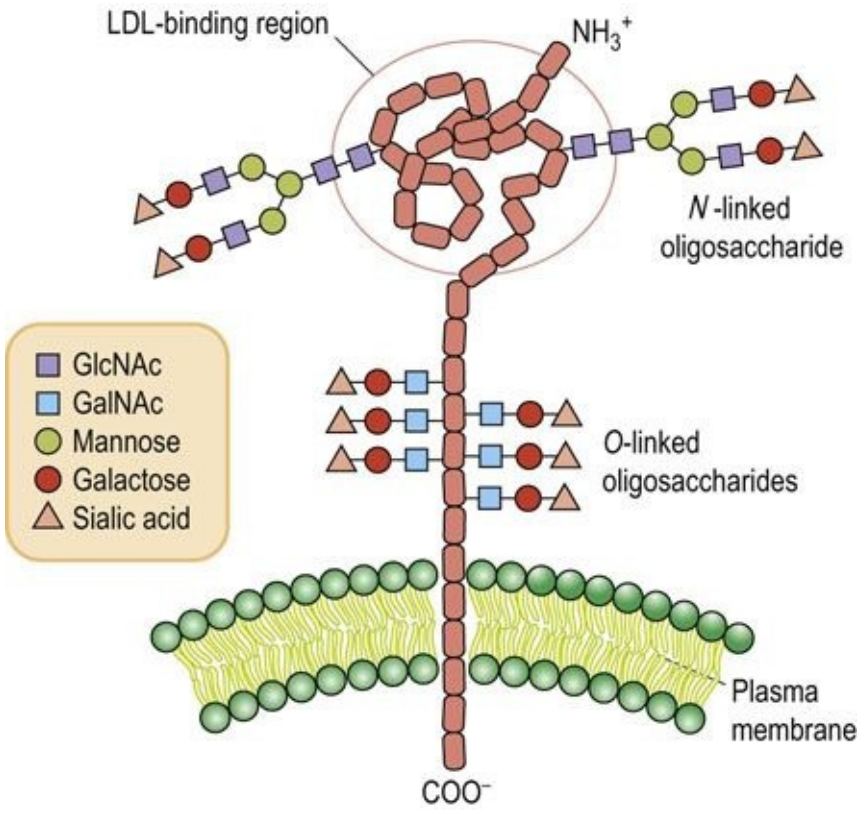


FIG. 27.5 Model of the low-density lipoprotein (LDL) receptor.
(See also Chapter 18 and compare Fig. 18.2.)

Advanced concept box The structure of *N*-glycans depends on the enzyme complement of the cell

The final structure of the *N*-glycan chain of a given glycoprotein is not coded in the genes for the proteins, but depends on the enzyme complement of the cell making that oligosaccharide. All cells appear to have the necessary enzymes to produce the lipid-linked saccharide precursor of the *N*-linked high-mannose chains, and can therefore glycosylate any membrane protein that has the appropriate asparagine in the right sequence and protein conformation. However, the glycosyltransferases and glycosidases involved in processing the oligosaccharide to its final complex structure are not so widely distributed, and a given glycosyltransferase may be present in one type of cell but not in

another. For example, one cell type may have the GlcNAc transferase (GlcNAc T-IV or V) necessary to attach a second GlcNAc on the 2-linked α-mannoses to make a triantennary or tetra-antennary chain, whereas another cell may not have these GlcNAc transferases. Such a cell will only make biantennary chains. Enveloped viruses, such as the influenza virus or HIV, are examples of this phenomenon, since their *N*-glycan structures reflect that of the cell in which they are grown; viruses use the cellular machinery to make all of their structures, and therefore their glycoproteins will have carbohydrate structures characteristic of the infected cell. For the virus this is beneficial since their proteins will not be recognized as foreign proteins and will escape immune surveillance. In addition, it allows the virus to attach to host cell receptors and fuse with host cell membranes by interacting with host lectins. In the biotechnology industry, this means that, although a given protein will have the identical amino acid sequence regardless of cell type, it will have different oligosaccharide structures, depending on the cell in which it is expressed. These differences in carbohydrate structure may affect the conformation and functional properties of the protein and limit its use for protein or enzyme replacement therapy. In fact, many cells used to express 'human' proteins are bioengineered to contain the complement of enzymes needed to properly glycosylate the target protein.

Structure–function relationships in mucin glycoproteins

Mucins are glycoproteins that are secreted by epithelial cells lining the respiratory, gastrointestinal and genitourinary tracts. These proteins are very large in size with subunits having molecular weights of over one million daltons, and having as much as 80% of their weight as carbohydrate. **Mucins are uniquely designed for their function**, with about one-third of the amino acids being serines or threonines, and most of these being substituted with an *O*-

glycan. Because most of these oligosaccharides carry a negatively charged sialic acid and these negative charges are in close proximity, they repel each other and prevent the protein from folding, causing it to remain in an extended state. Thus, the protein solution is highly viscous, forming a protective barrier on the epithelial surface, providing lubrication between surfaces and facilitating transport processes, such as the movement of food through the gastrointestinal system. There is a wide range of complex linear and branching oligosaccharide structures on mucins, including blood group antigens (see Chapter 28). Some oligosaccharides participate in interaction and binding to various bacterial cell surfaces. This property may play a significant role in bacterial sequestration and elimination, limiting colonization and infection.

Interconversions of dietary sugars

Cells can use glucose to make all the other sugars they need

Humans have a dietary requirement for some essential fatty acids and amino acids and vitamins (Chapter 11), but all the sugars that they need to make glycoconjugates can be synthesized from blood sugar, *i.e.* D-glucose. Figure 27.6 presents an overview of the sequence of reactions involved in interconversion of sugars in mammalian cells. All these sugar interconversion reactions involve sugar phosphates or sugar nucleotides.

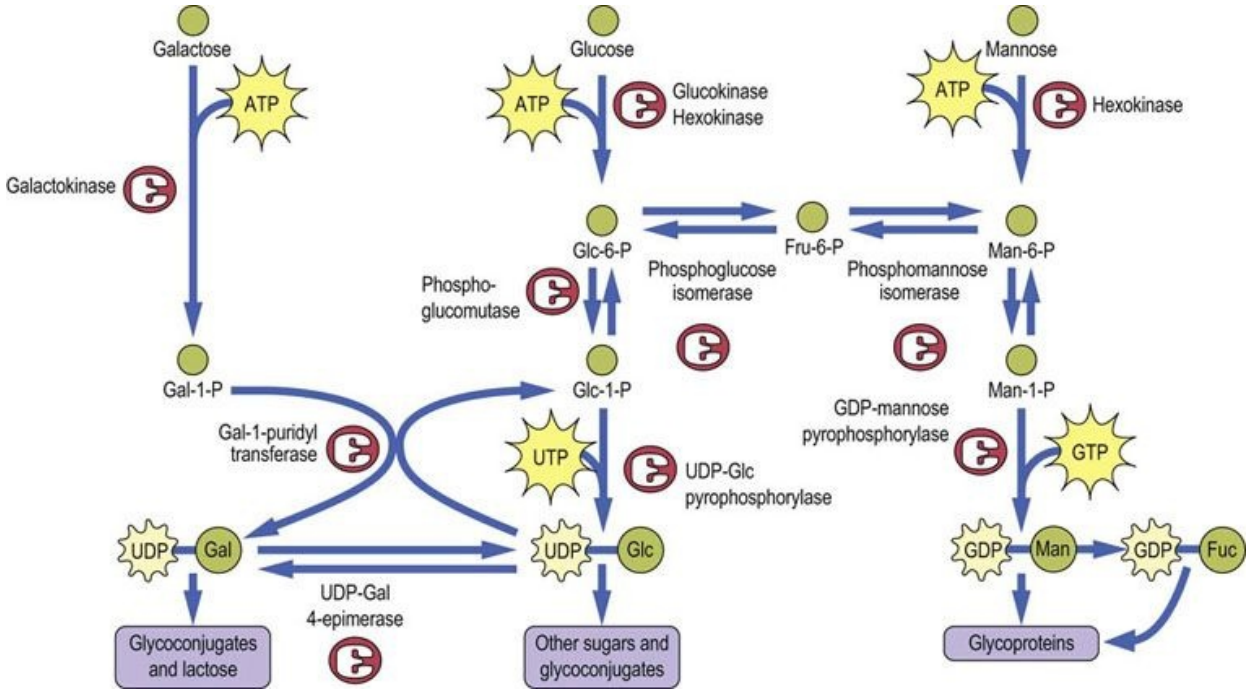


FIG. 27.6 Interconversions of glucose, mannose, galactose, and their nucleotide sugars.
Fuc, fucose; Gal-1-P, galactose-1-phosphate; Glc-1-P, glucose-1-phosphate; Man-1-P, mannose-1-phosphate.

Formation of galactose, mannose and fucose from glucose

Glucose is phosphorylated by hexokinase (or glucokinase in liver) as it enters the cell. Glucose-6-phosphate (Glc-6-P) can then be converted by a mutase (phosphoglucose mutase) to form Glc-1-P, which reacts with uridine triphosphate (UTP) to form UDP-Glc, catalyzed by the enzyme UDP-Glc pyrophosphorylase (Fig. 27.6). This enzyme is a pyrophosphorylase, named for the reverse reaction in which the phosphate of Glc-1-P acts to cleave the pyrophosphate bond of UTP to form UDP-Glc and pyrophosphate (PPi); cleavage of the high energy bond of PPi by pyrophosphatase provides the driving force for the reaction. This is the same pathway used in liver and muscle for incorporation of glucose into glycogen. UDP-Glc is epimerized to UDP-galactose (UDP-Gal) by UDP-Gal 4-epimerase, providing UDP-Gal for synthesis of glycoconjugates.

Glc-6-P may also be converted to fructose-6-phosphate (Fru-6-P) by the glycolytic enzyme phosphoglucose isomerase, and Fru-6-P may be isomerized to mannose-6-phosphate (Man-6-P) by phosphomannose isomerase. The Man-6-P is then converted by a mutase (phosphomannose mutase) to Man-1-P, which condenses with GTP to form GDP-Man. GDP-Man is the form of mannose that is used in the formation of *N*-linked oligosaccharides and is also the precursor for the formation of activated fucose, GDP-L-fucose. Mannose is present in our diet in small amounts but it is not an essential dietary component, since it can be readily produced from glucose. However, dietary mannose can be phosphorylated by hexokinase to Man-6-P, and then enters metabolism through phosphomannose isomerase.

Metabolism of galactose

Although normal animal cells can make all the galactose they need from glucose, galactose is still an important component of our diet because it is one of the sugars that make up the milk disaccharide, lactose. The pathway of galactose metabolism requires three enzymes (see Fig. 27.6). Dietary galactose is transported to the liver where it is phosphorylated by a specific kinase, galactokinase, that attaches phosphate to the hydroxyl group on carbon-1, rather than carbon-6, to form galactose-1-phosphate (Gal-1-P). Humans lack a UDP-Gal pyrophosphorylase, so the conversion of Gal-1-P to Glc-1-P involves the participation of a sugar nucleotide, UDP-Glc. The enzyme Gal-1-P uridyltransferase catalyzes an exchange between UDP-Glc and Gal-1-P to form UDP-Gal and Glc-1-P (see Fig. 27.6). The UDP-Gal is used for synthesis of

glycoconjugates, while the Glc-1-P can be converted to Glc-6-P by phosphoglucomutase, and thus the original galactose molecule enters glycolysis.

UDP-Glc is present in only micromolar concentrations in cells so that its availability for galactose metabolism would be quickly exhausted were it not for the presence of a third enzyme, UDP-Gal-4-epimerase. This enzyme catalyzes the equilibrium between UDP-Glc and UDP-Gal, providing a constant source of UDP-Glc during galactose metabolism. The reactions catalyzed by (1) **galactokinase**, (2) **Gal-1-P uridyltransferase**, and (3) **UDP-Gal 4-epimerase** are summarized below, illustrating the roundabout way by which galactose enters mainstream metabolic pathways.

(1) Gal + ATP → Gal-1-P + ADP

(2) Gal-1-P + UDP-Glc → Glc-1-P + UDP-Gal

(3) UDP-Gal ⇄ UDP-Glc

Net: Gal + ATP → Glc-1-P + ADP

Metabolism of fructose

Fructose is a component of sucrose and also accounts for about half the sugar in high fructose corn syrup

Fructose can be metabolized by two pathways in cells as shown in Figure 27.7. It can be phosphorylated by hexokinase, an enzyme that is present in all cells; however, hexokinase has a strong preference for glucose as a substrate, and glucose, which is present at a concentration of about 5 mmol/L in blood, is a strong competitive inhibitor of the phosphorylation of fructose. The primary pathway of fructose metabolism, which is especially important in liver after a meal, involves the enzyme fructokinase. This enzyme is a very specific kinase that phosphorylates fructose at carbon-1 (rather than the 6-position like hexokinase) to give fructose-1-phosphate (Fru-1-P). The liver aldolase is called **aldolase B**, and it is different in substrate specificity from the muscle aldolase A, because aldolase B can cleave both Fru-1-P and Fru-1,6-P_2, whereas aldolase A will only cleave Fru-1,6-P_2. Thus, in liver, the products of fructose cleavage by aldolase B are dihydroxyacetone phosphate and glyceraldehyde (not glyceraldehyde-3-P).The glyceraldehyde must then be phosphorylated by triose

kinase in order to be metabolized in glycolysis.

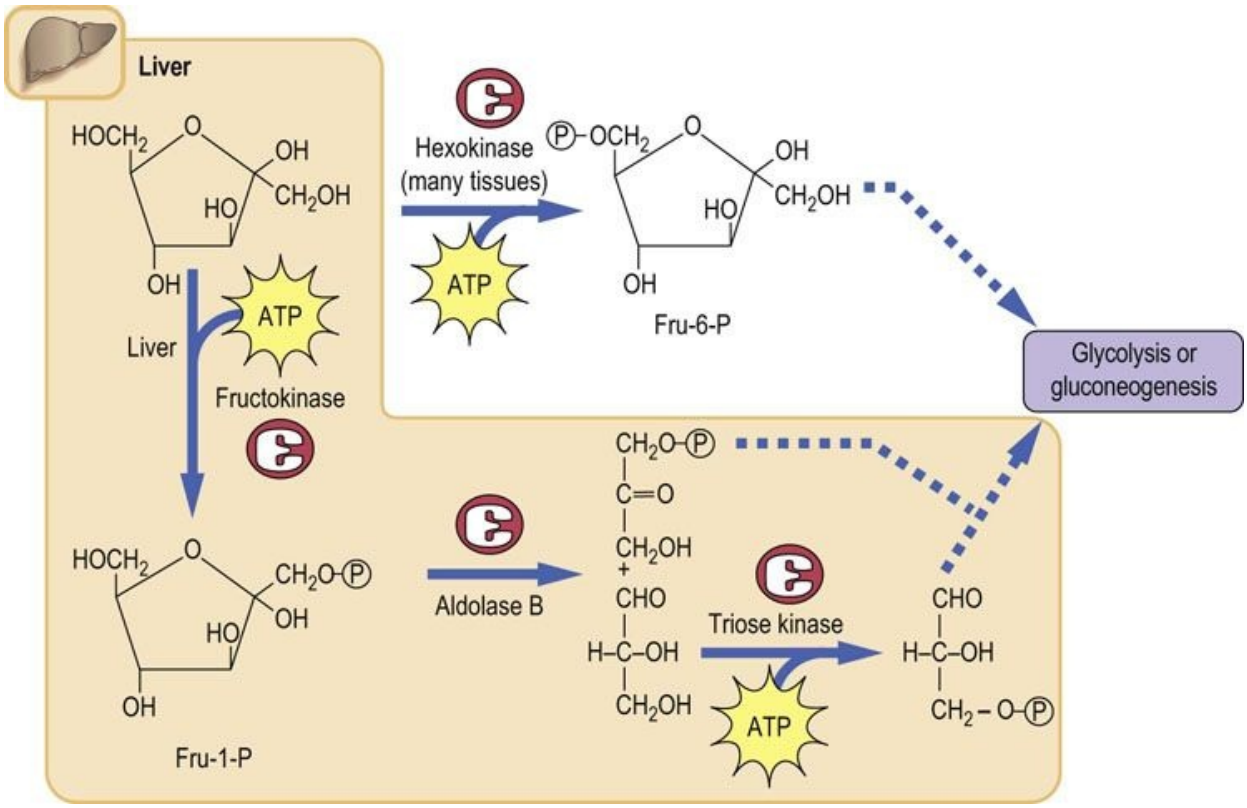


FIG. 27.7 Metabolism of fructose by fructokinase or hexokinase.

It should be noted that in the liver, fructose enters glycolysis at the level of the triose phosphate intermediates, rather than as Fru-6-P as in muscle. Thus, in liver, the ingested fructose is not subject to regulation by the usual control points for regulatory enzymes, hexokinase and phosphofructokinase. By circumventing these two rate-limiting steps, fructose provides a rapid source of energy in both aerobic and anaerobic cells. This is part of the rationale behind the development of high-fructose drinks such as Gatorade®. The significance of the fructokinase, as opposed to the hexokinase, pathway of fructose metabolism is indicated by the pathology of hereditary fructose intolerance (see Clinical Box below).

Advanced concept box

Biosynthesis of lactose

Lactose synthase and α-lactalbumin.

Lactose (galactosyl-β1,4-glucose) is synthesized from UDP-Gal and glucose in mammary glands during lactation. Lactose synthase is formed by the binding of **α-lactalbumin** to the galactosyl transferase that normally participates in biosynthesis of *N*-linked glycoproteins. α-Lactalbumin, which is expressed only in the mammary glands during lactation, converts galactosyl transferase to lactose synthase by lowering the enzyme's K_m for glucose by about three orders of magnitude, from 1 mol/L to 1 mmol/L, leading to preferential synthesis of lactose, rather than galactosylation of glycoproteins. α-Lactalbumin is the only known example of a **'specifier' protein** that alters the substrate specificity of an enzyme.

Clinical box Galactosemia: a baby who developed jaundice after breast-feeding

An apparently normal newborn baby began to vomit and develop diarrhea after breast-feeding. These problems, together with dehydration, continued for several days, when the child began to refuse food and developed jaundice indicative of liver damage, followed by hepatomegaly and then lens opacifications (cataracts). Measurement of glucose in the blood by a glucose oxidase assay (Chapter 3) indicated that the concentration of glucose was low, consistent with failure to absorb foods. However, glucose measured by a colorimetric method that measures total reducing sugar (i.e. any sugar that is capable of reducing copper) indicated that the concentration of sugar was quite high in both blood and urine. The reducing sugar that accumulated was eventually identified as galactose, indicating an abnormality in galactose metabolism known as galactosemia. This finding was consistent with the observation that, when milk was removed from the diet and replaced with an infant formula containing sucrose rather than

lactose, the vomiting and diarrhea stopped, and hepatic function gradually improved.

Comment.

The accumulation of galactose in the blood most often is the result of a deficiency in the enzyme Gal-1-P uridyl transferase (classic form of galactosemia), which prevents the conversion of galactose to glucose and leads to the accumulation of Gal and Gal-1-P in tissues. The accumulated Gal-1-P interferes with phosphate and glucose metabolism, leading to widespread tissue damage, organ failure and mental retardation. In addition, accumulation of galactose in tissues results in galactose conversion, through the **polyol pathway**, to galactitol, which in the lens results in osmotic stress and development of cataracts (compare diabetic cataracts, Chapter 21). Another form of galactosemia is caused by galactokinase deficiency, but in this case Gal-1-P does not accumulate and complications are milder.

Clinical box Hereditary fructose intolerence: a child who developed hypoglycemia after eating fruit

A child was brought into the emergency room suffering from nausea, vomiting and symptoms of hypoglycemia along with sweating, dizziness, and trembling. The parents indicated that these attacks occurred shortly after eating fruits (fructose) or candy (sucrose). As a result of these symptoms, the child was developing a strong aversion to fruits so the mother was providing a large supplementation of multivitamin preparations. The child was below normal weight, but he had not exhibited any of the above unusual symptoms during the period of time when he was breast-feeding. A series of clinical tests demonstrated some cirrhosis of the liver, and a normal glucose tolerance test. However, reducing substances were detected in the urine and these reducing

substances did not react in the glucose oxidase test (i.e. they were not due to glucose). A **fructose tolerance test** was ordered using 3 g fructose/m^2 of body surface area, given intravenously in a single and rapid push. Within 30 minutes, the child displayed symptoms of hypoglycemia. Blood glucose analysis confirmed this and revealed that the hypoglycemia was greatest after 60–90 min. Fructose concentrations reached a maximum (3.3 mmol/L) after 15 min and gradually decreased to zero by 3 h. Plasma phosphate concentration fell by 50% and tests for the enzymes alanine aminotransferase and aspartate aminotransferase indicated that they were elevated after about 90 min. The urine was also positive for fructose.

Comment.

The results of a fructose tolerance test demonstrate the accumulation of fructose and its derivatives in blood and urine. The elevation of liver enzymes, alanine and aspartate aminotransferase, as well as jaundice and other symptoms indicate liver damage and suggest that Fru-1-P affects metabolism in a manner similar to that of Gal-1-P in galactosemia.

Other pathways of sugar nucleotide metabolism

UDP-GlcUA

UDP-Glc is the precursor of a number of other sugars, such as glucuronic acid, xylose and galactose, which are required for proteoglycan and/or glycoprotein synthesis. The reactions that lead to the formation of these other sugars are outlined in Figures 27.6 and 27.8. A two-step oxidation by the enzyme UDP-Glc dehydrogenase leads to the formation of the activated form of glucuronic acid (UDP-GlcUA) (Fig. 27.8). This sugar nucleotide is the donor of glucuronic acid, both for the formation of proteoglycans (see Chapter 29) and for the detoxification and conjugation reactions that occur in the liver to remove bilirubin, drugs and xenobiotics (see Chapter 30). UDP-GlcUA is also the precursor of UDP-xylose, a pentose sugar nucleotide (Fig. 27.8). UDP-GlcUA undergoes a decarboxylation reaction that removes carbon-6 to form UDP-xylose, the activated form of xylose. Xylose is the linkage sugar between protein and glycan in proteoglycans (Figure 27.1B and Chapter 29). Xylose is also present on many plant glycoproteins, as part of their *N*-linked oligosaccharides, and is partly responsible for allergic reactions to peanut and nut proteins.

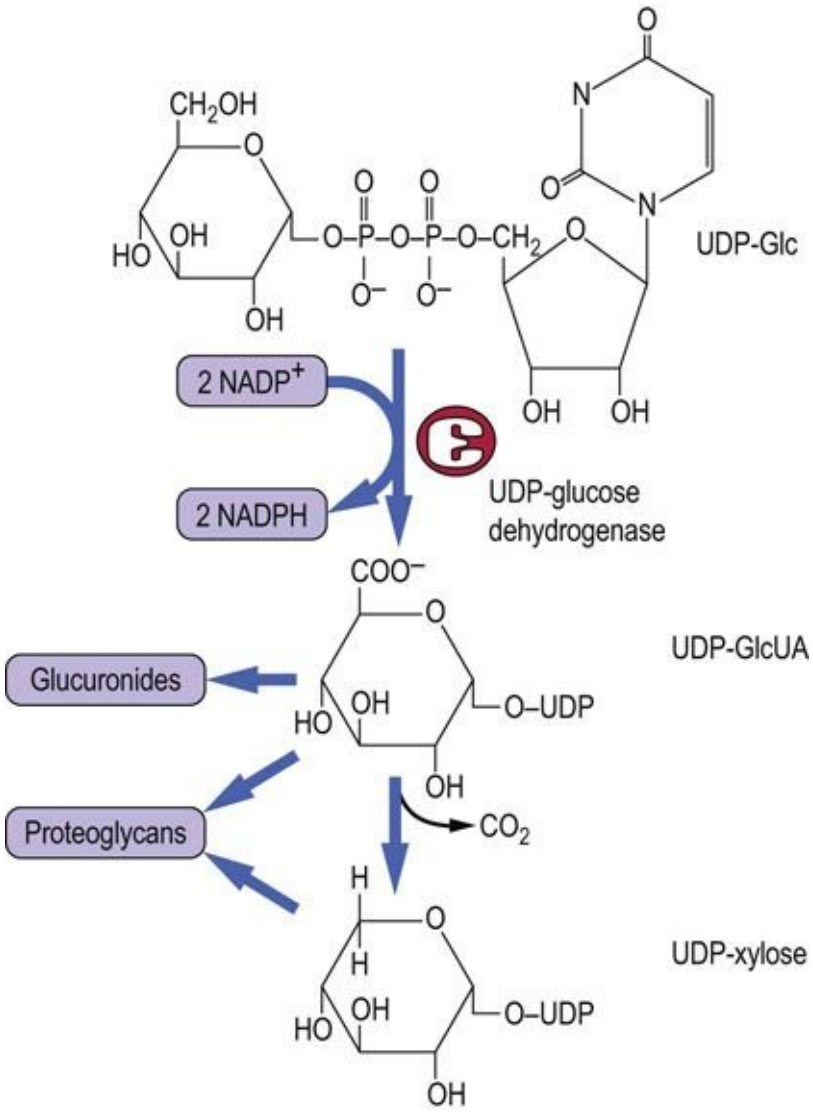


FIG. 27.8 Conversion of UDP-Glc to UDP-glucuronic acid (UDP-GlcUA) and UDP-xylose.
Note that oxidation of UDP-Glc is a two-step reaction, from alcohol to aldehyde, then to an acid. Both reactions are catalyzed by UDP-Glc dehydrogenase.

GDP-Man and GDP-Fuc

Guanosine diphospate-mannose (GDP-Man) is the donor substrate for most mannosyltransferases. As shown in Figure 27.6, it is produced from Man-6-P and is also the precursor to GDP-L-fucose (GDP-Fuc), which is the donor substrate for all fucosyltranferases. Fucose is a 6-deoxyhexose that is an important sugar participating in many recognition reactions in biological events,

such as inflammatory response (Fig. 27.9). The conversion of GDP-Man to GDP-Fuc involves a complex series of oxidative and reductive steps, as well as epimerizations. Deficiency in the GDP-Fuc transporter that translocates GDP-Fuc from the cytosol into the Golgi lumen is associated with a defective inflammatory response and increased susceptibility to infection (Leukocyte adhesion deficiency II: LADII). Resulting depletion of GDP-Fuc in the Golgi lumen blocks the biosynthesis of the **leukocyte recognition signal, sialyl Lewis-X** structure (see Fig. 27.9).

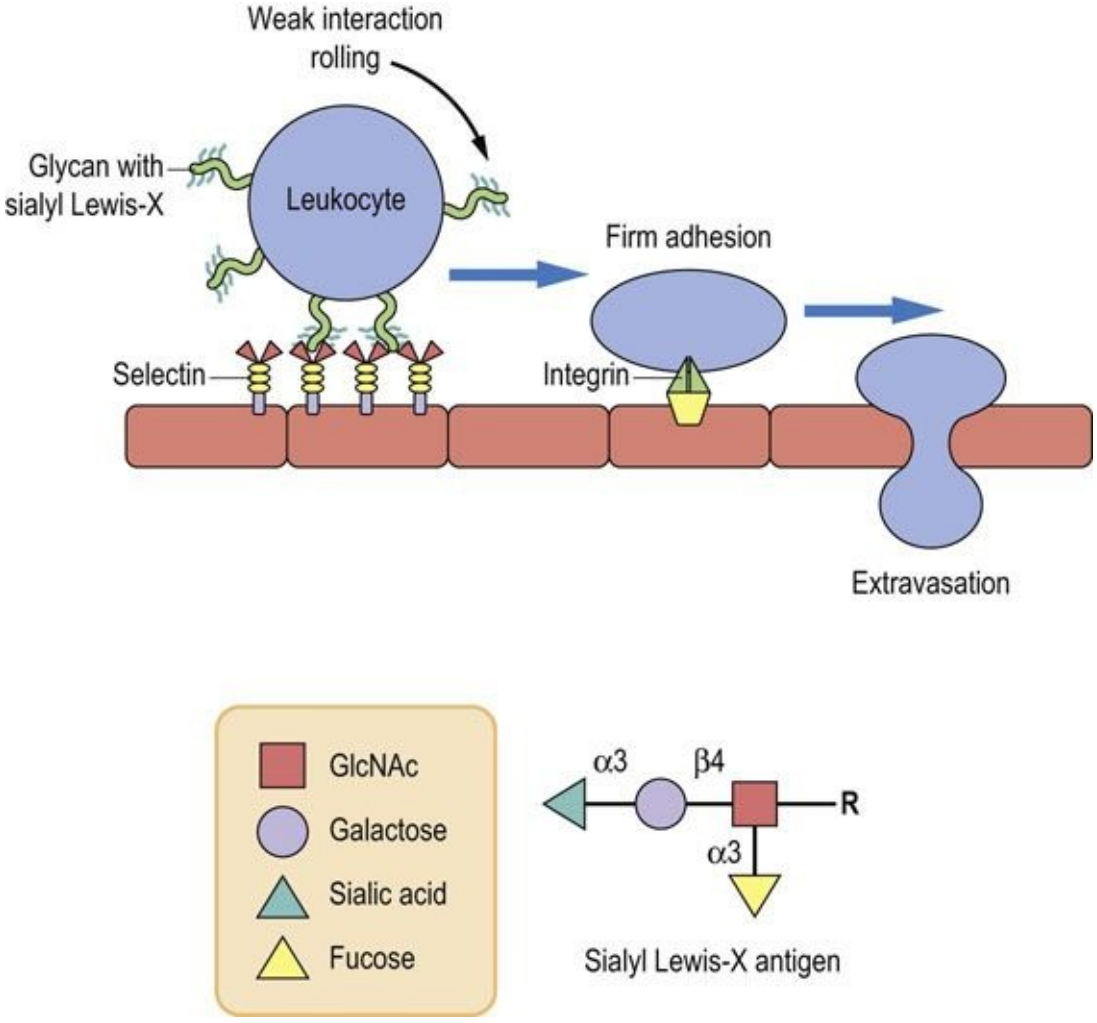


FIG. 27.9 Carbohydrate-dependent cell–cell interactions in inflammation. Sialyl Lewis-X, a tetrasaccharide antigen that forms part of the membrane structure of leukocytes, is recognized by a carbohydrate-binding protein, selectin (Sel), on the surface of endothelial cells. The sialyl Lewis-X-selectin interaction mediates the initial, weak binding that results in rolling of the leukocytes along the endothelial monolayer. It

facilitates the firm adhesion mediated by the protein–protein interactions leading to extravasation.

Advanced concept box
Carbohydrate-dependent cell–cell interactions

An important example of carbohydrate-dependent cell–cell interactions occurs during inflammation. An injury, or infection, to the vascular endothelial cells elicits an inflammatory response that causes the release of cytokines, such as tumor necrosis factor-α (TNF-α) and interleukin-1 (IL-1), from the damaged tissue. These cytokines attract leukocytes to the site of injury or infection to remove the damaged tissue or the invading organisms. Leukocytes must be able to stop or exit from the blood flow and attach to the injured tissue. They are able to do this because they have a carbohydrate ligand on their surface that is recognized by a **lectin** (carbohydrate-binding protein) that becomes exposed on the surface of the damaged endothelial cells. The carbohydrate ligand is a tetrasaccharide called the sialyl Lewis-X antigen (Fig. 27.9), which is a component of a glycoprotein or glycolipid on the surface of the leukocyte. The sialyl Lewis-X antigen is recognized by lectins, **E-selectin and P-selectin**, that are expressed on the surface of the endothelial cells by the stimulation of cytokines. Figure 27.9 schematically shows sequential events during vascular adhesion and extravasation of leukocytes to inflamed tissue. The sialyl Lewis-X–selectin interaction mediates the initial step of the leukocytes–endothelium interaction, which is described as tethering, followed by rolling of leukocytes along the endothelial cell surface. Although this carbohydrate–protein binding is weak and transient, it. is able to slow down the leukocytes circulating under a strong shear force in blood and promote the firm protein–protein interactions between the integrins and their receptors. Eventually, the leukocytes migrate through the endothelium into the underlying tissue.

On the other hand, interaction of L-selectin expressed on the

lymphocytes with a sialyl Lewis-X-like oligosaccharide expressed on the endothelial cells of the high endothelial venules (HEV) enables lymphocytes circulating in the bloodstream to enter a lymph node by a similar mechanism. This process is called lymphocyte homing.

While these carbohydrate–protein interactions play a critical role in the immune system, they can be dangerous and life threatening under other circumstances. Some cancer cells utilize such a carbohydrate–protein interaction in their metastasis through the bloodstream. There is active research on developing drugs whose structure is similar to carbohydrates (glycomimetic) and will block the vascular adhesion of tumor cells and prevent metastasis.

Comment.

This example of a protein–carbohydrate interaction is just one of many that occur in vivo. Each of these kinds of interactions involves a different lectin, each of which has a specific carbohydrate recognition site. Once the carbohydrate structure is known and the protein-binding site has been mapped, it may be possible for chemists to design compounds that mimic the carbohydrate structure. These synthetic compounds should bind at the carbohydrate-binding site of the lectin and block the natural interaction. One of the difficulties with this approach is that the individual binding interactions are weak and multiple cell–cell contacts are required; these may be difficult to block by small-molecule drugs. A second problem is that synthesis of specific oligosaccharides is difficult and expensive, and that large quantities may have to be injected into the blood for effective therapy.

Amino sugars

Fru-6-P is the precursor of amino sugars

Figure 27.10 shows the pathway of formation of GlcNAc, GalNAc and sialic

acid. The initial reaction involves the transfer of an amino group from the amide nitrogen of glutamine to Fru-6-P to produce glucosamine-6-P (GlcN-6-P). An acetyl group is then transferred from acetyl-CoA to the amino group of the GlcN-6-P to form GlcNAc-6-P, which is converted to its activated form, UDP-GlcNAc, by sequential mutase and pyrophosphorylase reactions. In addition to its role as a GlcNAc donor, UDP-GlcNAc can also be epimerized to UDP-GalNAc. With few exceptions, all amino sugars in glycoconjugates are acetylated; thus they are neutral and do not contribute any ionic charge to the glycoconjugates.

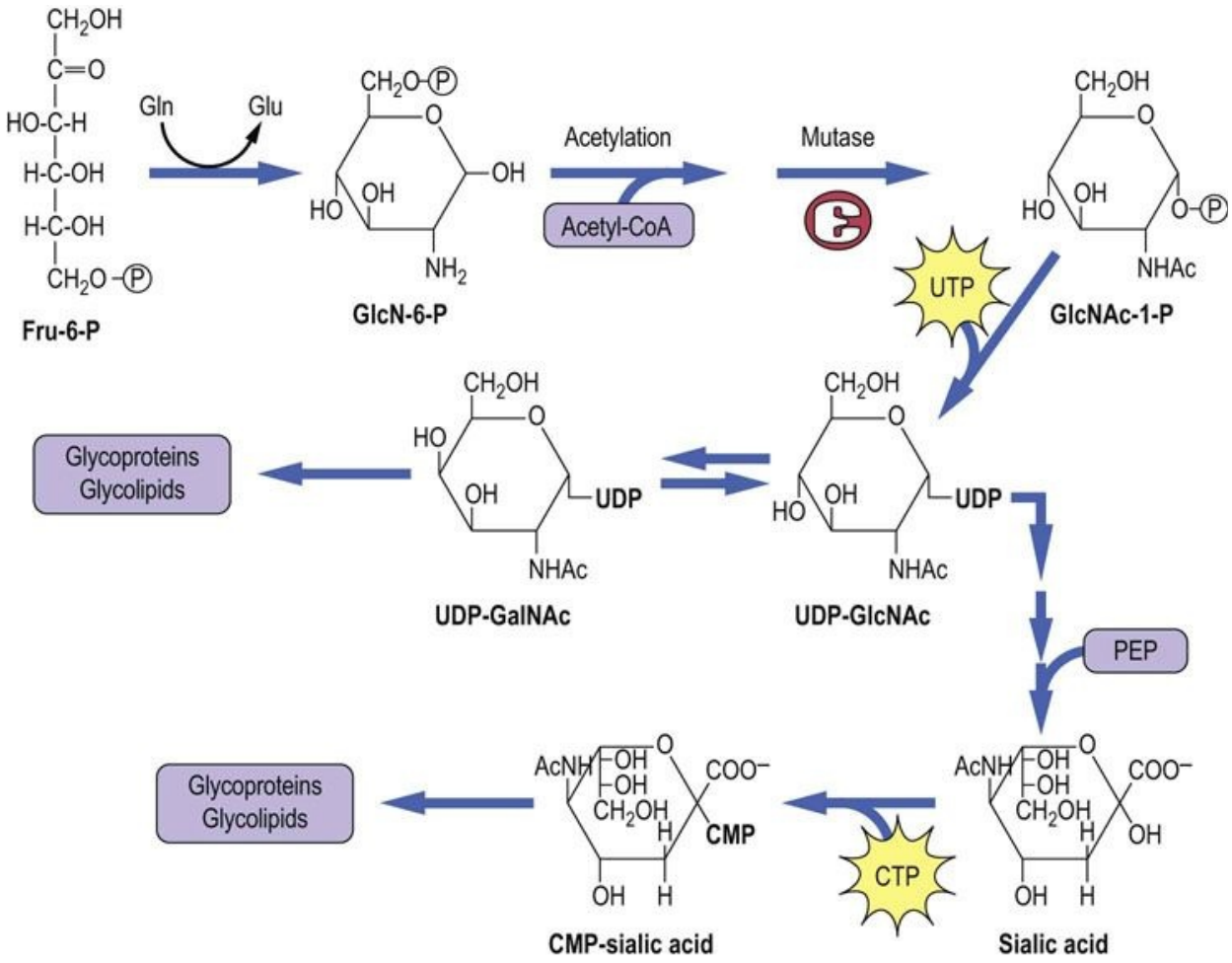


FIG. 27.10 Synthesis of amino sugars and sialic acid.
Acetyl-CoA, acetyl coenzyme A; GlcN-6-P, glucosamine-6-phosphate; GlcNAc-6P, *N*-acetylglucosamine-6-phosphate; GalNAc, *N*-acetylgalactosamine; HNAc, AcHN, acetamide group; PEP, phosphoenolpyruvate.

Sialic acid

UDP-GlcNAc is the precursor of *N*-acetylneuraminic acid (NeuAc), also referred to as sialic acid. Sialic acid, a 9-carbon *N*-acetylamino-ketodeoxyglyconic acid, is produced by the condensation of an amino sugar with **phosphoenolpyruvate** (see Fig. 27.10). Cytidine monophosphate neuraminic acid (CMP-NeuAc) is the activated form of sialic acid and is the sialic acid donor in biosynthetic reactions. CMP-sialic acid is the only nucleoside monophosphate sugar donor in glycoconjugate metabolism.

Biosynthesis of oligosaccharides

N-glycan assembly begins in the endoplasmic reticulum

The pathway of assembly of the *N*-glycans begins with the transfer of two GlcNAc residues to a membrane-bound lipid, called dolichyl phosphate. Mannose and glucose residues are added to build a lipid-linked oligosaccharide intermediate, which is transferred "*en bloc*" to protein in the lumen of the endoplasmic reticulum (Fig. 27.11). **Dolichols** are long-chain polyisoprenol derivatives usually having about 120 carbon atoms (about 22–26 isoprene units) with a phosphate group at one end. They are synthesized in membranes using the same machinery that is used to make cholesterol, but in contrast to cholesterol, the dolichols remain as long straight chains. The length of the chain requires it to snake through the phospholipid bilayer, providing a strong anchor for the growing oligosaccharide chain.

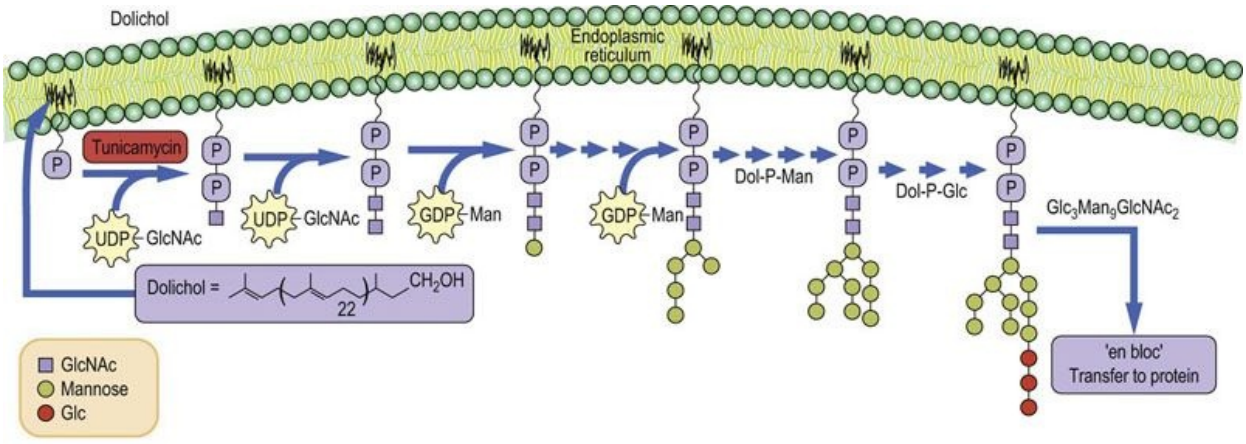


FIG. 27.11 Synthesis of *N*-linked oligosaccharides in the endoplasmic reticulum. Tunicamycin is an inhibitor of the GlcNAc phosphotransferase that the catalyzes the first step in glycan synthesis. GlcNAc, *N*-acetylglucosamine; Dol, dolichol; Man, mannose.

The first sugar to be added to dolichyl-P from UDP-GlcNAc is GlcNAc-1-P by a GlcNAc-1-P transferase to produce dolichyl-P-P-GlcNAc. A second GlcNAc is linked to the first GlcNAc followed by addition of 4–5 mannose residues from GDP-Man. Dolichyl-P-Man and dolichyl-P-Glc serve as glycosyl donors for the remaining mannoses and the three glucoses residues. Each of the

sugars is transferred by a specific glycosyltransferase located in or on the endoplasmic reticulum membrane. The glucoses are not found on any of the *N*-linked oligosaccharides on glycoproteins, but are removed by glucosidases in the endoplasmic reticulum. Why are they added in the first place? They serve two very important functions. First of all, the presence of glucoses on the lipid-linked oligosaccharide has been shown to expedite the transfer of oligosaccharide from lipid to protein – the transferring enzyme (oligosaccharide transferase) has a preference for oligosaccharides that contain three glucoses and transfers those oligosaccharides to protein much faster. Secondly, the glucoses are important in directing protein folding in the endoplasmic reticulum (below).

Intermediate processing continues in the endoplasmic reticulum (ER) and Golgi apparatus

In a series of trimming or pruning reactions (Fig. 27.12), all three glucoses are removed in the ER. The oligosaccharide may then remain as a high-mannose oligosaccharide or it may be further processed to a complex oligosaccharide structure. One or more mannoses may be removed in the ER, and the folded protein is then translocated to the Golgi apparatus where three or four additional mannoses may be removed to leave the core structure of three mannose and two GlcNAc residues. This core oligosaccharide is then elongated in the Golgi apparatus.

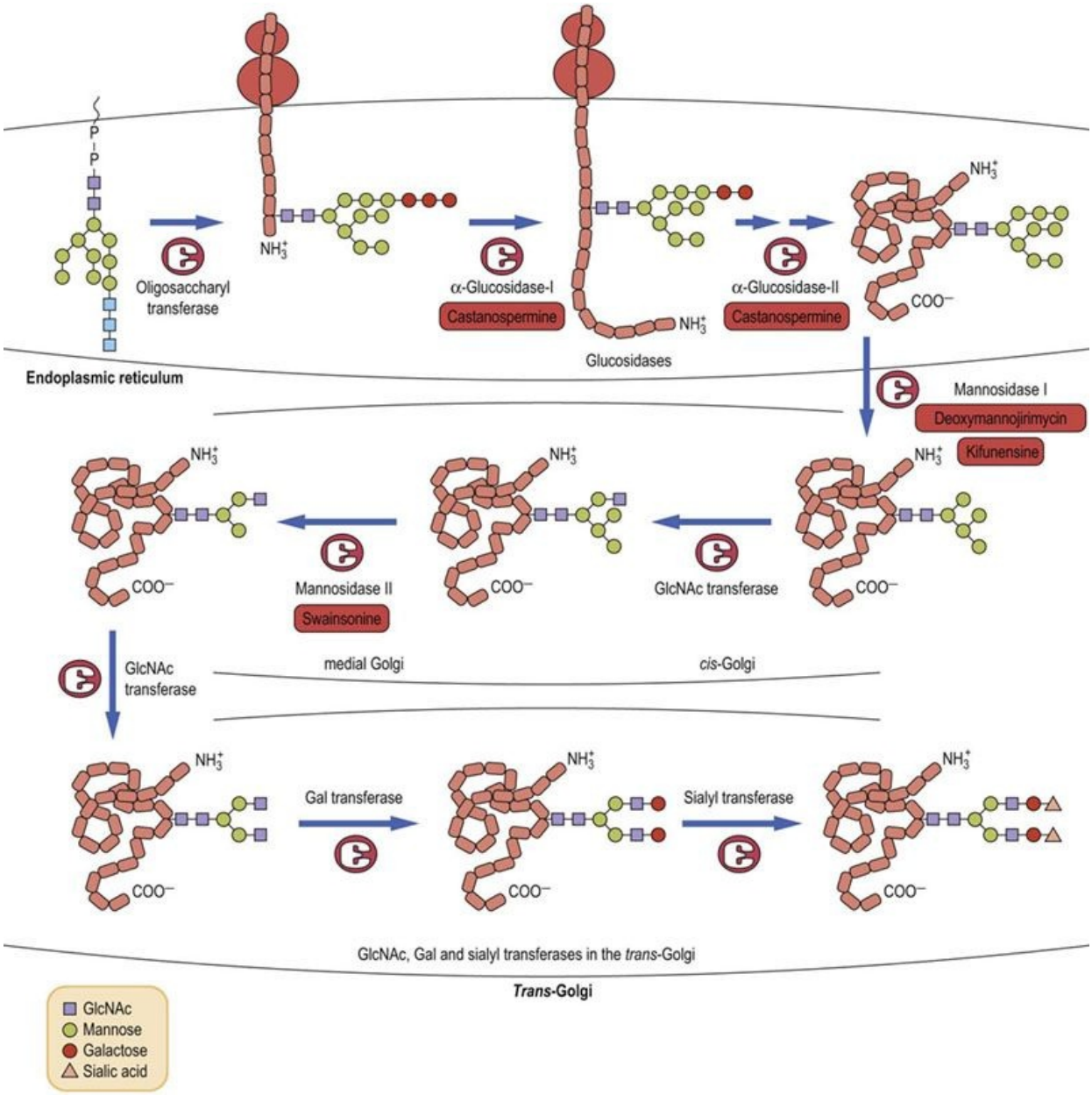


FIG. 27.12 Processing of *N*-linked oligosaccharides from high-mannose to complex forms.
Glycoproteins are transported between the endoplasmic reticulum and Golgi compartments in vesicles. Inhibitors of glycan processing enzymes are shown in red. GlcNAc, *N*-acetylglucosamine.

Final modifications of the N-linked oligosaccharides occur in the Golgi apparatus

After the pruning reactions in the ER and proper folding, the protein with one or

more $Man_8GlcNAc_2$ oligosaccharides is transported to the Golgi apparatus where other modification reactions occur (see Fig. 27.12). Usually in the *cis*-Golgi, several other mannoses are removed by α-mannosidases to leave the core structure. Also in the *cis*-Golgi, GlcNAc residues are added to each of the mannoses. Then the protein enters the *trans*-Golgi fraction where the remaining sugars of the trisaccharide sequences, *i.e.* galactose, sialic acid and fucose, can be added to make a variety of different complex chains. The final structure of the oligosaccharide chains depends on the glycosyltransferases complement of the cell.

O-glycans

Oglycans are synthesized in the Golgi apparatus

In contrast to the biosynthesis of *N*-glycans, the synthesis of the *O*-glycans occurs only in the Golgi apparatus by the stepwise addition of sugars from their sugar nucleotide derivatives to the protein. No lipid intermediates are involved in *O*-glycan formation. Figure 27.13 outlines the stepwise sequence of reactions that are involved in the assembly of an oligosaccharide chain on salivary mucin. In this sequence, GalNAc is first transferred from UDP-GalNAc to serine or threonine residues on the protein by a GalNAc transferase in the Golgi apparatus. The resulting GalNAc-serine-protein serves as the acceptor for galactose and then sialic acid, transferred from their sugar nucleotides (UDP-Gal and CMP-sialic acid) by Golgi galactosyltransferases and sialyltransferases. Other Golgi glycosyltransferases are involved in the stepwise biosynthesis of more complex mucin oligosaccharides and in the synthesis of *O*-glycans in proteoglycans and collagens (see Chapter 29). There are more than 100 glycosyltransferases involved in glycoconjugate synthesis in a typical cell.

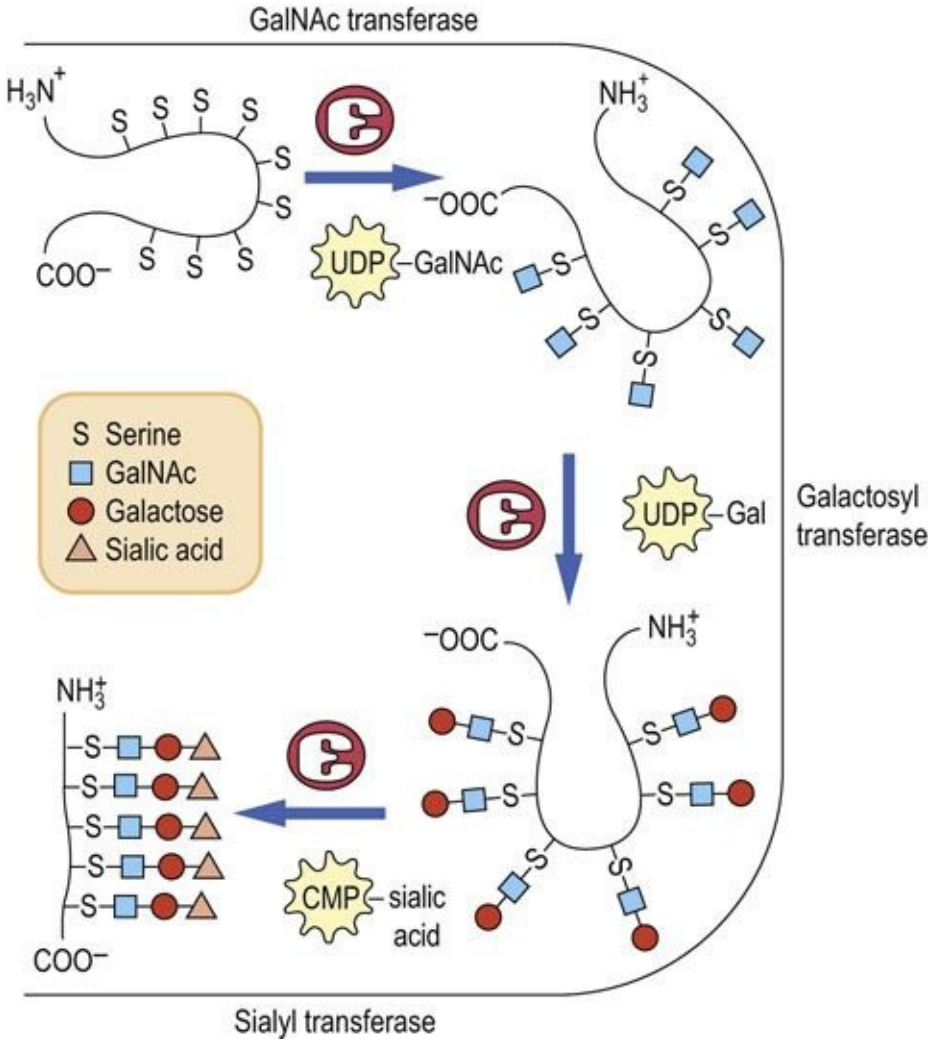


FIG. 27.13 Biosynthesis of *O*-linked oligosaccharides of mucins in the Golgi apparatus. GalNAc, *N*-acetylgalactosamine.

Advanced concept box Inhibitors of glycoprotein biosynthesis

A number of inhibitors of the biosynthesis of *N*-glycans have been identified and these compounds have proven to be valuable reagents for studies on the role of specific carbohydrate structures in glycoprotein function. **Tunicamycin** is a glycoside antibiotic that inhibits the first step in the synthesis of *N*-glycans, *i.e.* the formation of dolichyl-PP-GlcNAc (see Fig. 27.11). Tunicamycin has varied effects on glycoprotein synthesis and on cells, from benign to profound. In some cases, the protein portion of the

glycoprotein is synthesized, but without its carbohydrate it is misfolded, aggregates and is degraded in the cell. Thus, treatment of cells with tunicamycin frequently induces endoplasmic reticulum (ER) stress (see Chapter 34).

Other inhibitors inhibit specific steps in the processing pathway. Many are plant alkaloids that structurally resemble the sugars glucose and mannose, and inhibit the pruning glycosidases (see Fig. 27.12). Castanospermine inhibits the ER glucosidases, whereas kifunensine, deoxymannojirimycin and swainsonine each inhibit a different processing mannosidase. These drugs prevent the formation of complex chains and therefore are useful to evaluate structure–function relationships. Some compounds have been tested against HIV and against some cancers, and have shown positive inhibitory effects. However, they also have adverse effects on enzymes in normal cells and are therefore not usable for drug therapy. With more specific compounds, it may be possible to manipulate glycan structures for therapeutic purposes.

Functions of the oligosaccharide chains of glycoproteins

N-glycans have an important role in protein folding

Resident proteins in the endoplasmic reticulum, known as **chaperones**, assist newly synthesized proteins to fold into their proper conformations. Two of these chaperones, calnexin and calreticulin, bind to unfolded glycoproteins by recognition of high-mannose oligosaccharides that still contain a single glucose remaining on their structure, after the glucosidases have removed two of the three glucoses. Not all of the glycoproteins synthesized in the cell require assistance in folding but for those that do, the rate of folding is greatly accelerated by the chaperones. Incorrectly folded or unfolded proteins do not undergo normal transport to the Golgi apparatus and if they do not fold properly, they precipitate in the endoplasmic reticulum and are subsequently degraded by the ubiquitin–proteasome system in the cytoplasm.

Oligosaccharides containing Man-6-P target lysosomal enzymes to the lysosome

Lysosomes are subcellular organelles involved in the hydrolysis and turnover of many cellular organelles and proteins. They contain a variety of hydrolytic enzymes with acidic pH optima. Most of these lysosomal enzymes are *N*-linked glycoproteins that are synthesized and glycosylated in the endoplasmic reticulum and Golgi apparatus. The sorting of lysosomal enzymes occurs in the *cis*-Golgi. Proteins destined to be transported to the lysosomes contain a cluster of lysine residues that come together as a result of the protein folding into its proper conformation. As shown in Figure 27.14, this cluster of lysine residues serves as a docking site for an enzyme, GlcNAc-1-P transferase, that transfers a GlcNAc-1-P from UDP-GlcNAc to terminal mannose residues on the high-mannose chains of the lysosomal enzymes. A second enzyme, called an uncovering enzyme, then removes the GlcNAc, leaving the phosphate residues still attached to the mannoses on the high-mannose chains. The resulting Man-6-P residues on the high-mannose structure are now recognized by a Golgi protein called a Man-6-P receptor that directs the enzyme to the lysosomes. Thus, the **Man-6-P residues are a targeting signal** used by the cell to sort out those proteins that

are destined to go to the lysosomes, and separate them from other proteins being synthesized in the Golgi apparatus. The Man-6-P receptor is also present on the cell surface, so that even extracellular enzymes that have this signal are endocytosed and transported to the lysosomes.

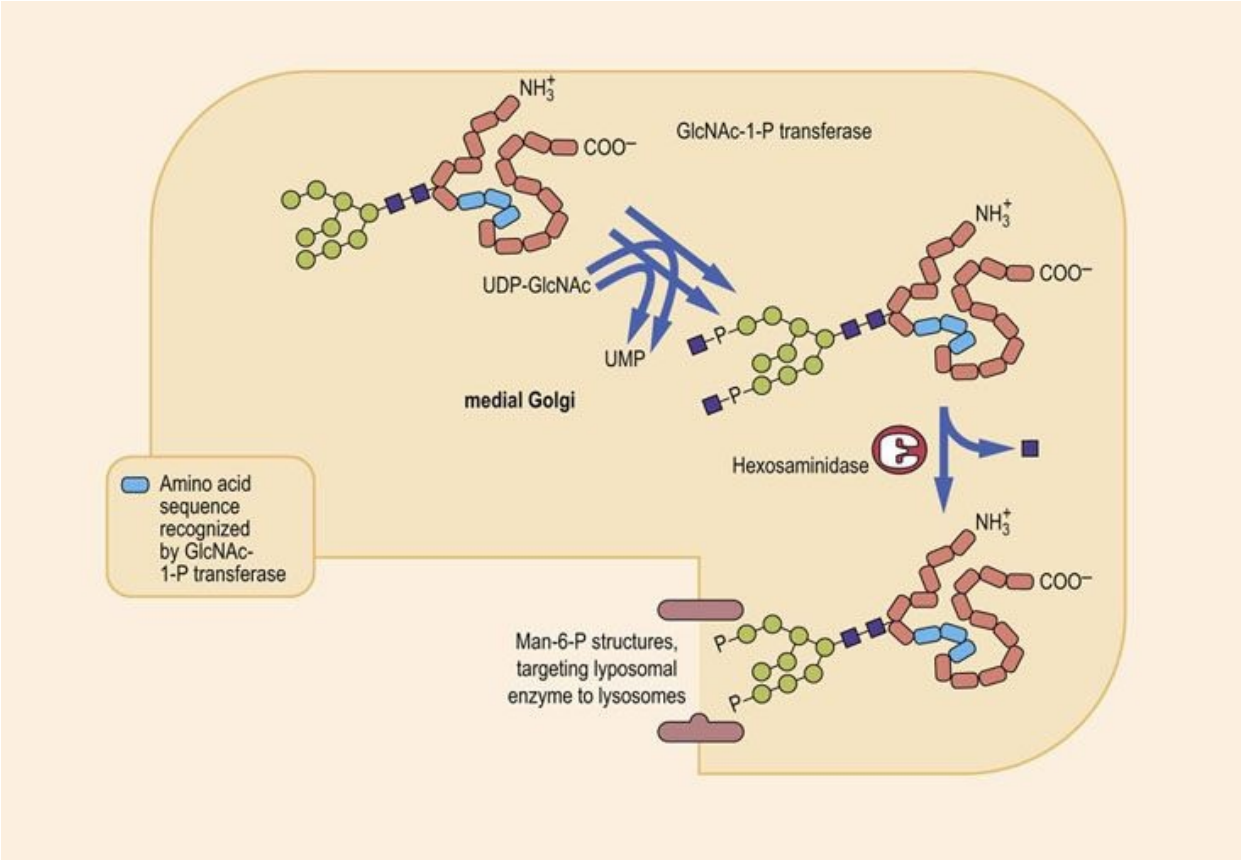


FIG. 27.14 Targeting of lysosomal enzymes to lysosomes.
GlcNAc, *N*-acetylglucosamine; Man, mannose.

The oligosaccharide chains of glycoproteins frequently increase the solubility and stability of proteins

Clinical box Deficiencies in glycoprotein synthesis

The **congenital disorders of glycosylation (CDG)** are a recently described group of rare genetic diseases that affect the biosynthesis

of glycoproteins. All patients show multisystem pathology, with severe involvement of the nervous system. Three distinct classes have been identified thus far and are characterized by a deficiency in the structure of the carbohydrate moiety of serum glycoproteins, lysosomal enzymes or membrane glycoproteins. The diagnosis of the disease is routinely made by electrophoresis of serum transferrin. In CDG, the transferrin contains less sialic acid and therefore the protein migrates more slowly. The decrease in sialic acid results from a defect in biosynthesis of the underlying oligosaccharide structure. While a change in the migration of serum transferrin indicates that the patient is suffering from one of the CDG, it does not identify the specific lesion. That can only be done by either characterizing the structure of the altered oligosaccharide chain(s) to determine what sugars or structures are missing, or doing a profile of key enzymes in the biosynthetic pathways, since the absence of any of these enzymes will affect the final oligosaccharide structure.

Comment.

The basic defects in this group of diseases appears to be in the synthesis or processing of *N*-glycans. However, defects in phosphomannose isomerase and phosphomannose mutase have also been identified as causes of CDG.

Because oligosaccharides are hydrophilic, they increase the solubility of proteins in the aqueous environment. Thus, most of the proteins that are secreted from cells are glycoproteins, including plasma proteins, excepting plasma albumin. These glycoproteins and enzymes generally also have high stability to heat, chemical denaturants, detergents, acids and bases. Enzymatic removal of the carbohydrate from many of these proteins greatly reduces their stability to stress. Indeed, when glycoproteins are synthesized in cells in the presence of glycosylation inhibitors, such as tunicamycin, which prevents the production and therefore the attachment of the *N*-glycan chain, many of these proteins become insoluble and form inclusion bodies in the cells as a result of incorrect folding and/or decreased hydrophilicity.

Sugars are involved in chemical recognition interactions with lectins

Clinical box I-cell disease

I-cell disease (mucolipidosis II) and pseudo-Hurler polydystrophy (mucolipidosis III) are rare inherited diseases that are caused by deficiencies in the machinery that targets lysosomal enzymes to lysosomes. Clinical presentation includes severe psychomotor retardation, coarse facial features and skeletal abnormalities; death usually occurs in the first decade. In cultured fibroblasts taken from patients suffering from mucolipidosis II, newly synthesized lysosomal enzymes are secreted into the extracellular medium rather than being targeted correctly to the lysosomes. Mesenchymal cells, especially fibroblasts, contain numerous membrane-bound vacuoles in the cytoplasm containing fibrillogranular material. These deposits are called inclusion bodies and this is the origin of the name I-cell disease.

Comment.

I-cell disease results from a deficiency in synthesis of the targeting signal, Man-6-P residues on high-mannose oligosaccharides. The mutation is most commonly an absence of GlcNAc-1-P phosphotransferase, but defects in the uncovering enzyme also occur. It is likely that absence of the Man-6-P receptor protein would yield the same phenotype. In I-cell disease, the lysosomes, lacking the full spectrum of hydrolase enzymes, become engorged with indigestible substances.

N-glycans on the mammalian cell surface play critical roles in cell–cell interactions and other recognition processes. One cell may contain on its cell surface a carbohydrate-recognizing protein, known as a lectin, that binds to a specific oligosaccharide structure on the surface of the complementary cell. The interaction between these two chemical interfaces mediates a specific chemical recognition between the cells, and such a process is a key factor in fertilization,

inflammation, infection, development and differentiation.

Carbohydrate–protein interactions are also important in non-self interactions. Many pathogens use this mechanism to recognize their target cells. *E. coli*, for example, and some other Gram-negative enteric bacteria have short hair-like projections called pili on their surfaces. These pili have mannose-binding lectins at their tip that can recognize and bind to high-mannose oligosaccharides on the brush border membranes of intestinal epithelial cells. This interaction allows the bacteria to be retained in the intestine. The influenza virus uses a hemagglutinin protein on its surface to bind to sialic acid residues on glycoproteins and glycolipids on the surfaces of target cells.

Variations in mucin structure appear to have a role in the specificity of fertilization, cell differentiation, development of the immune response, and virus infectivity. Glycoprotein ZP3, which is present on the zona pellucida of the mouse egg, functions as a receptor for sperm during fertilization. Enzymatic removal of *O*-glycans from ZP3 results in loss of sperm receptor activity, whereas removal of the *N*-glycans has no effect on sperm binding. The isolated *O*-glycans obtained from ZP3 also have sperm-binding activity and inhibit sperm–egg interaction and fertilization in vitro. Differences between the *O*-glycan structures of cytotoxic lymphocytes and helper cells involved in the immune response are also believed to be important in mediating cellular interactions during the immune response.

Advanced concept box Toxicity of ricin and other lectins

Lectins are found in a variety of foods, including beans, peanuts and dry cereals. Many plant lectins are toxic to animal cells. In edible plants, these may be less of a problem if the foods are cooked, since the lectins are denatured and therefore susceptible to intestinal proteases. On the other hand, lectins in uncooked plants are very resistant to proteases and can therefore cause serious problems. They bind to cells in the gastrointestinal tract, inhibiting enzyme activities, food digestion and nutrient absorption, and causing gastrointestinal distress and allergic reactions.

Ricin, produced by the castor bean plant, is among the most poisonous proteins known to man. These types of toxic lectins are usually composed of several subunits, one of which is the

carbohydrate-recognizing or -binding site, while the other subunit is an enzyme that can catalytically inactivate ribosomes. Thus a single molecule of this catalytic subunit entering a cell can completely block protein synthesis in that cell. Other toxic lectins include modeccin, abrin and mistletoe lectin I.

Clinical box Changes in sugar composition and/or structure can be diagnostic markers of some types of cancer

Changes in glycosylation of both proteins and lipids have been consistently reported on cell surface carbohydrates of various types of cancer cells, including melanomas, ovarian cancer, and hepatocellular carcinoma. While these changes are not the cause of the disease, they are being evaluated as diagnostic tools for early detection of disease. Increased levels of the enzyme GlcNAc transferase V (the transferase involved in adding a second (branching) GlcNAc residue to a mannose residue to make a triantennary complex chain) is highly expressed in some transformed cells, resulting in increased branching and production of larger *N*-linked oligosaccharides. Changes in *O*-linked oligosaccharides have also been reported: for example, increased levels of sialyl Lewis-X antigen, which is thought to contribute to metastasis. Changes in the amount and sialylation of mucins are also associated with metastasis of lung and colon carcinoma cells and are being studied for their usefulness as diagnostic or prognostic biomarkers.

There is also evidence that changes in the level of fucose on some glycoproteins regulate the biological phenotype of cancer cells, and in fact, fucosylation of the protein α-fetoprotein (AFP-L3) has been used clinically as a marker for hepatocellular carcinoma.

Comment.

The structure and composition of glycoproteins and glycolipids are altered in tumor cells, compared to normal cells. While these changes may not cause the cancer, they may have a significant effect on clinical outcome, *e.g.* if they limit leukocyte infiltration, assist in evading immune surveillance or facilitate metastasis. Analysis of oligosaccharide structures may be useful for early detection and diagnostic purposes and manipulation of oligosaccharide structure may prove useful in treatment of some cancers.

Summary

■ Glycosylation is the major posttranslational modification of tissue proteins.
■ Glycosylation is a multicompartment activity, involving sugar interconversions and activation in the cytosolic compartment, building of complex structures on lipid intermediates in the ER, and glycosylation and pruning reactions in the ER and Golgi apparatus. The outcome is an amazingly diverse range of oligosaccharide structures on proteins.
■ Sugars on glycoconjugates can serve a number of different functions, including:
■ modification of the physical properties of the protein (solubility, stability and/or viscosity);
■ aiding in the folding of the protein;
■ participating in the targeting of the protein to its proper location in the cell;
■ mediating cell–protein and cell–cell recognition during fertilization, development, inflammation and other processes.
■ A number of human diseases involve defects in sugar metabolism, including galactosemia and hereditary fructose intolerance, leukocyte adhesion deficiency, congenital disorders of glycosylation (CDG) and lysosomal storage diseases.

Active learning

1. Why do eukaryotic cells use lipid-linked oligosaccharides as intermediates in synthesis of *N*-glycans but not *O*-glycans?
2. Do animal cells need amino sugars in the diet in order to synthesize complex carbohydrates? If not, why not? Review the use of glucosamine-chondroitin supplements for the treatment of arthritis.
3. Describe the role of carbohydrate-dependent cell–cell interactions during development of the nervous or immune system.

Further reading

Asano, N. Sugar-mimicking glycosidase inhibitors: bioactivity and application. *Cell Mol Life Sci*. 2009; 66:1479–1492.

Etzioni, A. Genetic etiologies of leukocyte adhesion defects. *Curr Opin Immunol*. 2009; 21:481–486.

Kato, K, Kamiya, Y. Structural views of glycoprotein-fate determination in cells. *Glycobiology*. 2007; 17:1031–1044.

Martin, PT. Congenital muscular dystrophies involving the *O*-mannose pathway. *Curr Mol Med*. 2007; 7:417–425.

Schachter, H, Freeze, HH. Glycosylation diseases: Quo vadis? *Biochim Biophys Acta*. 2009; 1792:925–930.

Taylor, ME, Drickmer, K. *Introduction to glycobiology*, ed 2. New York: Oxford University Press; 2006.

Varki, A, Cummings, RD, Esko, JD, *et al*. *Essentials of glycobiology*, ed 2. New York: Cold Spring Harbor Press; 2009.

Zeidan, Q, Hart, GW. The intersections between O-GlcNAcylation and phosphorylation: implications for multiple signaling pathways. *J Cell Sci*. 2010; 123:13–22.

Websites

Congenital disorders of glycosylation. www.familyvillage.wisc.edu/lib_cdgs.htm.

Galactosemia. http://galactosemia.org/.

Glycoprotein biosynthesis. http://www.ccrc.uga.edu/~lwang/bcmb8020/N-glycans-A.pdf.

Glycoprotein topics. www.glycoforum.gr.jp/science/word/glycoprotein/GP_E.html.

Hereditary fructose intolerance. www.bu.edu/aldolase/HFI/.

I-cell disease. http://emedicine.medscape.com/article/945460-overview.

Plant lectins. www.ansci.cornell.edu/plants/toxicagents/lectins.html.

CHAPTER 28

Complex Lipids

†Alan D. Elbein and Koichi Honke

Learning objectives

After reading this chapter you should be able to:

■ Describe how the various glycerol-based phospholipids are synthesized and how they are interconverted.

■ Describe the multiple roles of cytidine nucleotides in activation of intermediates in phospholipid synthesis.

■ Describe the various types of sphingolipids and glycolipids that occur in mammalian cells and their functions.

■ Explain the etiology of lysosomal storage diseases, their pathology, and the rationale for enzyme replacement therapy for treatment of these diseases.

Introduction

Complex lipids encompass the glycerophospholipids, introduced in Chapter 3, and the sphingolipids. These molecules are found mostly in two locations, either embedded in biological membranes or in circulating lipoproteins. The sphingolipids are almost exclusively in cell membranes, primarily the plasma membrane. They carry a wide range of carbohydrate structures which face into the exterior environment and, like glycoproteins, have a range of recognition functions. A major difference between these two classes of lipids is that glycerophospholipids are saponifiable (except plasmalogens), while **sphingolipids** contain no alkali-labile ester bonds. Thus, it was convenient to isolate sphingolipids from tissues by saponification, and then extract the remaining lipids into organic solvent. Once isolated, the characterization of the glycan structure of the sphingolipids was technically challenging. Therefore, the structures were, for a long time, unknown and mysterious, leading to their name: sphinx-like or sphingolipids.

This chapter discusses the structure, biosynthesis, and function of the two major classes of polar lipids: glycerophospholipids and sphingolipids. In preparation for this chapter, it might help to review the structure of phospholipids in Chapter 3.

Synthesis and turnover of glycerophospholipids

Synthesis of glycerophospholipids

There are many species of glycerophospholipids with a distinct composition of polar head groups and hydrophobic acyl groups (see Chapter 3). With regard to the acyl groups, saturated fatty acids are usually esterified at the *sn*-1 position, whereas unsaturated fatty acids are esterified at the *sn*-2 position. Biosynthesis of glycerophospholipids first proceeds by the de novo pathway, and subsequently the originally attached fatty acids in the de novo pathway are replaced with new ones in the remodeling pathway. Through this remodeling pathway, diversity and asymmetry of the acyl groups are generated.

De novo pathway

Phospholipids are in a constant state of synthesis, turnover, and remodeling

The de novo pathway begins with sequential reactions, in which glycerol-3-P is acylated by transfer of two fatty acids from acyl-CoA to produce **phosphatidic acid** (PA) via the intermediate, lysophosphatidic acid (see Fig. 28.1). Then PA is dephosphorylated to **diacylglycerol** (DAG) by a specific cytosolic phosphatase. Alternatively, PA reacts with cytidine triphosphate (CTP) to yield the activated phosphatidic acid, cytidine diphosphate (CDP)-DAG. Phosphatidic acid and DAG are common intermediates in the synthesis of both triglycerides (triacylglycerols) and phospholipids. All animal cells, except for erythrocytes, are able to synthesize phospholipids de novo, whereas triglyceride synthesis occurs mainly in liver, adipose tissue, and intestinal cells. The starting material, **glycerol-3-phosphate**, is formed in most tissues by reduction of the glycolytic intermediate dihydroxyacetone phosphate (DHAP). In liver, kidney and intestine, glycerol-3-P can also be formed directly via phosphorylation of glycerol by a specific kinase. DHAP may also be acylated by addition of a fatty acid to the 1-hydroxyl group; this intermediate is then reduced and acylated to PA.

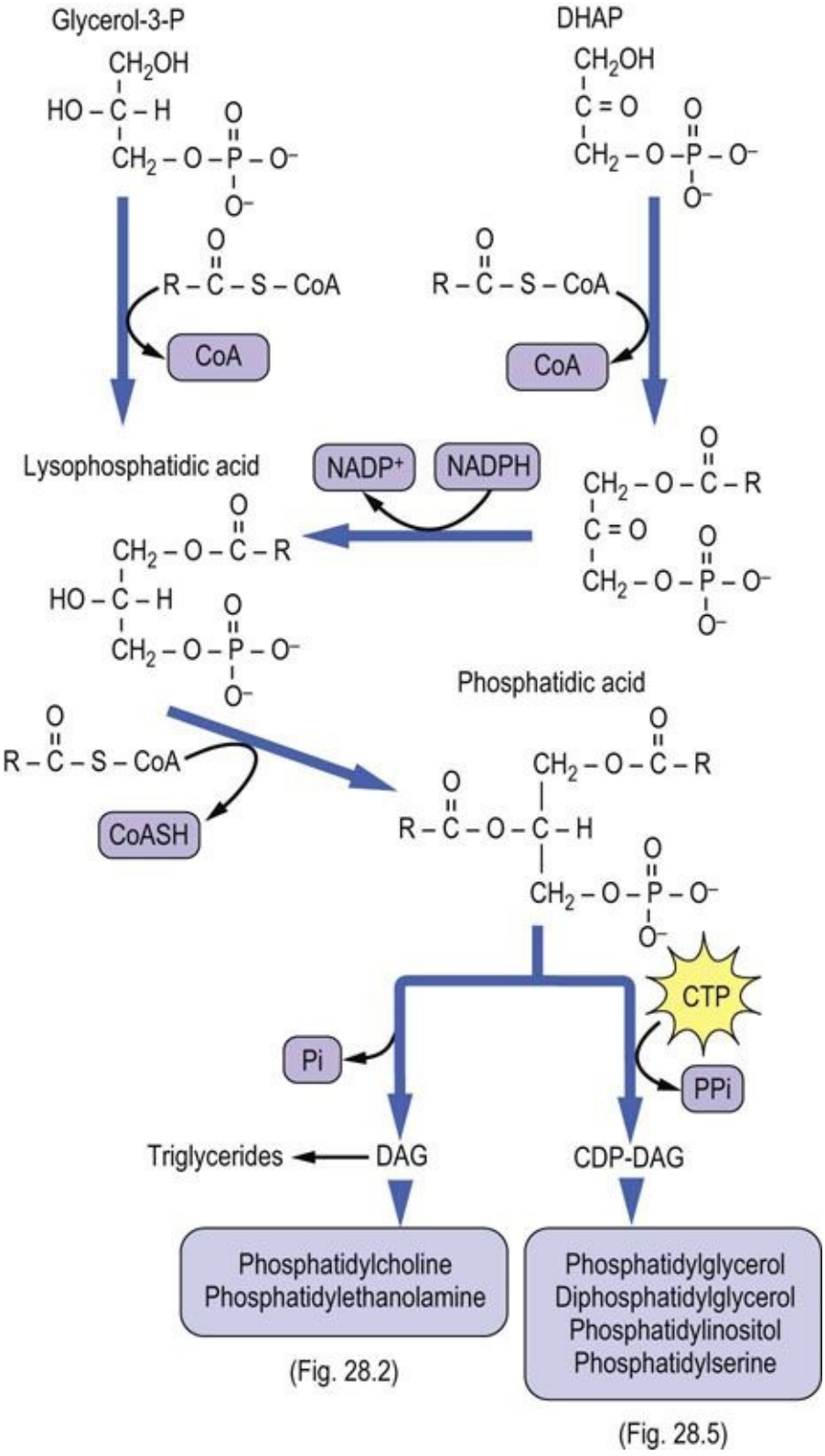


FIG. 28.1 De novo pathway to synthesis of glycerophospholipids.
CDP, cytidine diphosphate; CDP-DAG, CDP-diacylglycerol; CTP, cytidine triphosphate; CoAS, coenzyme A; DHAP, dihydroxyacetone phosphate; Pi, inorganic phosphate; PPi,

inorganic pyrophosphate.

The biosynthesis of the major phospholipid phosphatidylcholine (PC; also known as **lecithin**) from DAG requires activation of choline to CDP-choline. In this series of reactions, shown in Figure 28.2, the choline 'head group' is converted to phosphocholine and then activated to CDP-choline by a pyrophosphorylase reaction. The pyrophosphate bond is cleaved and phosphocholine (choline phosphate) is transferred to DAG to form PC. This reaction is analogous to the transfer of GlcNAc-6-P to dolichol or the high-mannose core of lysosomal enzymes – both the sugar and a phosphate are transferred from the nucleotide derivative. Phosphatidylethanolamine (PE) is formed by a similar pathway using CTP and phosphoethanolamine, to form CDP-ethanolamine. Both PC and PE can react with free serine by an exchange reaction to form phosphatidylserine (PS) and the free base, choline or ethanolamine (Fig. 28.3).

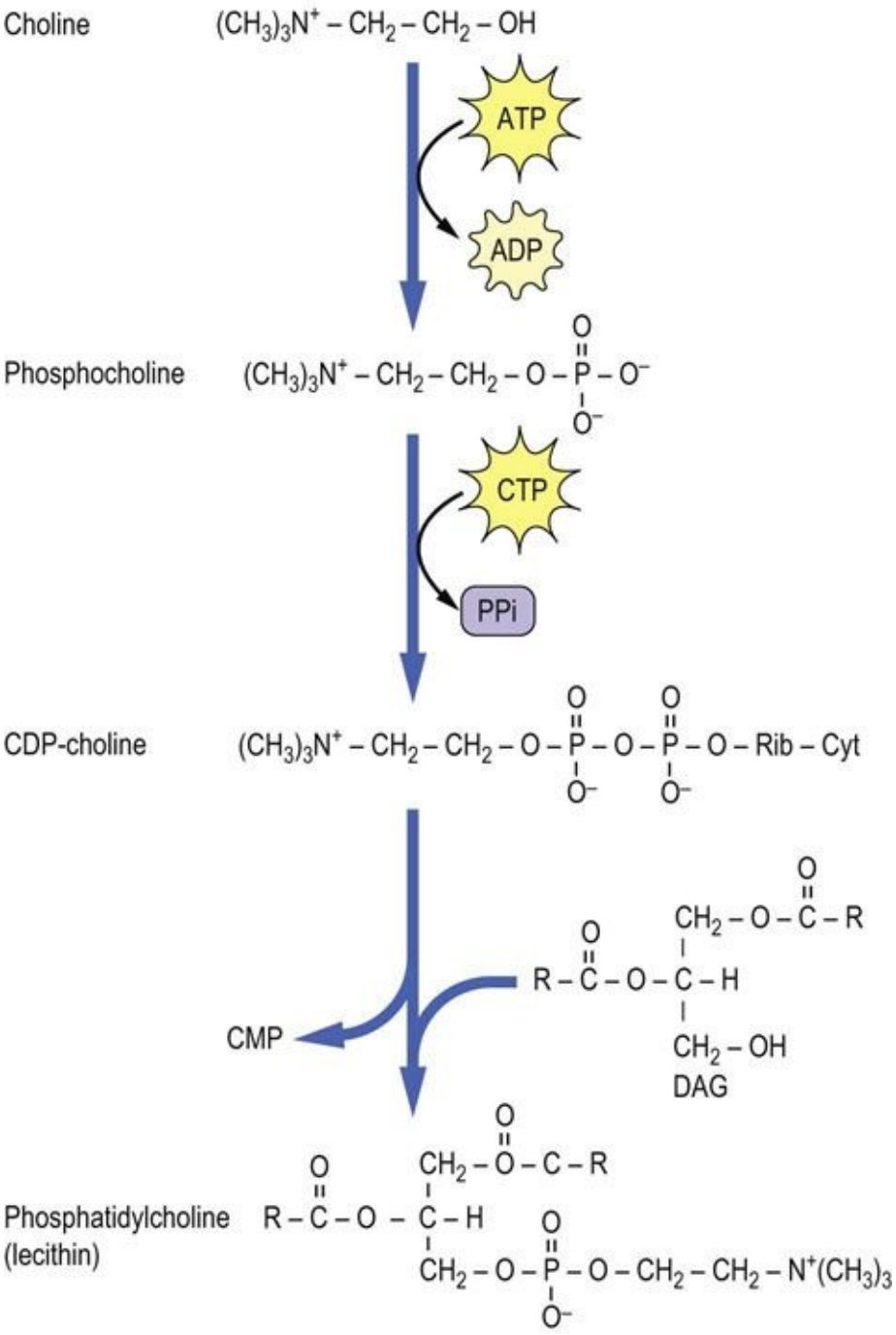


FIG. 28.2 Formation of phosphatidylcholine by the CDP-choline pathway.
This pathway is an extension of the bottom left side of Figure 28.1. Cyt, cytosine; CDP, cytidine diphosphate; CMP, cytidine monophosphate; DAG, diacylglycerol; Rib, ribose; CTP, cytidine triphosphate.

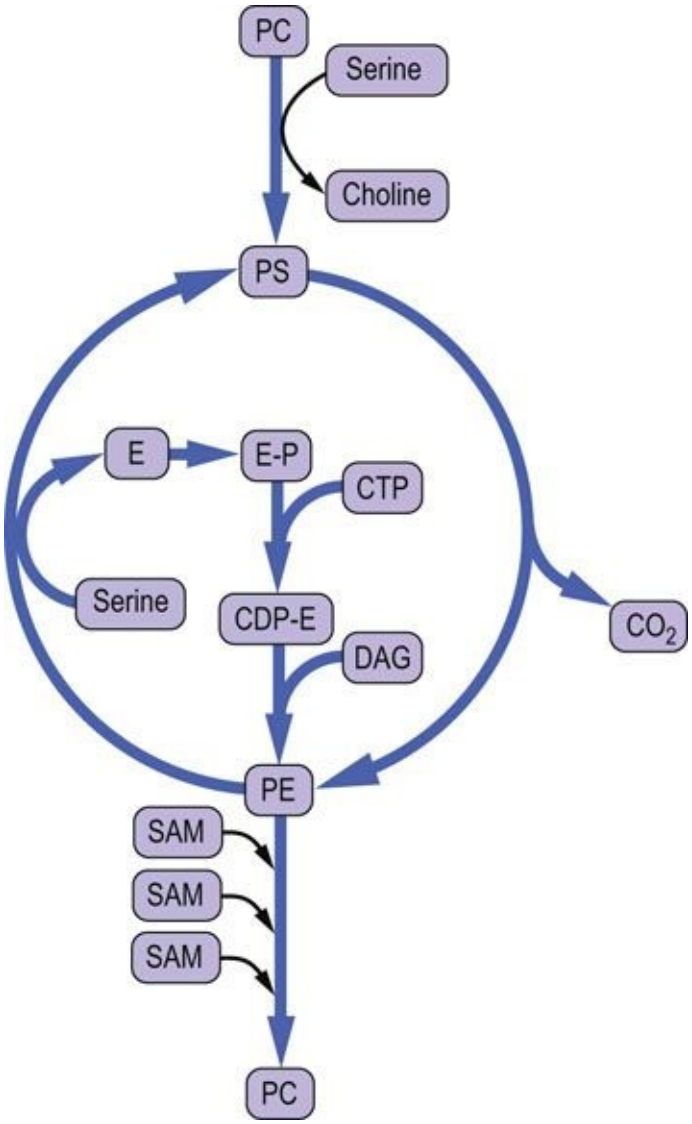


FIG. 28.3 Pathways of interconversion of phospholipids by exchange of head groups, by methylation or by decarboxylation.
CDP, cytidine diphosphate; CTP, cytidine triphosphate; DAG, diacylglycerol; E, ethanolamine; PC, phosphatidylcholine; PE, phosphatidylethanolamine; PC, phosphatidylserine; SAM, *S*-adenosylmethionine.

Large amounts of PC are needed in liver for biosynthesis of lipoproteins and bile in liver. In a secondary pathway, which is a necessary supplement to the above CDP-choline pathway in liver in times of starvation, PC can also be formed by methylation of PE with the methyl donor **S-adenosylmethionine (SAM)** (Figs 28.3 and 28.4). This methylation pathway involves the sequential transfer of three activated methyl groups from three molecules of SAM. PE used in this pathway is supplied from PS by a specific mitochondrial decarboxylase

(Fig. 28.3).

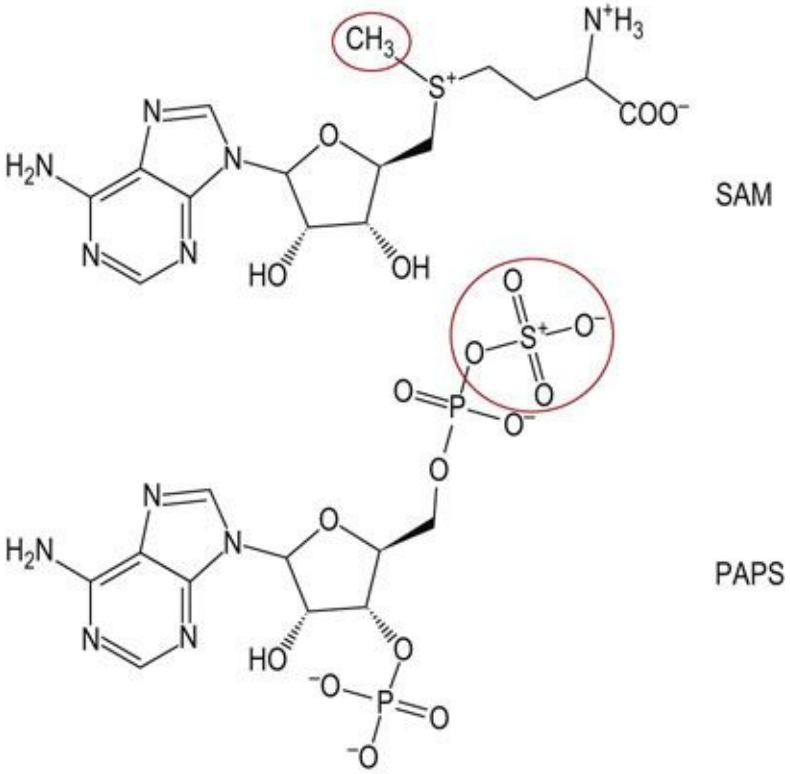


FIG. 28.4 Structures of the methyl and sulfate donors involved in the synthesis of membrane lipids.
SAM, *S*-adenosylmethionine; PAPS, 3'-phosphoadenosine-5'-phosphosulfate (active sulfate).

PS and other phospholipids with an alcohol head group, *e.g.* phosphatidylglycerol (PG) and phosphatidylinositol (PI), are synthesized by an alternative pathway. In this case, PA is activated by CTP, yielding CDP-DAG (see Fig. 28.5); the PA group is then transferred to free serine, glycerol or inositol, to form PS, PG or PI, respectively. A second PA may also be added to phosphatidylglycerol to form 1,3-diphosphatidylglycerol (DPG). This lipid, known commonly as **cardiolipin**, is found almost exclusively in the inner mitochondrial membrane; it represents about 20% of phospholipids in heart mitochondria and is required for efficient activity of electron transport complexes III and IV and the ATP:ADP translocase.

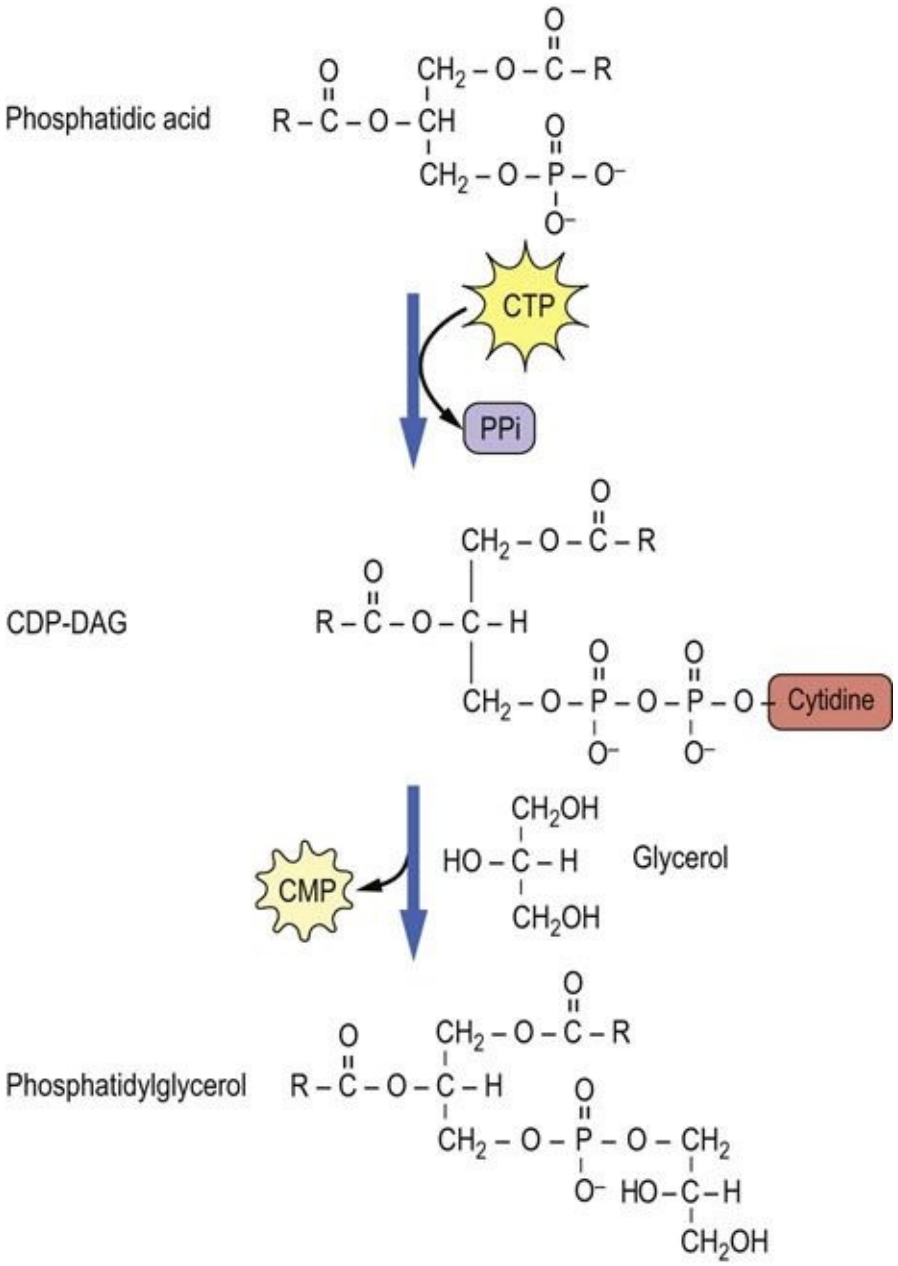


FIG. 28.5 Formation of phosphatidylglycerol by activation of phosphatidic acid to form CDP-DAG, and transfer of DAG to glycerol.
This pathway is an extension of the lower right side of Figure 28.1. CMP, cytidine monophosphate; CTP, cytidine triphosphate.

Plasmalogens are a second major class of mitochondrial lipids, and are enriched in nerve and muscle tissue; in the heart they account for nearly 50% of total phospholipids. The biosynthesis of plasmalogens proceeds from DHAP: it is first acylated at C-1; then the acyl group exchanges with a lipid alcohol to form an ether lipid. The ether lipid is desaturated, leading eventually to a 1-

alkenylether-2-acyl-phospholipid. The function of plasmalogens versus diacylphospholipids is not clear, but there is some evidence that they are more resistant to oxidative damage, which may provide protection against oxidative stress in tissues with active aerobic metabolism (see Chapter 37).

Clinical box Surfactant function of phospholipids: the acute respiratory distress syndrome

Acute respiratory distress syndrome (ARDS) accounts for 15–20% of neonatal mortality in Western countries. The disease affects premature infants and its incidence is directly related to the degree of prematurity.

Comment.

Immature lungs do not have enough type II epithelial cells to synthesize sufficient amounts of the phospholipid **dipalmitoylphosphatidylcholine (DPPC)**. This phospholipid makes up more than 80% of the total phospholipids of the extracellular lipid layer that lines the alveoli of normal lungs. DPPC decreases the surface tension of the aqueous surface layer of the lungs, facilitating opening of the alveoli during inspiration. Lack of surfactant causes the lungs to collapse during the expiration phase of breathing, leading to ARDS. The maturity of the fetal lung can be assessed by measuring the lecithin:sphingomyelin ratio in amniotic fluid. If there is a potential problem, a mother can be treated with a glucocorticoid to accelerate maturation of the fetal lung. ARDS is also seen in adults in whom the type II epithelial cells have been destroyed as a result of the use of immunosuppressive drugs or certain chemotherapeutic agents.

Remodeling pathway

The acyl groups of glycerophospholipids are highly diverse and distributed in an

asymmetric manner between the *sn*-1 and *sn*-2 position of glycerol; polyunsaturated fatty acids, such as arachidonate, are found predominately at the *sn*-2 position. The composition of fatty acyl groups in phospholipids also varies among tissues and membranes, and with the nature of the head group: choline, ethanolamine, serine, inositol or glycerol. The diversity and asymmetry of phospholipids are not explained by the de novo pathway since phosphatidic acid and DAG are common precursors of both triglycerides and phospholipids. Instead, the redistribution of fatty acids in phospholipids is accomplished by remodeling pathways through the concerted action of phospholipase A_2 (PLA_2) and lysophospholipid acyltransferases (LPLAT), which remove, replace and, in the process, redistribute fatty acids in phospholipids. Not until the last decade have the LPLAT enzymes been identified; they play an essential role in (re)incorporation of polyunsaturated fatty acids into phospholipids.

Turnover of phospholipids

Phospholipids are in a continuous state of turnover in most membranes. This occurs as a result of oxidative damage, during inflammation, and through activation of phospholipases, particularly in response to hormonal stimuli. As shown in Figure 28.6, there are a number of phospholipases that act on specific bonds in the phospholipid structure. Phospholipase A_2 (PLA_2) and phospholipase C (PLC) are particularly active during the inflammatory response and in signal transduction. Phospholipase B (not shown) is a lysophospholipase that removes the second acyl group after action of PLA_1 or PLA_2. The lysophospholipids may be degraded or recycled (reacylated).

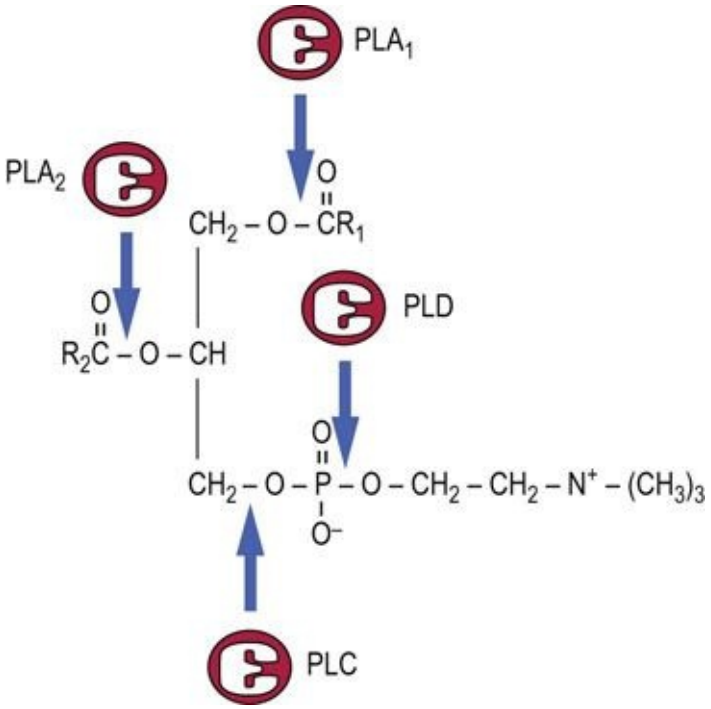


FIG. 28.6 Sites of action of phospholipases on phosphatidylcholine. PLA_1, PLA_2, PLC, PLD are phospholipases A_1, A_2, C, and D, respectively.

Advanced concept box
Glycosylphosphatidylinositol membrane anchors

Phosphatidylinositol is an integral component of the glycosylphosphatidylinositol (GPI) structure that anchors various proteins to the plasma membrane (Fig. 28.7). In contrast to other membrane phospholipids, including most of the membrane phosphatidylinositol, GPI has a glycan chain containing glucosamine and mannose attached to the inositol. Ethanolamine connects the GPI-glycan to the carboxyl terminus of the protein. Many membrane proteins in eukaryotic cells are anchored by a GPI structure, including alkaline phosphatase and acetylcholinesterase, which have roles in bone mineralization and nerve transmission, respectively. In contrast to integral or peripheral membrane proteins, GPI-anchored proteins may be released from the cell surface by phospholipase C in response to

regulatory processes.

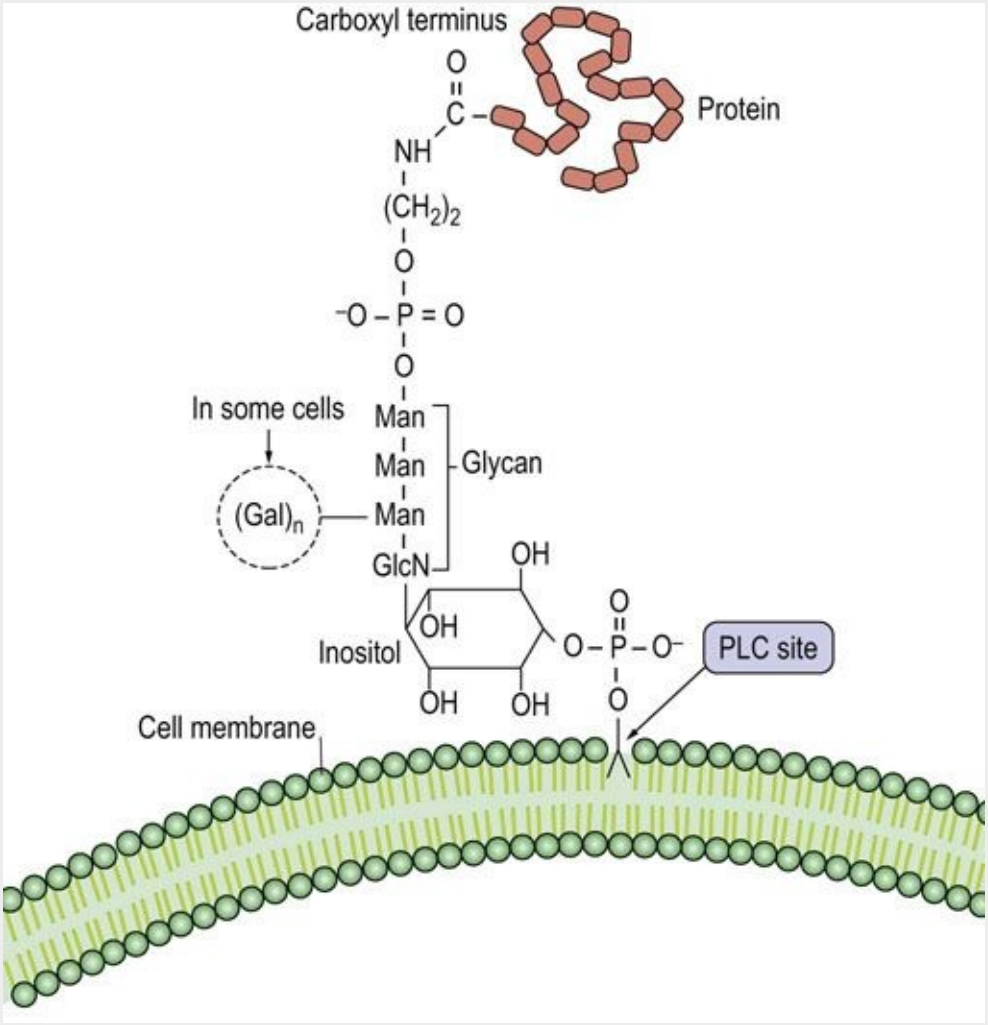


FIG. 28.7 Structure of the glycosylphosphatidylinositol (GPI) anchor and its attachment to proteins.
Gal, galactose; GlcN, glucosamine; Man, mannose; PLC, phospholipase C.

Sphingolipids

Structure and biosynthesis of sphingosine

Sphingolipids are a complex group of amphipathic, polar lipids. They are built on a core structure of the long-chain amino alcohol sphingosine, which is formed by oxidative decarboxylation and condensation of palmitate with serine. In all sphingolipids, the long-chain fatty acid is attached to the amino group of the sphingosine in an amide linkage (Fig. 28.8). Because of the alkaline stability of amides, compared to esters, sphingolipids are nonsaponifiable, which facilitates their separation from alkali-labile glycerolipids.

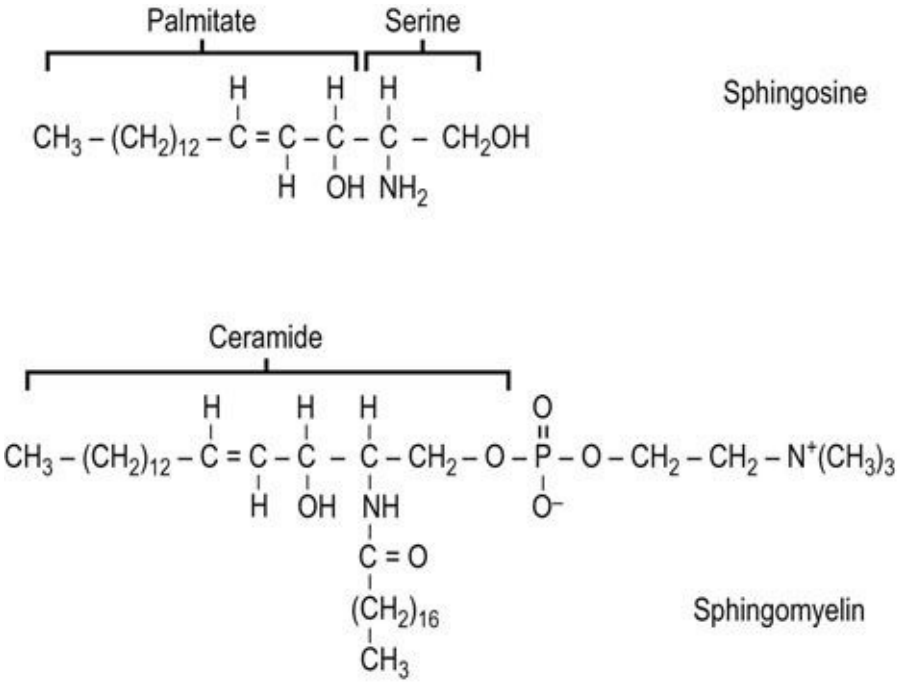
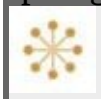


FIG. 28.8 Structures of sphingosine and sphingomyelin.

The synthesis of the sphingosine base of sphingolipids involves condensation of palmitoyl-CoA with serine, in which the carbon-1 of serine is lost as carbon dioxide. The product of this reaction is converted in several steps to sphingosine, which is then *N*-acylated to form ceramide (*N*-acylsphingosine). Ceramide (see Fig. 28.9) is the precursor and backbone structure of both sphingomyelin and

glycosphingolipids.



Advanced concept box Variable surface antigens of trypanosomes

The parasitic trypanosome that causes sleeping sickness, *Trypanosoma brucei*, has a protein called the variable surface antigen bound to its cell surface by a GPI anchor. This variable surface antigen elicits the formation of specific antibodies in the host, and these antibodies can attack and kill the parasite. However, some of the parasites evade immune surveillance by shedding this antigen, as if they were shedding a coat.

Comment.

Trypanosomes and some other pathogens are able to shed their surface antigens because they have an enzyme, phospholipase C, that cleaves the GPI anchor at the phosphate–diacylglycerol bond, releasing the protein-glycan component into the external fluid. Surviving cells rapidly make a new coat with a different antigenic structure that will not be recognized by the antibody. Of course, this new coat will elicit the formation of new specific antibodies but the parasite can again shed this coat, and so on, in a random sequence to evade the host immune system.

Clinical box Defects in GPI anchoring in hematopoietic cells: paroxysmal nocturnal hemoglobinuria

Paroxysmal nocturnal hemoglobinuria (PNH) is a complex hematologic disorder characterized by hemolytic anemia, venous thrombosis in unusual sites, and deficient hematopoiesis. The diagnosis of this disease is based on the unusual sensitivity of the red blood cells to the hemolytic action of complement (Chapter 38), because red cells from patients with PNH lack several proteins

that are involved in regulating the activation of complement at the cell surface.

Comment.

One of these cell surface proteins is decay accelerating factor, a GPI-anchored protein that inactivates a hemolytic complex formed during complement activation; in its absence, there is increased hemolysis. PNH is an acquired genetic disease due to a hematopoietic stem cell mutation defect. One of these mutations involves a defect in the GlcNAc transferase that adds *N*-acetylglucosamine to the inositol moiety of phosphatidylinositol, the first step in GPI anchor formation (see Fig. 28.7).

Sphingomyelin

Sphingomyelin is the only sphingolipid that contains phosphate and is the major phospholipid in the myelin sheath of nerves

Sphingomyelin (see Fig. 28.8) is found in plasma membranes, subcellular organelles, endoplasmic reticulum, and mitochondria. It comprises 5–20% of the total phospholipids in most cell types, and is mostly localized in the plasma membrane. The phosphocholine group in sphingomyelin is transferred to the terminal hydroxyl group of sphingosine by a transesterification reaction with phosphatidylcholine. The fatty acid composition varies, but long-chain fatty acids are common, including lignoceric (24 : 0), cerebronic (2-hydroxylignoceric) and nervonic (24 : 1) acids.

Glycolipids

Sphingolipids containing covalently bound sugars are known as glycosphingolipids or glycolipids. As with glycoconjugates, in general, the structure of the oligosaccharide chains is highly variable. In addition, the glycosyltransferase distribution and glycosphingolipid content of cells varies during development and in response to regulatory processes.

Glycolipids can be classified into three main groups: neutral glycolipids, **sulfatides** and gangliosides. In all of these compounds, the polar head-group, comprising the sugars, is attached to ceramide by a glycosidic bond at the terminal hydroxyl group of sphingosine. Figure 28.9 illustrates the structure and biosynthesis of some of the simpler glycolipids. Neutral glycolipids contain only neutral and amino sugars. Glucosylceramide (GlcCer) and galactosylceramide (GalCer) are the smallest members of this class of compounds and serve as the nucleus for elaboration of more complex structures. Sulfatides are formed by addition of sulfate from the sulfate donor, **3'-phosphoadenosine-5'-phosphosulfate (PAPS)** (Fig. 28.4), yielding, for example, GalCer 3-sulfate. Finally, glycolipids containing sialic acids (*N*-acetylneuraminic acid, NeuAc) are termed gangliosides.

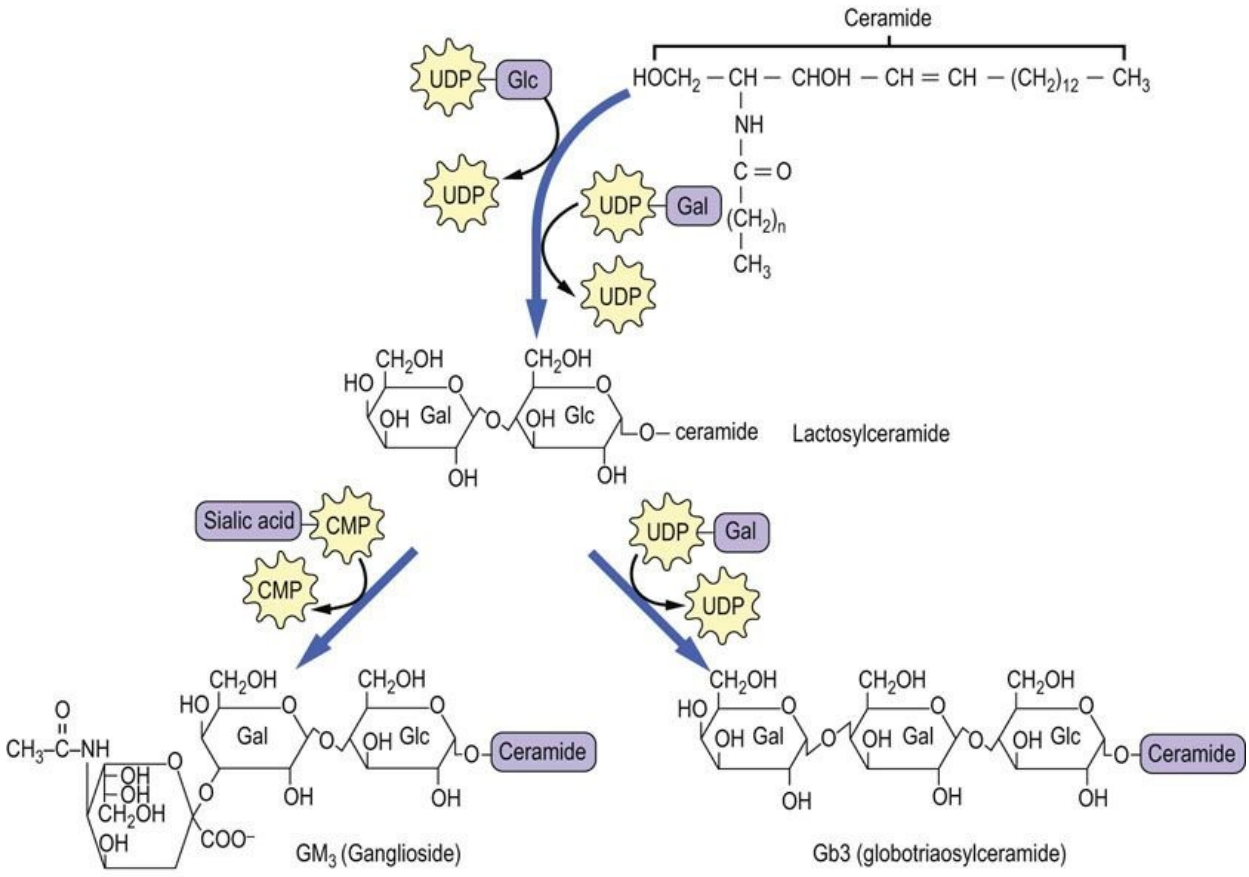


FIG. 28.9 An outline of transferase reactions for elongation of glycolipids and formation of gangliosides.

Structure and nomenclature of gangliosides

Gangliosides are glycosphingolipids containing sialic (N-acetylneuraminic) acid

The term ‘ganglioside’ refers to glycolipids that were originally identified in high concentrations in ganglionic cells of the central nervous system. In general, more than 50% of the sialic acid in these cells is present in gangliosides. Gangliosides are also found in the surface membranes of cells of most extraneural tissues, but in these tissues they account for less than 10% of the total sialic acid.

The nomenclature used to identify the various gangliosides is based on the number of sialic acid residues contained in the molecule, and on the sequence of the carbohydrates (Fig. 28.10). GM refers to a ganglioside with a single (mono) sialic acid, whereas GD, GT and GQ would indicate two, three and four sialic acid residues in the molecule, respectively. The number after the GM, *e.g.* GM_1, refers to the structure of the oligosaccharide. These numbers were derived from the relative mobility of the glycolipids on thin layer chromatograms; the larger, GM_1, gangliosides migrate the most slowly.

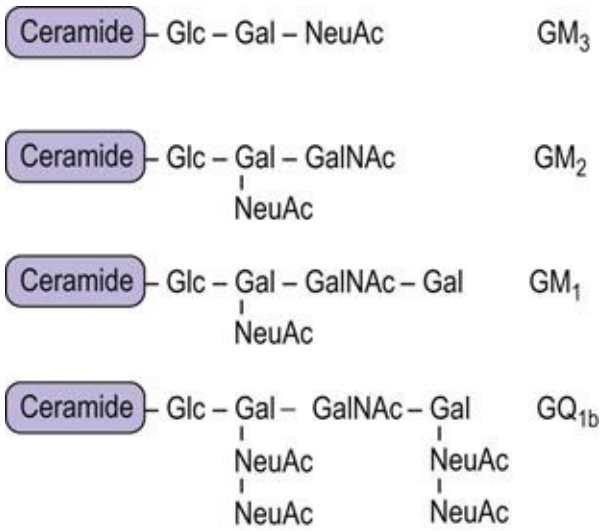


FIG. 28.10 Generalized structures of gangliosides.
Glc, glucose; Gal, galactose; NeuAc, *N*-acetylneuraminic acid; GalNAc, *N*-acetylgalactosamine.

Clinical box Sphingolipidoses and gangliosidoses

Tay–Sachs disease is a gangliosidosis in which ganglioside GM_2 accumulates as a result of an absence of hexosaminidase A (see Fig. 28.11). Individuals with this disease usually have mental retardation and blindness and die between 2 and 3 years of age. **Fabry's disease** is a sphingolipidosis resulting from deficiency of α-galactosidase and accumulation of globotriaosylceramide (Gb3) (see Table 28.1). The symptoms of Fabry's disease are skin rash, kidney failure, and pain in the lower extremities. Patients with this condition benefit from kidney transplants and usually live into early to mid-adulthood. Most of these lysosomal storage diseases appear in several forms (variants), resulting from different mutations in the genome. Some lysosomal storage diseases and some variants are more severe and debilitating than others. Although lysosomal storage diseases are relatively rare, they have had a major impact on our understanding of the function and importance of lysosomes.

Table 28.1

Some lipid storage diseases

Disease	Symptoms	Major storage product	Deficient enzymes
Tay–Sachs	Blindness, mental retardation, death between 2nd and 3rd year	GM_2 ganglioside	Hexosaminidase A
Gaucher's	Liver and spleen enlargement, mental retardation in infantile form	Glucocerebroside	β-Glucosidase
Fabry's	Skin rash, kidney failure, pain in lower extremities	Ceramide trihexoside	α-Galactosidase
Krabbe's	Liver and spleen enlargement, mental retardation	Galactocerebroside	β-Galactosidase

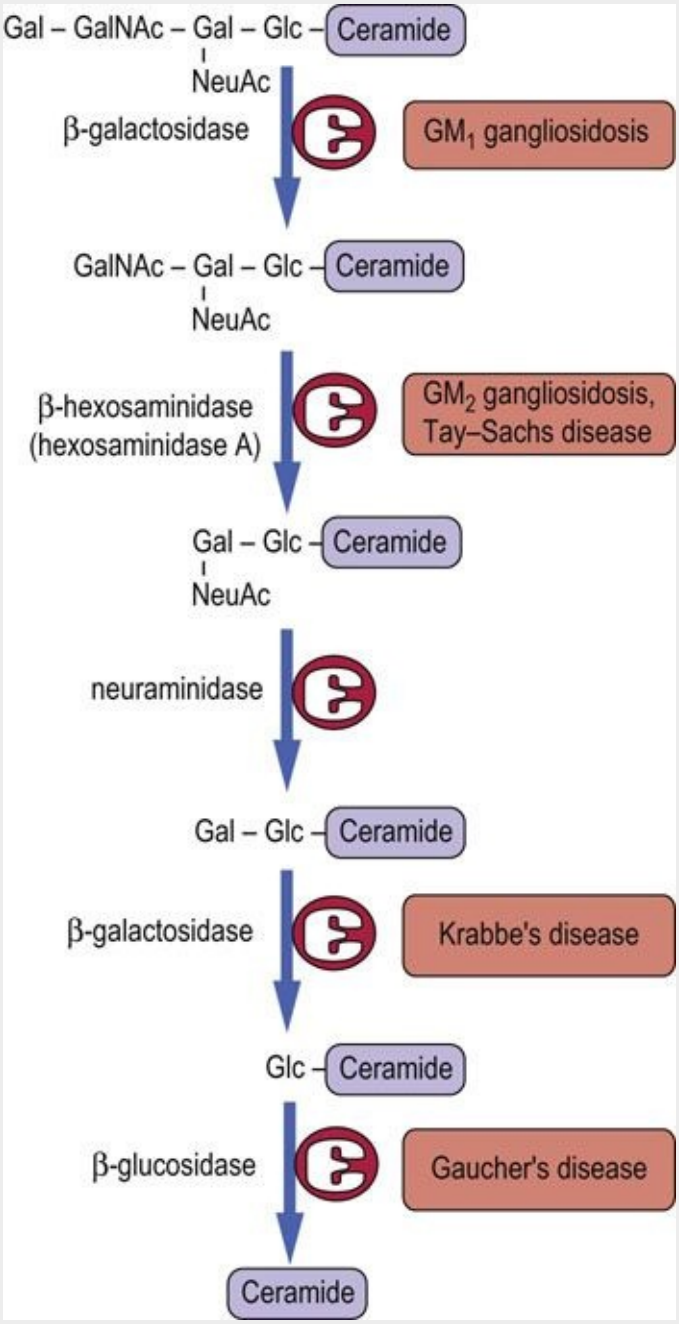


FIG. 28.11 Lysosomal pathway for turnover (degradation) of ganglioside GM_1 in human cells.

Various enzymes may be missing in specific lipid storage diseases, as indicated in Table 28.1. Gal, galactose; GalNAc, *N*-acetylgalactosamine; Glc, glucose; NeuAc, *N*-acetylneuraminic acid.

Comment.

When cells die, biomolecules, including glycosphingolipids and

glycoproteins, are degraded to their individual components. Figure 28.11 presents the pathway for the degradation of ganglioside GM_1 in the lysosomes. A number of lysosomal diseases result from the absence of an essential glycosidase (see Table 28.1). The sphingolipidoses are characterized by lysosomal accumulation of the substrate of the missing enzyme, which interferes with normal lysosomal function in turnover of biomolecules.

Clinical box Gaucher's disease: a model for enzyme replacement therapy

Gaucher's disease is a lysosomal storage disease in which afflicted individuals are missing the enzyme β-glucosidase (also known as glucocerebrosidase). This enzyme removes the final sugar from the ceramide, allowing the lipid portion to be further degraded in the lysosomes. This disease is characterized by hepatomegaly and neuro-degeneration, but there are milder variants that are amenable to treatment by enzyme replacement therapy.

For treatment of Gaucher's disease, exogenous β-glucosidase was successfully targeted to the lysosomes of macrophages. To do this, it was necessary to produce the recombinant replacement enzyme with *N*-glycan chains containing terminal mannose residues. This was done by cleaving the glycans of the enzyme produced in mammalian cells with a combination of sialidase (neuraminidase), β-galactosidase and β-hexosaminidase to trim the complex chains down to the mannose core. An alternative recombinant glucosidase has been produced in a baculovirus-infected insect cell system. Both enzymes have a high mannose oligosaccharide; although they do not contain Man-6-P residues, they are recognized by a macrophage cell surface receptor for high-mannose oligosaccharides; the enzymes are endocytosed and end up in the lysosomal compartment where they hydrolyze glucosyl-ceramide. The recombinant enzymes are administered intravenously. The success in using recombinant

glucocerebrosidase for treatment of Gaucher's disease has stimulated the development of other lysosomal hydrolases for treatment of lysosomal storage diseases.

Lysosomal storage diseases resulting from defects in glycolipid degradation

The complex oligosaccharides on glycolipids are built up, one sugar residue at a time, in the Golgi apparatus and are degraded in a similar stepwise fashion but in the opposite direction by a series of exoglycosidases in lysosomes (Fig. 28.11). Defects in sequential degradation of glycolipids lead to a number of **lysosomal storage diseases**, known as sphingolipidoses and gangliosidoses (Table 28.1). These diseases are autosomal recessive in inheritance. Heterozygotes are asymptomatic, indicating that a single copy of the gene for a functional enzyme is sufficient for apparently normal turnover of glycolipids. Like I-cell disease (Chapter 27), the sphingolipidoses are characterized by accumulation of undigested lipids in inclusion bodies in the cells.

Clinical box Fabry's disease (incidence 1 in 100,000)

A 30-year-old man was found to have proteinuria at an insurance medical examination. He had been seen over a number of years from around age 10 with headaches, vertigo and shooting pains in his arms and legs. No diagnosis was made and he had grown accustomed to these problems. The physician carefully examined his perineum and scrotum, identifying small, raised, red angiokeratoma.

Comment.

This man was diagnosed with Fabry's disease, which often takes years before a diagnosis is confirmed by measuring α-galactosidase A activity. The principal endothelial depositions of a ceramide trihexoside (Gal-α1-4-Gal-β1-4-Glc-β-ceramide; Gb3) occur in the kidney (leading to proteinuria and renal failure), the heart and brain (leading to myocardial infarction and stroke), and around blood vessels supplying nerves (leading to painful paresthesiae). Historically, most patients experienced end-stage renal disease, requiring transplantation. However, recombinant enzyme replacement therapy appears to clear the deposited Gb3

and initial studies suggest that renal function is maintained.

ABO blood group antigens

Blood transfusion replenishes the oxygen-carrying capacity of blood in persons who suffer from blood loss or anemia (see Chapter 5). The term '**blood transfusion**' is something of a misnomer, because it involves only the infusion of washed and preserved red cells. The membranes of red blood cells contain a number of blood group antigens, of which the ABO blood group system is the best understood and most widely studied.

The ABO blood group antigens are complex carbohydrates present as components of glycoproteins or glycosphingolipids of red cell membranes (Fig. 28.12). The H locus codes for a fucosyltransferase, which adds fucose to a galactose residue in a glycan chain. Individuals with type A blood have, in addition to the H substance, an A gene that codes for a specific GalNAc transferase that adds GalNAc α1,3 to the galactose residue of **H substance**, to form the A antigen. Individuals with type B blood have a B gene that codes for a galactosyl transferase that adds galactose α1,3 to the galactose residue of H substance, to form the B antigen. Individuals with type AB blood have both the GalNAc and the Gal transferases, and their red blood cells contain a mixture of A and B substances. Those with type O blood have only H substance on their red cell membranes; they do not make either enzyme. Enzymes such as coffee bean α-galactosidase can remove the galactose from type B red cells, an approach that is being tested for increasing the supply of type O (universal donor) red cells.

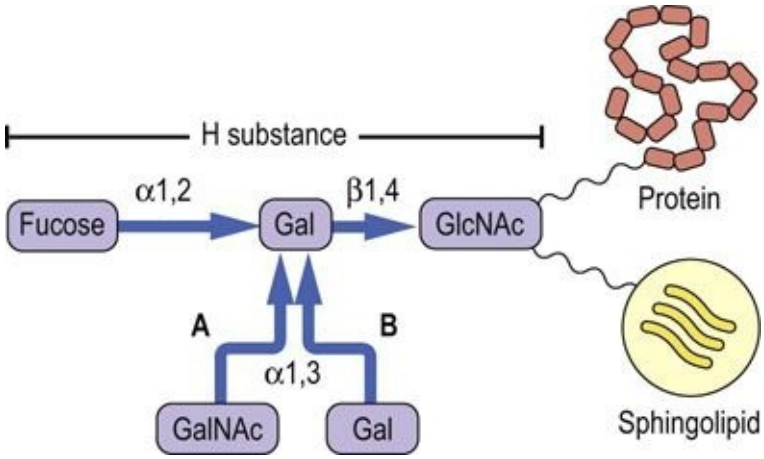


FIG. 28.12 Relationship between the H, A, and B blood group substances.

The terminal oligosaccharide is linked through other sugars to proteins and lipids of the red cell membrane. GlcNAc, *N*-acetylglucosamine; GalNAc, *N*-acetylgalactosamine; Gal, galactose.

An individual may have type A, type B, type AB or type O blood. Individuals with type A cells develop natural antibodies in their plasma that are directed against and will agglutinate type B and type AB red blood cells; those with type B red cells develop antibodies against A substance and will agglutinate type A and type AB blood. Persons with type AB blood have neither A nor B antibodies and are called '**universal recipients**', as they can be transfused with cells of either blood type. Individuals with type O blood have only H substance, not A or B substance, on their red blood cells and are '**universal donors**', as their red blood cells are not agglutinated by either A or B antibodies; however, they may accept blood only from a type O donor.

The ABO antigens are present on most cells in the body but they are referred to as blood group antigens because of their association with transfusion reactions. The transfusion reaction is the result of reaction of host antibodies with transfused red cells, resulting in complement-mediated hemolysis (see Chapter 38). While the transfusion reaction demonstrates the role of carbohydrates in recognition of the foreign red cells, the physiologic function of blood group substances is unclear. Persons with an O genotype are generally as healthy as those with A or B genotype. However, there is some evidence that specific phenotypes may confer differential resistance to disease; for example, people with type A and type O blood appear to be more susceptible to smallpox and cholera, respectively.

Other blood group substances

Advanced concept box Glycolipids are binding sites for bacteria and bacterial toxins

Bacteria have evolved proteins called adhesins that recognize and interact with specific carbohydrate structures on glycolipids, glycoproteins and even proteoglycans. Many of the bacterial adhesins are protein subunits of pili, hair-like structures on the surfaces of the bacteria. The carbohydrate recognition domains are usually located at the tip of the pili. Most bacteria also have several different kinds of adhesins on their surfaces, each having different carbohydrate recognition sites, and these adhesins define the range of susceptible tissues that the bacteria can bind to and perhaps invade. Each individual adhesin binding is of low affinity and the binding is weak, but there are many copies of a given adhesin on the bacterial surface and they cluster together, so the total interaction is polyvalent rather than monovalent and binding becomes quite strong. The interaction of the adhesin with its receptor can activate signal transduction pathways and lead to events that are critical for colonization and perhaps infection.

A number of bacterial adhesins target Galβ1-4Glc-containing oligosaccharides. This is the disaccharide structure that is found in the glycolipid lactosylceramide, and this structure may be present as such or it may be capped with other sugars, as in the ABO blood group antigens. But some bacteria secrete enzymes (glycosidases) that can remove these terminal sugars to expose the lactose structure for binding to their adhesin. The epithelial cells of the large intestine express lactosylceramide, whereas the cells lining the small intestine do not express this glycolipid. As a result, *Bacteroides*, *Clostridium*, *Escherichia coli* and *Lactobacillus* only colonize the large intestine under normal conditions.

In addition to the bacterial binding, a number of toxins that are secreted from bacterial cells also bind to specific glycolipids. The

best studied of these toxins is cholera toxin, the toxin produced by *Vibrio cholerae*, which binds to GM_1. The toxin from *Shigella dysenteriae* also binds to cells of the large intestine, but it recognizes a different glycolipid, in this case Gb3. These two examples show quite clearly how subtle changes in the structures of carbohydrate molecules can be recognized by different proteins, and why nature has selected carbohydrates molecules as providers of chemical recognition information. There are other toxins, such as the tetanus toxin produced by *Clostridium tetani* or botulinum toxin by *Clostridium botulinum*, which also bind to glycolipids on nerve cell membranes. These toxins recognize much more complicated glycolipids. For example, tetanus toxin binds to the ganglioside GT_{1b}.

Clinical box Ganglioside receptor for cholera toxin

The galactose-containing glycolipids in the plasma membranes of intestinal epithelial cells are binding sites for bacteria. The glycolipids appear to assist in retention of normal intestinal flora (symbionts) in the intestine but, conversely, binding of pathogenic bacteria to these and other glycolipids is believed to facilitate infection of the epithelial cells. The difference between symbiotic and parasitic bacteria depends, in part, on their ability to secrete toxins or to penetrate the host cell after the binding reaction.

Intestinal mucosal cells contain ganglioside GM_1 (see Fig. 28.10). This ganglioside serves as the receptor to which cholera toxin binds as the first step in its penetration of intestinal cells. Cholera toxin is a hexameric protein secreted by the bacterium *Vibrio cholerae*. The protein is composed of one A subunit and five B subunits. The protein binds to gangliosides by multiple interactions through the B subunits, which enables the A subunit to enter the cell and activate adenylate cyclase on the inner surface of

the membrane. The cyclic AMP that is formed then stimulates intestinal cells to export chloride ions, leading to osmotic diarrhea, electrolyte imbalances, and malnutrition. **Cholera remains the number one killer of children in the world today.**

The **Lewis blood group** antigens correspond to a set of fucosylated glycan structures. The Lewis-A antigen (Lewis^a) is synthesized by a fucosyltransferase that transfers a fucose residue to a GlcNAc residue in a glycan chain (Fig. 28.13), while the Lewis^b antigen is synthesized by the action of a second fucosyltransferase which transfers fucose to the terminal galactose residue in the same glycan chain. Note the similarity of these structures to the sialyl Lewis-X antigen in Figure 27.9 and the ABO antigens in Figure 28.12. There are 13 fucosyltransferase genes in the human genome. Changes in fucosylation of glycans are associated with differentiation, development, carcinogenesis and metastasis.

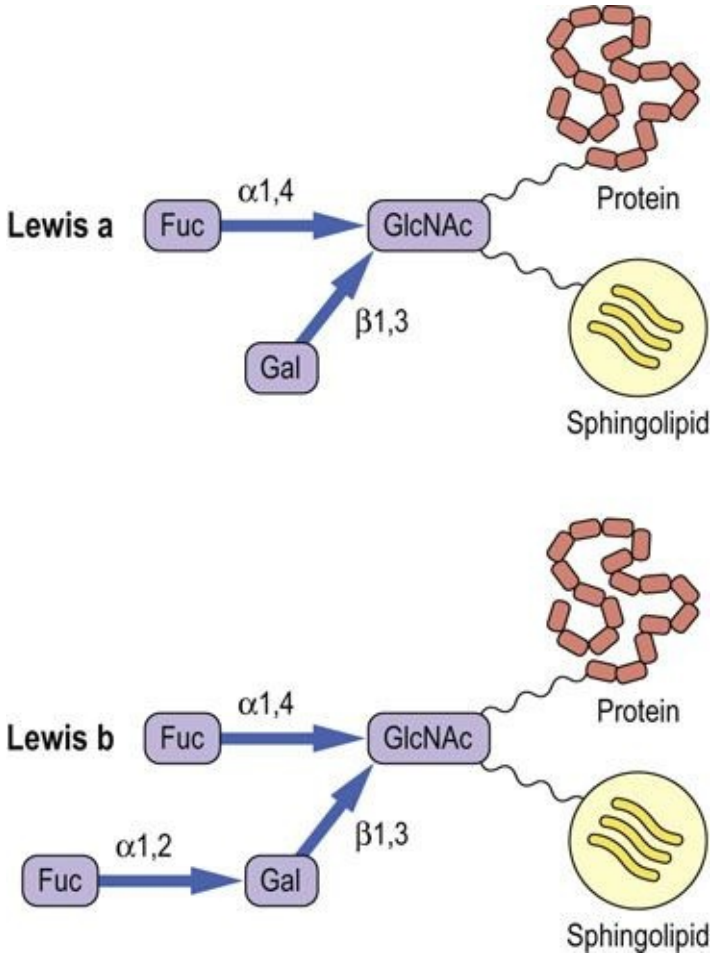


FIG. 28.13 Structure of Lewis blood group antigens.

The **P blood group** antigens expressed on the globo-series glycosphingolipids are distributed in red cells and other tissues. Again, the glycans in this blood group are synthesized by the sequential action of distinct glycosyltransferases. The physiologic function of these blood groups is unknown, but P antigens are associated with urinary tract infections and parvovirus infections. Uropathogenic strains of *E. coli* express lectins that bind to the Galα1,4Gal moiety of the P^k and P_1 antigens. More work is needed to understand the genetics and biochemistry of these and other blood group antigens as well as their roles in physiology and disease.

Summary

■ Complex polar lipids are essential components of all cell membranes.
■ Phospholipids are the major structural lipids of all membranes, but they also have important functional properties as surfactants, as cofactors for membrane enzymes, and as components of signal transduction systems.
■ The primary route for de novo biosynthesis of phospholipids involves the activation of one of the components (either DAG or the head group) with CTP to form a high-energy intermediate, such as CDP-diacylglycerol or CDP-choline.
■ Phospholipids undergo maturation in the remodeling pathway, in which acyl groups at the *sn*-2 position are replaced with new ones, yielding diversity and asymmetry of the hydrophobic moiety of phospholipids.
■ Glycosphigolipids function as receptors for cell–cell recognition and interactions, and as binding sites for symbiotic and pathogenic bacteria and for viruses. Carbohydrate structures on glycosphingolipids of red cell membranes are also antigenic determinants responsible for the ABO and other blood types.
■ Glycosphingolipids are degraded in lysosomes by a sequence of reactions that involve a stepwise removal of sugars from the nonreducing end of the molecule, with each step involving a specific lysosomal exoglycosidase. A number of inherited lysosomal storage diseases result from defects in degradation of sphingolipids.

Active learning

1. Describe the significance of the remodeling pathway of glycerophospholipids in various biological events.
2. Compare the role of plasmalogens vs diacylglycerol-phospholipids in cell membranes.
3. Discuss the challenges in development of a vaccine to protect against trypanosomiasis.
4. Review current therapeutic approaches for diagnosis and treatment of acute respiratory distress syndrome (ARDS).
5. Review the mechanisms of host–pathogen interaction, focusing on the role of lectins in microbial pathogenicity.

Further reading

Beck, M. Therapy for lysosomal storage diseases. *IUBMB Life*. 2010; 62:33–40.

Brown, JR, Crawford, BE, Esko, JD. Glycan antagonists and inhibitors: a fount for drug discovery. *Crit Rev Biochem Mol Biol*. 2007; 42:481–515.

Cipolla, L, Araujo, AC, Bini, D, *et al.* Discovery and design of carbohydrate-based therapeutics. *Expert Opin Drug Discov*. 2010; 5:721–737.

Desnick, RJ, Schuchman, EH. Enzyme replacement therapy for lysosomal diseases: lessons from 20 years of experience and remaining challenges. *Annu Rev Hum Genet*. 2012; 13:307–335.

Kulkarni, AA, Weiss, AA, Iyer, SS. Glycan-based high affinity ligands for toxins and pathogen receptors. *Med Res Review*. 2010; 30:327–393.

Mignani, R, Feriozzi, S, Schaefer, RM, *et al.* Dialysis and transplantation in Fabry disease: indications for enzyme replacement therapy. *Clin J Am Soc Nephrol*. 2010; 5:379–385.

Sindou, H, Shimizu, T. Acyl-CoA:lysophospholipid acyltransferases. *J Biol Chem*. 2009; 284:1–5.

Van Meer, G, Voelker, DR, Feigenson, GW. Membrane lipids: where they are and how they behave. *Nat Rev Mol Cell Biol*. 2008; 9:112–124.

Yamamoto, F, Cid, E, Yamamoto, M, Blancher, A. ABO research in the modern era of genomics. *Transfus Med Rev*. 2012; 26:103–118.

Xu, YH, Barnes, S, Sun, Y, Grabowski, GA. Multi-system disorders of glycosphingolipid and ganglioside metabolism. *J Lipid Res*. 2010; 51:1643–1675.

Websites

Blood groups. www.bloodbook.com/type-sys.html.

Gaucher's disease. www.ninds.nih.gov/disorders/gauchers/gauchers.htm.

Paroxysmal nocturnal hemoglobinuria. http://emedicine.medscape.com/article/207468-overview.

Tay–Sachs disease. www.ninds.nih.gov/disorders/taysachs/taysachs.htm.

CHAPTER 29

The Extracellular Matrix

Gur P. Kaushal, †Alan D. Elbein and Wayne E. Carver

Learning objectives

After reading this chapter you should be able to:

■ Describe the composition, structure and function of the extracellular matrix (ECM) and its components, including collagens, noncollagenous proteins and proteoglycans.

■ Outline the sequence of steps in the biosynthesis and post-translational modification of collagens and elastin, including the structure and synthesis of crosslinks.

■ Discuss the functional roles of the ECM in tissues.

■ Describe the pathways of biosynthesis and turnover of proteoglycans.

■ Discuss the structure and function of integrins as receptors for ECM components.

■ Describe pathologies involving ECM components.

Introduction

The extracellular matrix (ECM) is a complex network of secreted macromolecules located in the extracellular space. Historically, the ECM has been described as simply providing a three-dimensional framework for the organization of tissues and organs; however, it has become increasingly clear that it plays a central role in regulating basic cellular processes, including proliferation, differentiation, migration and even survival. The macromolecular network of the ECM is made up of collagens, elastin, glycoproteins and proteoglycans that are secreted by a variety of cell types including fibroblasts, chondrocytes, osteoblasts and others. The components of the ECM are in intimate contact with their cells of origin and form a three-dimensional gelatinous bed in which the cells thrive. Proteins in the ECM are also bound to the cell surface, so that they transmit mechanical signals resulting from stretching and compression of tissues. The relative abundance, distribution, and molecular organization of ECM components vary enormously, depending upon tissue type, developmental stage and pathologic status. Variations in the composition, accumulation and organization of the ECM dramatically impact the structural and functional properties of the tissue. Changes in these ECM characteristics are associated with chronic diseases, such as arthritis, atherosclerosis, cancer, and fibrosis.

Collagens

Collagens are the major proteins in the ECM

The collagens are a family of proteins that comprise about 30% of total protein mass in the body. As the primary structural components of the ECM in connective tissues, collagens have an important role in tissue architecture and integrity, and in mediating a variety of cell–cell and cell–matrix interactions. To date, more than 25 different types of collagens have been identified. They are composed of related, but distinct, peptide chains and vary greatly in their distribution, organization and function in tissues.

Triple-helical structure of collagens

The left-handed triple helical structure of collagen is unique among proteins

The collagens are heterotrimeric proteins composed of three individual peptide chains. The structural hallmark of collagens is their triple-helical structure, formed by folding of the three component peptide chains. X-ray diffraction analysis indicates that three left-handed helical chains are wrapped around one another in a rope-like fashion, to form a superhelix structure (Fig. 29.1). The left-handed helix is more extended than the α-helix of globular proteins, having nearly twice the rise per turn and only three, rather than 3.6, amino acids per turn of the helix. Every third amino acid is glycine, because only this amino acid, with the smallest side chain, fits into the crowded central core. The characteristic, repeating sequence of collagen is **Gly-X-Y**, where X and Y can be any amino acid but most often X is proline and Y is hydroxyproline. Because of their restricted rotation and bulk, proline and hydroxyproline confer rigidity to the helix. The intra-and interchain helices are stabilized by hydrogen bonds, largely between peptide NH and C═O groups. The side chains of the X and Y amino acids point outward from the helix, and thus are on the surface of the protein, where they form lateral interactions with other triple helices or proteins.

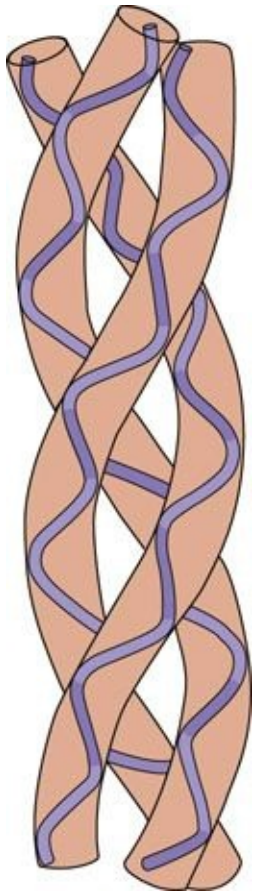


FIG. 29.1 Three-dimensional structure of collagen.
Collagen monomer strands assume a left-handed, **α**-helical tertiary structure. They then associate to form a triple-stranded, right-handed superhelical quarternary structure.

Types of collagen

Some representative collagens are listed in Table 29.1. The collagen family of proteins can be divided into two main types: the fibril-forming (fibrillar) and the nonfibrillar collagens.

Table 29.1

Members of the collagen family. Classification and distribution of different collagen types

Type	Class	Distribution
I	Fibrillar	Widely distributed including skin, tendon, bone, heart
II	Fibrillar	Cartilage, developing cornea and vitreous humor
III	Fibrillar	Extensible connective tissue, *e.g.* skin, lung and vascular system
IV	Network	Basement membranes, kidney, vascular wall

V	Fibrillar	Liver, cornea and mucosa
VI	Beaded filament	Most connective tissue
IX	Facit	Cartilage, vitreous humor
XI	Fibril forming	Cartilage, bone, placenta
XII	Facit	Embryonic tendon and skin
XIII	Transmembrane domain	Widely distributed
XIV	Facit	Fetal skin and tendons

FACIT, fibril-associated collagen with interrupted triple helices.

Fibril-forming collagens

Fibrillar collagens provide tensile strength to tendons, ligaments and skin

Fibril-forming collagens include types I, II, III, V, and XI (see Table 29.1). Collagen fibrils can be formed from a mixture of different fibrillar collagens. For instance, dermal collagen fibrils are hybrids of type I and type III collagen, and fibrils in corneal stroma are hybrids of type I and type IV collagen. Type I is the most abundant fibrillar collagen and occurs in a wide variety of tissues; others have a more limited tissue distribution (see Table 29.1). Type I and related fibrillar collagens form well-organized, banded fibrils and provide high-tensile strength to skin, tendons, and ligaments. As indicated above, collagens are heterotrimers composed of three α-helical peptide chains (see Fig. 29.1). The type I collagen heterotrimer is composed of two α1(I) chains and one α2(I) chain. Each of these peptide chains contains about 1000 amino acids and has a triple-helical domain structure along almost the entire length of the molecule. The collagen fibrils are formed by lateral association of triple helices in a ‘quarter-staggered’ alignment in which each molecule is displaced by about one-quarter of its length relative to its nearest neighbor (Fig. 29.2). The **quarter-staggered array** is responsible for the banded appearance of collagen fibrils in connective tissues. The fibrils are stabilized by both noncovalent forces and interchain crosslinks derived from lysine residues (see below).

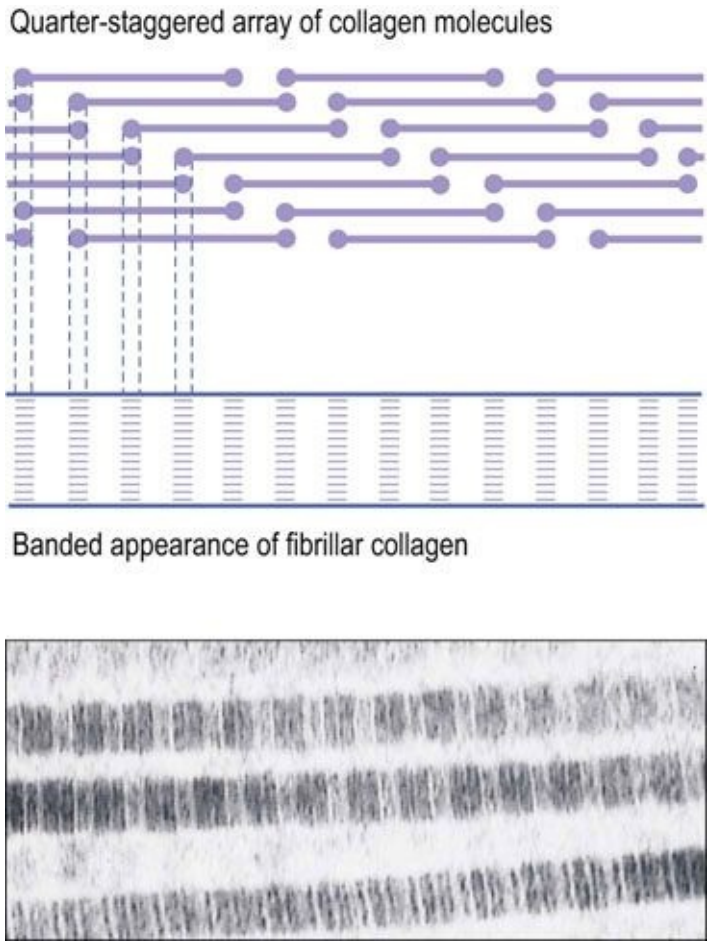


FIG. 29.2 Formation of the quarter-staggered array of collagen molecules in a fibril. The regular overlap of the short, nonhelical termini of the collagen chains yields a regular, banded pattern in the collagen fiber. Electron micrograph courtesy of Dr Trevor Gray.

Nonfibrillar collagens

Nonfibrillar, lattice-forming collagens are major structural components of basement membranes

Clinical box Osteogenesis imperfecta (incidence 1 in 30,000–50,000)

A 6-year-old boy was seen in the casualty department with broken tibia and fibula occurring during a soccer game. His 6-foot-tall

father explained that he had broken his legs four times while at school. The father's teeth were slightly transparent and discolored.

Comment.

Osteogenesis imperfecta (OI), also called brittle bone disease, is a congenital disease caused by multiple genetic defects in the synthesis of type I collagen. It is characterized by fragile bones, thin skin, abnormal teeth and weak tendons. The majority of individuals with this disease have mutations in genes encoding α1(I) or α2(I) collagen chains. Many of these mutations are single-base substitutions that convert glycine in the Gly-X-Y repeat to bulky amino acids, preventing the correct folding of the collagen chains into a triple helix and their assembly to form collagen fibrils. The dominance of type 1 collagen in bone explains why bones are predominantly affected. However, there is remarkable clinical variability characterized by bone fragility, osteopenia, variable degrees of short stature, and progressive skeletal deformities. The most common form of OI, with a presentation that is sometimes mistaken for child abuse, has a good prognosis, with fractures decreasing after puberty, though the general reduction in bone mass means that lifetime risk remains high. Patients frequently develop deafness due to osteosclerosis, partly from recurrent fractures of the stapes. Bisphosphonate drugs (see Chapter 26), which inhibit osteoclast activity and thereby inhibit normal bone turnover, have reduced the incidence of fractures. Long-term follow-up studies are under way.

Nonfibrillar collagens are a heterogeneous group containing triple-helical segments of variable length, interrupted by one or more intervening nonhelical (noncollagenous) segments. This group includes basement membrane collagens (the type IV family), fibril-associated collagens with interrupted triple helices (FACITs), and collagens with multiple triple-helical domains with interruptions, known as multiplexins. Nonfibrillar collagens associate with the fibrillar collagens, forming microfibrils and network or mesh-like structures.

Basement membranes are relatively thin layers of ECM found on the basal

aspect of epithelial cells and surrounding some other cell types including myocytes, Schwann cells and adipocytes. The basement membrane has a number of functions including anchorage of cells to surrounding connective tissue and filtration.

Type IV collagen assembles into a flexible mesh-like network. This collagen contains a long triple-helical domain interrupted by short noncollagenous sequences. These interruptions in the helical domain block continued association of two triple helices, oblige them to find another partner, and thus contribute to formation of a lattice-type structure. In the kidney, the thickened basement membrane (100–200 nm thick) on the basal aspect of the glomuerular capillary endothelial cells plays an essential role as a macromolecular filter (see Chapter 23). The meshwork of ECM proteins in the basement membrane restricts the passage of large molecules from the blood into the urine. In addition, the inclusion of negatively charged proteoglycans (described later in this chapter) in the glomerular basement membrane restricts the passage of charged molecules. Anomalies in type IV collagen in the glomerular basement membrane result in several glomerular diseases including **Goodpasture's syndrome** and **Alport syndrome**. Goodpasture's syndrome is a rare autoimmune disease caused by the production of antibodies that specifically bind to type IV collagen of basement membranes. This inflammatory condition leads to progressive worsening of basement membrane function in the kidney and sometimes in the lung. Alport syndrome results from mutations in the type IV collagen chains which cause defective collagen scaffold assembly within the basement membrane. The symptoms of both of these syndromes progress from blood in the urine (hematuria) to urine containing excessive protein (proteinuria) and eventually to kidney failure.

Synthesis and post-translational modification of collagens

Collagen synthesis begins in the rough endoplasmic reticulum (RER)

After synthesis in the RER, the nascent collagen polypeptide undergoes extensive modification, first in the RER, then in the Golgi apparatus, and finally in the extracellular space, where it is modified to a mature extracellular collagen fibril (Fig. 29.3). A nascent polypeptide chain, preprocollagen, is synthesized

initially with a hydrophobic signal sequence that facilitates binding of ribosomes to the endoplasmic reticulum (ER) and directs the growing polypeptide chain into the lumen of the ER. Post-translational modification of the protein begins with removal of the signal peptide in the ER, yielding procollagen. Three different hydroxylases then add hydroxyl groups to proline and lysine residues, forming 3-and 4-**hydroxyprolines** and δ-**hydroxylysine**. These hydroxylases require ascorbate (vitamin C) as a cofactor (Fig. 29.3, step 1). Vitamin C deficiency leads to **scurvy** as a result of alterations in collagen synthesis and crosslinking (see Chapter 11).

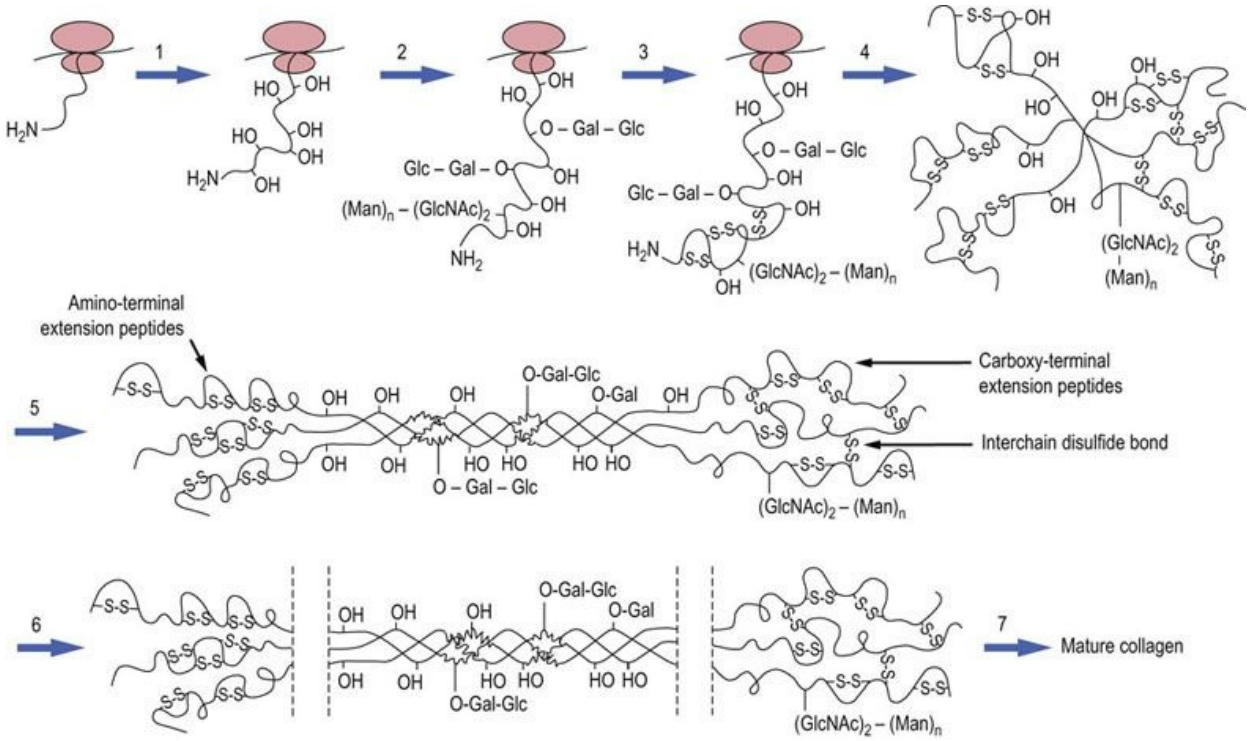


FIG. 29.3 Biosynthesis and post-translational processing of collagen in the endoplasmic reticulum.
Collagen is synthesized in the RER, post-translationally modified in the Golgi apparatus, then secreted, trimmed of extension peptides, and finally assembled into fibrils in the extracellular space. (1) Hydroxylation of proline and lysine residues. (2) Addition of *O*-linked and *N*-linked oligosaccharides. (3) Formation of intrachain disulfide bonds at the *N*-terminal of the nascent polypeptide chain. (4) Formation of interchain disulfides in the *C*-terminal domains, which assist in alignment of chains. (5) Formation of triple-stranded, soluble tropocollagen, and transport to Golgi vesicles. (6) Exocytosis and removal of *N*- and *C*-terminal propeptides. (7) Final stages of processing, including lateral association of triple helices, covalent crosslinking and collagen fiber formation. Gal, galactose; Glc, glucose; GlcNAc, *N*-acetylglucosamine; Man, mannose.

O-linked glycosylation occurs by the addition of galactosyl residues to hydroxylysine by galactosyl transferase; a disaccharide is formed by addition of glucose to galactosyl hydroxylysine by a glucosyl transferase (Fig. 29.3, step 2). These enzymes have strict substrate specificity for hydroxylysine or galactosyl hydroxylysine, and they glycosylate only those peptide sequences that are in noncollagenous domains. *N*-linked glycosylation also occurs on specific asparagine residues in nonfibrillar domains. The nonfibrillar collagens, with a greater extent of nonhelical domains, are more highly glycosylated than fibrillar collagens. Thus, the extent of glycosylation may influence fibril structure, interrupting fibril formation and promoting interchain interactions required for a meshwork structure. Intra-and interchain disulfide bonds are formed in the *C*-terminal domains by a protein disulfide isomerase, facilitating the association and folding of peptide chains into a triple helix (Fig. 28.3, steps 3–5). At this stage, the **procollagen** is still soluble and contains additional, nonhelical extensions at its *N*- and *C*-terminals.

Procollagen is finally modified to collagen in the Golgi apparatus

Clinical box Lathyrism: the result of lysyl oxidase inhibition

Lathyrism is a diet-induced disease characterized by deformation of the spine, dislocation of joints, demineralization of bones, aortic aneurysms, and joint hemorrhages. These problems develop as a result of inhibition of lysyl oxidase, an enzyme required for the crosslinking of collagen chains. Lathyrism can be caused by chronic ingestion of the sweet pea *Lathyrus odoratus*, the seeds of which contain β-aminopropionitrile, an irreversible inhibitor of lysyl oxidase. Penicillamine, a sulfhydryl agent used for chelation therapy in heavy-metal toxicity, also causes lathyrism, because of either chelation of copper required for lysyl oxidase activity or reaction with aldehyde groups of (hydroxy)allysine, inhibiting collagen crosslinking reactions.

After assembly into the triple helix, the procollagen is transported from the RER to the Golgi apparatus, where it is packaged into cylindrical aggregates in secretory vesicles, then exported to the extracellular space by exocytosis. The nonhelical extensions of the procollagen are now removed in the extracellular space, by specific *N*- and *C*-terminal procollagen proteinases (Fig. 29.3, step 6). The '**tropocollagen**' molecules then self-assemble into insoluble collagen fibrils, which are further stabilized by the formation of aldehyde-derived intermolecular crosslinks. **Lysyl oxidase** (not to be confused with lysyl hydroxylase involved in formation of hydroxylysine) oxidatively deaminates the amino group from the side chains of some lysine and hydroxylysine residues, producing reactive aldehyde derivatives, known as **allysine and hydroxyallysine**. The aldehyde groups now form aldol condensation products with neighboring aldehyde groups, generating crosslinks both within and between triple-helical molecules. They may also react with the amino groups of unoxidized lysine and hydroxylysine residues to form Schiff base (imine) crosslinks (Fig. 29.4). The initial products may rearrange, or be dehydrated, or reduced to form stable crosslinks, such as **lysinonorleucine**. Studies with β-aminopropionitrile, which inhibits the enzyme lysyl oxidase, have illustrated that collagen crosslink formation is a major determinant of tissue mechanical properties and strength.

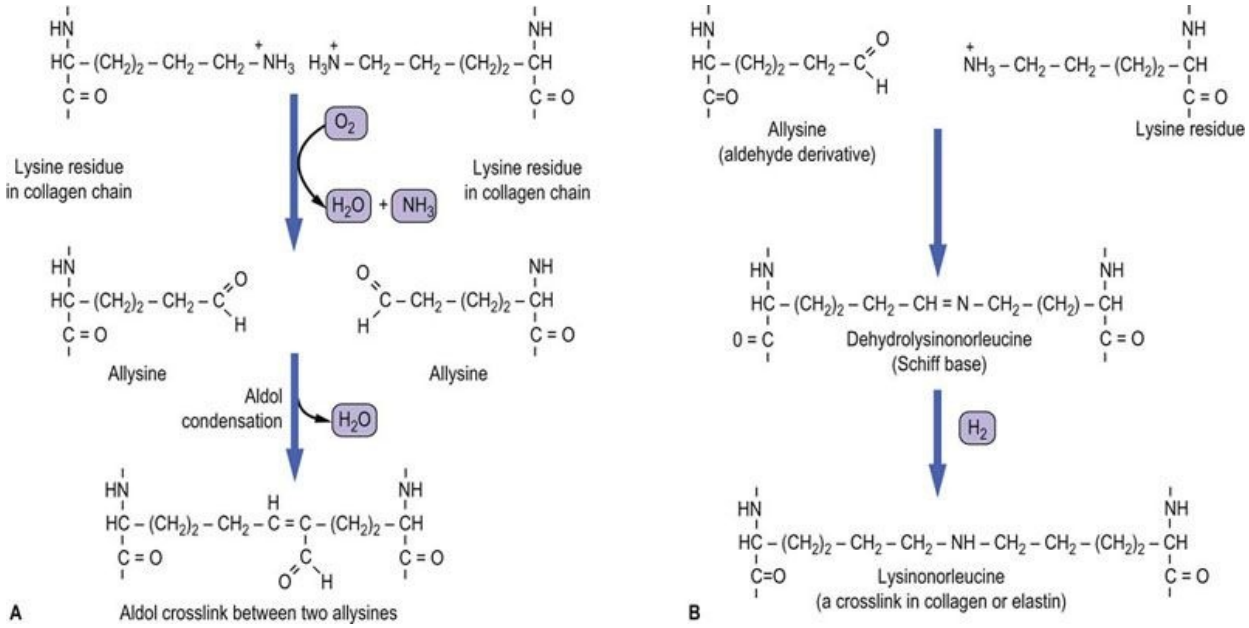


FIG. 29.4 Collagen crosslink formation.

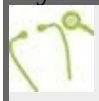
Allysine (and hydroxyallysine) are precursors of collagen crosslink formation by **(A)** aldol condensation and **(B)** Schiff base (imine) intermediates.

Noncollagenous proteins in the extracellular matrix

Elastin

Weak hydrophobic interactions between valine residues permit the flexibility and extensibility of elastin

The flexibility required for function of blood vessels, lungs, ligaments and skin is imparted by a network of elastic fibers in the ECM of these tissues. The predominant protein of elastic fibers is elastin. Unlike the multigene collagen family, there is only one gene for elastin, coding for a polypeptide about 750 amino acids long. In common with collagens, it is rich in glycine and proline residues but elastin is more hydrophobic: one in seven of its amino acids is a valine. Unlike collagens, elastin contains little hydroxyproline and no hydroxylysine or carbohydrate chains, and does not have a regular secondary structure. Its primary structure consists of alternating hydrophilic and hydrophobic lysine and valine-rich domains. The lysines are involved in intermolecular crosslinking, while the weak interactions between valine residues in the hydrophobic domains impart elasticity to the molecule.



Clinical box Marfan syndrome: result from mutations in fibrillin gene

The ultrastructure of elastic fibers reveals elastin as an insoluble, polymeric, amorphous core covered with a sheath of microfibrils that contribute to the stability of the elastic fiber. The predominant constituent of microfibrils is the glycoprotein fibrillin. Marfan syndrome is a relatively rare genetic disease of connective tissues caused by mutations in the fibrillin gene (frequency: 1 in 10,000 births). People with this disease have typically tall stature, long arms and legs, and arachnodactyly (long, ‘spidery’ fingers). The disease in a mild form causes loose joints, deformed spine, floppy mitral valves (leading to cardiac regurgitation), and eye problems such as lens dislocation. In severely affected individuals, the aortic wall is prone to rupture because of defects in elastic fiber

formation.

The soluble monomeric form of elastin initially synthesized on the RER is called tropoelastin. Except for some hydroxylation of proline, tropoelastin does not undergo post-translational modification. During the assembly process in the extracellular space, lysyl oxidase generates allysine in specific sequences: -Lys-Ala-Ala-Lys- and -Lys-Ala-Ala-Ala-Lys-. As with collagen, the reactive aldehyde of allysine condenses with other allysines or with unmodified lysines. Allysine and dehydrolysinonorleucine on different tropoelastin chains also condense to form pyridinium crosslinks – heterocyclic structures known as **desmosine** or **isodesmosine** (Fig. 29.5). Because of the way in which elastin monomers are crosslinked in polymers, elastin can stretch in two dimensions.

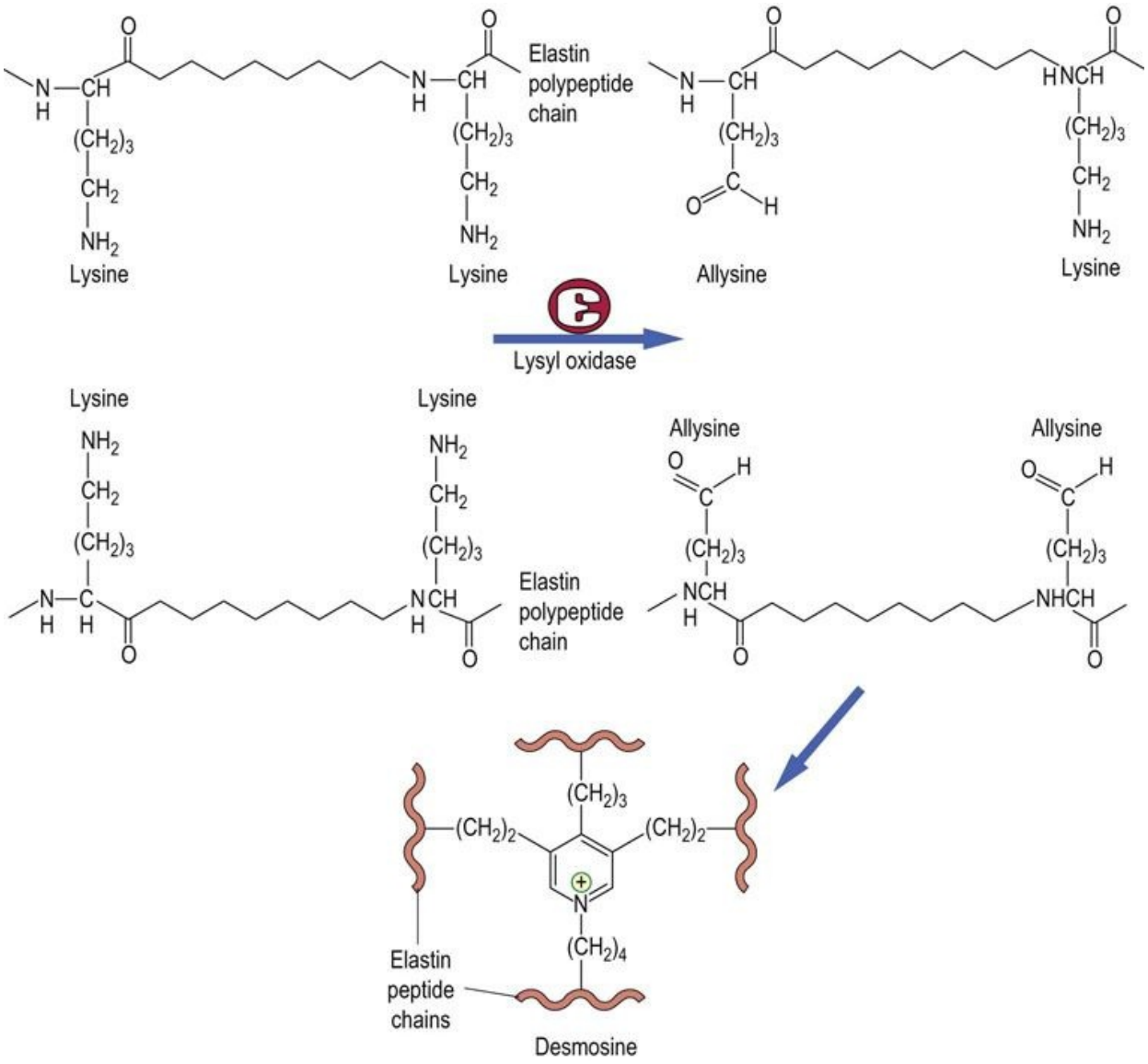


FIG. 29.5 Desmosine – a multichain crosslink in elastin.
Allysine and dehydrolysinonorleucine residues in adjacent elastin chains react to form the three-dimensional elastic polymer, crosslinked by desmosine.

Fibronectin

Fibronectin and laminin have multiple binding sites for ECM proteins and proteoglycans

Fibronectin is a glycoprotein present as a structural component of the ECM and also in plasma as a soluble protein. Fibronectin is a dimer of two identical

subunits, each of 230 kDa, joined by a pair of disulfide bonds at their *C*-terminals. Each subunit is organized into domains, known as type I, II, and III domains, and each of these has several homologous repeating units or modules in its primary structure: there are 12 type I repeats, two type II repeats, and 15–17 type III repeats. Each module is independently folded, forming a 'string of beads' type of structure. At least 20 different tissue-specific isoforms of fibronectin have been identified, all produced by alternative splicing of a single precursor messenger ribonucleic acid (mRNA). The alternative splicing is regulated not only in a tissue-specific manner but also during embryogenesis, wound healing, and oncogenesis. Plasma fibronectin, secreted mainly by liver cells, lacks two of the type III repeats that are found in cell-and matrix-associated forms of fibronectin. Because of its multidomain structure and its ability to interact with cells and with other ECM components, alterations in fibronectin expression affect cell adhesion and migration, embryonic morphogenesis, and cytoskeletal and ECM organization.

Functional domains in fibronectin have been identified by their binding affinity for other ECM components, including collagen, heparin, fibrin, and the cell surface. The type I modules interact with fibrin, heparin and collagen, type II modules have collagen-binding domains, and type III modules are involved in binding to heparin and the cell surface. The specific interactions have been further mapped to short stretches of amino acids. A short peptide containing **Arg-Gly-Asp (RGD)**, present in the tenth type III repeat of fibronectin, binds to the integrin family of proteins present on cell surfaces; this sequence is not unique to fibronectin but is also found in other proteins in the ECM. Another sequence, Pro-X-Ser-Arg-Asn (PXSRN), present in the ninth type III repeat, is implicated in integrin-mediated cell attachment. The loss of fibronectin from the surface of many tumor cells may contribute to their release into the circulation and penetration through the ECM, one of the first steps in tumor metastasis.

Laminins

Laminins are a family of noncollagenous glycoproteins found in basement membranes and expressed in variant forms in different tissues. They are large (850 kDa), heterotrimeric molecules, composed of α, β and γ chains. To date, five α, four β and three γ chains have been identified which can associate to produce at least 15 different laminin variants. The three interacting chains in a heterotrimer are arranged in an asymmetric cruciform or cross-shaped molecule,

held together by disulfide linkages. Laminins undergo reversible self-assembly in the presence of calcium to form polymers, contributing to the elaborate mesh-like network in the basement membrane. Biochemical and electron microscopic studies indicate that all full-length short arms of laminin are required for self-assembly and that the polymer is formed by joining the ends of the short arms. Like fibronectin, laminins interact with cells through multiple binding sites in several domains of the molecule. The α chains have binding sites for integrins and heparan sulfate (below). Laminin polymers are also connected to type IV collagen by a single-chain protein, **nidogen/entactin**, which has a binding site for collagen and, in common with fibronectin, also has an RGD sequence for integrin binding. Nidogen also binds to the core proteins of proteoglycans (below). It has a central role in formation of crosslinks between laminin and type IV collagen, generating a scaffold for anchoring of cells and ECM molecules in the basement membrane.

Clinical box Muscular dystrophies

Muscular dystrophies are a heterogeneous group of genetic disorders that result in progressive decline in muscle strength and structure. To date, mutations have been identified in more than 30 genes that result in muscular dystrophies. Many of the identified gene products are components of the ECM–cell surface–cytoskeletal complex of muscle cells. In particular, one class of muscular dystrophy is caused by mutations in the α2 chain of laminin-2. These mutations prevent normal polymer formation of laminin-2 and result in abnormal basement membrane organization surrounding skeletal muscle fibers of patients with this muscular dystrophy.

Clinical box Epidermolysis bullosa

Epidermolysis bullosa is a rare heritable disorder characterized by severe blistering of the skin and epithelial tissue. Three kinds are

known:

- **simplex:** blistering in the epidermis, caused by defects in keratin filaments
- **junctional:** blistering in the dermal–epidermal junction, caused by defects in laminin
- **dystrophic:** blistering in the dermis, caused by mutations in the gene encoding type VII collagen.

Epidermolysis bullosa illustrates the multifactorial nature of connective tissue diseases that have similar clinical features.

Proteoglycans

Proteoglycans are gel-forming components of the ECM and comprise what has classically been called the ‘ground substance’. Some proteoglycans are located on the cell surface, where they bind growth factors and other ECM components. They are composed of peptide chains containing covalently bound sugars. However, the peptide chains of proteoglycans are usually more rigid and extended than the protein portion of the glycoproteins, and the proteoglycans contain much larger amounts of carbohydrate – typically >95% carbohydrate. The sugar chains are linear, unbranched oligosaccharides that are much longer than those of the glycoproteins, and may contain more than 100 sugar residues in a chain. Furthermore, the oligosaccharide chains of proteoglycans have a repeating disaccharide unit, usually composed of a uronic acid and an amino sugar. Proteoglycan oligosaccharide chains are polyanionic because of the many negative charges of the carboxyl groups of the uronic acids, and from sulfate groups attached to some of the hydroxyl or amino groups of the sugars.

Structure of proteoglycans

Glycosaminoglycans are the polysaccharide components of proteoglycans

The general structures of the glycosaminoglycans (GAGs) are shown in Table 29.2. The **disaccharide repeat** is different for each type of GAG, but is usually composed of a hexosamine and a uronic acid residue, except in the case of keratan sulfate, in which the uronic acid is replaced by galactose. The amino sugar in GAGs is either glucosamine ($GlcNH_2$) or galactosamine ($GalNH_2$), both of which are present mostly in their *N*-acetylated forms (GlcNAc and GalNAc), although in some of the GAGs (heparin, heparan sulfate) the amino group is sulfated rather than acetylated. The uronic acid is usually D-glucuronic acid (GlcUA) but in some cases (dermatan sulfate, heparin) it may be **L-iduronic acid (IdUA)**. With the exception of hyaluronic acid and keratan sulfate, all the GAGs are attached to protein by a core trisaccharide, Gal-Gal-Xyl; the xylose is linked to a serine or threonine residue of a core protein. Keratan sulfate is also attached to protein, but in that case the linkage is either through an *N*-linked oligosaccharide (keratan sulfate I) or an *O*-linked oligosaccharide (keratan

sulfate II). Hyaluronic acid, which has the longest polysaccharide chains, is the only GAG that does not appear to be attached to a core protein.

Table 29.2

Structure and distribution of the proteoglycans

Proteoglycan	Characteristic disaccharide	Sulfation	Tissue location
Hyaluronic acid	[4GlcUAβ1–3GlcNAcβ1]	None	Joint and ocular fluids
Chondroitin sulfates	[4GlcUAβ1–3GalNAcβ1]	GalNAc	Cartilage, tendons, bone
Dermatan sulfate	[4IdUAα1–3GalNAcβ1]	IdUA, GalNAc	Skin, valves, blood vessels
Heparan sulfate	[4IdUAα1–4GlcNAcβ1]	GlcNAc	Cell surfaces
Heparin	[4IdUAα1–4GlcNAcβ1]	$GlcNH_2$, IdUA	Mast cells, liver
Keratan sulfates	[3Galβ1–4GlcNAcβ1]	GlcNAc	Cartilage, cornea

GalNAc, N-acetylgalactosamine; $GlcNH_2$, glucosamine; GlcUA, D-glucuronic acid; IdUA, L-iduronic acid.

Hyaluronic acid

Hyaluronic acid, the only nonsulfated glycosaminoglycan, has a unique role in proteoglycan assembly

Hyaluronic acid is composed of repeating units of GlcUA and GlcNAc. This polysaccharide chain is the longest of the GAGs, with molecular weight of 10^5–10^7 Da (250–25,000 repeating disaccharide units), and is the only nonsulfated GAG.

The chondroitin sulfates

The chondroitin sulfates are major components of cartilage. They contain GalNAc rather than GlcNAc as the amino sugar, and their polysaccharide chains are shorter: 2–5 × 10^5 Da. The chondroitin chains are attached to protein via the trisaccharide linkage region (Gal-Gal-Xyl), and they contain sulfate residues linked to either the 4-or 6-hydroxyl groups of GalNAc.

Dermatan sulfate

Dermatan sulfate was originally isolated from skin but is also found in blood vessels, tendon and heart valves. This GAG is similar in structure to chondroitin sulfate but has a variable amount of L-iduronic acid (IdUA), the *C*-5-epimer of D-GlcUA, formed in an unusual reaction by epimerization of GlcUA after it has been incorporated into the polymer. Dermatan sulfate has a higher charge density than the chondroitin sulfates, as it contains sulfate residues on the *C*-2 position of some IdUA residues, and on the 4-hydroxyl groups of GalNAc.

Heparin and heparan sulfate

Heparin is a small, highly charged GAG with strong anticoagulant activity

Heparin and heparan sulfate consist primarily of repeating disaccharide units of $GlcNH_2$ with IdUA or GlcUA, respectively. The linkage between the amino sugar and the uronic acid is uniformly 1–4, rather than the alternating 1–4/1–3 linkages seen in other GAGs. Most of the $GlcNH_2$ units of heparin are *N*-sulfated, whereas many of the IdUA residues are sulfated at the *C*-2 hydroxyl group, and the $GlcNH_2$ residues at the *C*-6 hydroxyl group. Heparin and heparan sulfate are the most highly charged of the GAGs. Although the structures of these two polymers are closely related, their distribution in the body and their functions are quite different: heparin is a small microheterogeneous molecule (~3000–30,000 Da), found intracellularly as a proteoglycan. It is released into the extracellular space as a free polysaccharide (GAG) and has strong anticoagulant activity (Chapter 7). In contrast, heparan sulfate is bound in the ECM or on the surface of cells, and has only weak anticoagulant activity.

Keratan sulfate

The final GAG structure shown in Table 29.2 is keratan sulfate (KS). This is a rather unusual GAG because it is linked to protein by either an *N*-linked (KS I) or an *O*-linked (KS II) oligosaccharide. Thus it has features common to both proteoglycans and glycoproteins. It is considered to be a proteoglycan, however, because the glycan portion has a repeating disaccharide unit and a long, linear chain. The repeating unit is composed of GlcNAc and galactose, instead of the uronic acid. Both the GlcNAc and the galactose are generally sulfated on the *C*-6 hydroxyl groups.

Synthesis and degradation of proteoglycans

The structure of glycosaminoglycans is determined by the cell's complement of glycosyl and sulfotransferases

Proteoglycans are synthesized by a series of glycosyl transferases, epimerases and sulfotransferases, beginning with the synthesis of the **core trisaccharide (Xyl→Gal→Gal)** while the protein is still in the RER. Synthesis of the repeating oligosaccharide and other modifications take place in the Golgi apparatus. As with the synthesis of glycoproteins and glycolipids, separate enzymes are involved in individual steps. For example, there are separate galactosyl transferases for each of the galactose units in the core, a separate GlcUA transferase for the core and repeating disaccharides, and separate sulfotransferases for the *C*-4 and *C*-6 positions of the GalNAc residues of chondroitin sulfates. Phosphoadenosine phosphosulfate (PAPS) is the sulfate donor for the sulfotransferases. These pathways are illustrated in Figure 29.6, for chondroitin-6-sulfate.

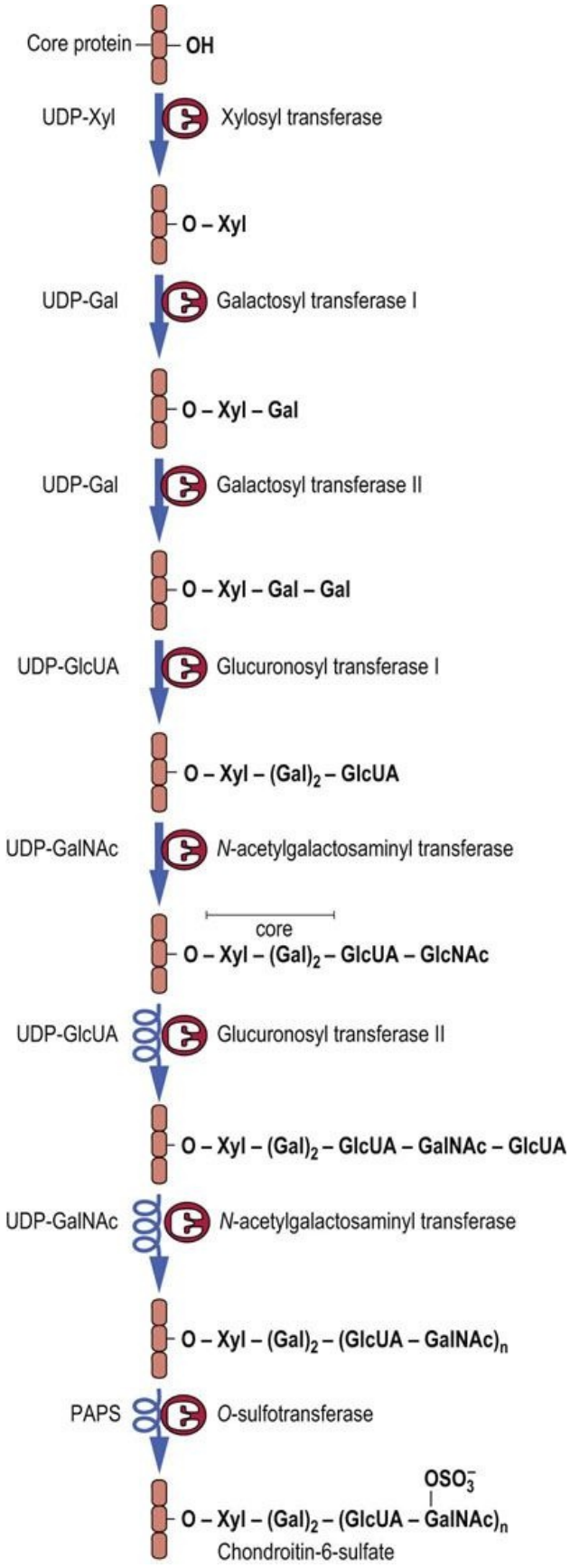

Core protein
OH
UDP-Xyl
Xylosyl transferase
O – Xyl
UDP-Gal
Galactosyl transferase I
O – Xyl – Gal
UDP-Gal
Galactosyl transferase II
O – Xyl – Gal – Gal
UDP-GlcUA
Glucuronosyl transferase I
O – Xyl – (Gal)2 – GlcUA
UDP-GalNAc
N-acetylgalactosaminyl transferase
core
O – Xyl – (Gal)2 – GlcUA – GlcNAc
UDP-GlcUA
Glucuronosyl transferase II
O – Xyl – (Gal)2 – GlcUA – GalNAc – GlcUA
UDP-GalNAc
N-acetylgalactosaminyl transferase
O – Xyl – (Gal)2 – (GlcUA – GalNAc)n
PAPS
O-sulfotransferase
OSO3−
O – Xyl – (Gal)2 – (GlcUA – GalNAc)n
Chondroitin-6-sulfate

FIG. 29.6 Synthesis of the proteoglycan, chondroitin-6-sulfate. Several enzymes participate in this pathway. Xyl, xylose.

Defects of proteoglycan degradation lead to mucopolysaccharidoses

The degradation of proteoglycans occurs in lysosomes. The protein portion is degraded by lysosomal proteases and the GAG chains are degraded by the sequential action of a number of different lysosomal acid hydrolases. The stepwise degradation of GAGs involves exoglycosidases and sulfatases, beginning from the external end of the glycan chain. This may involve the removal of sulfate by a sulfatase, then removal of the terminal sugar by a specific glycosidase, and so on. Figure 29.7 shows the steps in the degradation of heparan sulfate. As with degradation of glycosphingolipids, if one of the enzymes involved in the stepwise pathway is missing, the entire degradation process is halted at that point and the undegraded molecules accumulate in the lysosome. The lysosomal storage diseases resulting from accumulation of GAGs are known as **mucopolysaccharidoses** (Table 29.3), because of the original designation of GAGs as mucopolysaccharides. There are more than a dozen such mucopolysaccharidoses, resulting from defects in degradation of GAGs. In general, these diseases can be diagnosed by the identification of specific GAG chains in the urine, followed by assay of the specific hydrolases in leukocytes or fibroblasts.

Table 29.3

Enzymatic defects characteristic of various mucopolysaccharidoses

Syndrome	Deficient enzyme	Product accumulated in lysosomes and secreted in urine
Hunter's	Iduronate sulfatase	Heparan and dermatan sulfate
Hurler's	α-Iduronidase	Heparan and dermatan sulfate
Morquio's A	Galactose-6-sulfatase	Keratan sulfate
Morquio's B	β-Galactosidase	Keratan sulfate
Sanfilippo's A	Heparan sulfamidase	Heparan sulfate
Sanfilippo's B	*N*-Acetylglucosaminidase	
Sanfilippo's C	*N*-Acetylglucosamine-6-sulfatase	

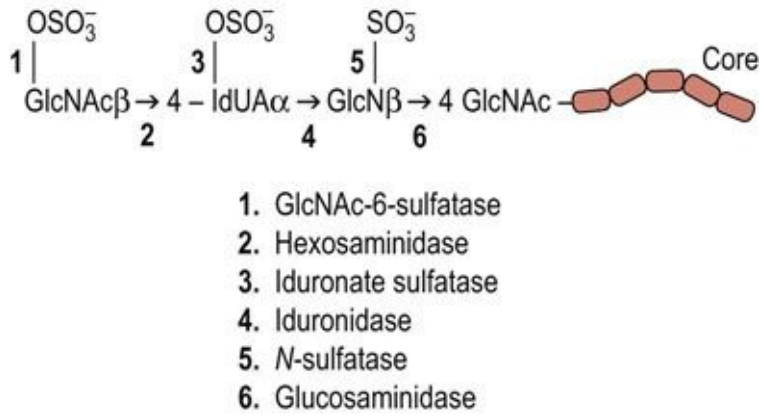


FIG. 29.7 Degradation of heparan sulfate.
This proceeds by a defined sequence of lysosomal hydrolase activities.

Advanced concept box
Mechanisms of the anticoagulant effect of heparin

Heparin is a heterogeneous (3000–30,000 kDa), polyanionic oligosaccharide activator of antithrombin III (AT) (Chapter 7). AT is a slow but quantitatively important inhibitor of thrombin (factor X) and other factors (IX, XI, XII) in the blood-clotting cascade. When heparin binds to AT, it converts AT from a slow inhibitor to a rapid inhibitor of coagulating enzymes. Heparin interacts with a lysine residue in AT and induces a conformational change that promotes covalent binding of AT to the active serine centers of coagulating enzymes, inhibiting their procoagulant activity. Heparin then dissociates from the ternary complex and can be recycled for anticoagulation.

The smallest, most active component of heparin is a pentasaccharide [GlcN-(*N*-sulfate-6-*O*-sulfate)-α1,4-GlcUA-β1,4-GlcN-(*N*-sulfate-3,6-di-*O*-sulfate)-α1,4-IdUA-(2-*O*-sulfate)-α-1,4-GlcN-(*N*-sulfate-6-*O*-sulfate], which has a K_d of ~10 μM for binding to ATlll. Heparin has an average half-life of 30 min in the circulation, so that it is commonly administered by infusion. Heparin does not have fibrinolytic activity; therefore, it will not lyse existing clots. In addition to its anticoagulant activity, heparin also releases several enzymes from proteoglycan binding sites on

the vascular wall, including lipoprotein lipase, which is often assayed as heparin-releasable plasma lipoprotein lipase activity or postheparin lipase. Lipoprotein lipase is inducible by insulin, and decreased activity of this enzyme delays plasma clearance of chylomicrons and VLDL, contributing to hypertriglyceridemia in diabetes (Chapter 18).

Functions of the proteoglycans

Bottlebrushes, silly putty and reinforced concrete

Proteoglycans are found in association with most tissues and cells. One of their major roles is to provide structural support to tissues, especially cartilage and connective tissue. In cartilage, large aggregates, composed of chondroitin sulfate and keratan sulfate chains linked to their core proteins, are noncovalently associated with hyaluronic acid via **link proteins**, forming a jelly-like matrix in which the collagen fibers are embedded. This macromolecule, a **'bottlebrush' structure** known as aggrecan (Fig. 29.8), provides both rigidity and stability to connective tissue. Because of their negative charge, the GAGs bind large amounts of monovalent and divalent cations: a cartilage proteoglycan molecule of 2×10^6 Da would have an aggregate negative charge of about 10,000. The maintenance of electrical neutrality consequently requires a high concentration of counter-ions. These ions draw water into the ECM, causing swelling and stiffening of the matrix, the result of tension between osmotic forces and binding interactions between proteoglycans and collagen. The structure and hydration of the ECM allow for a degree of rigidity, combined with flexibility and compressibility, enabling the tissue to withstand torsion and shock. The hyaluronic acid–proteoglycan–collagen aggregates in vertebral and articular disks have some of the viscoelastic properties of 'silly putty', bounce plus resilience, cushioning the impact between bones. These disks compress during the course of the day, expand elastically during the night, and deform gradually with age.

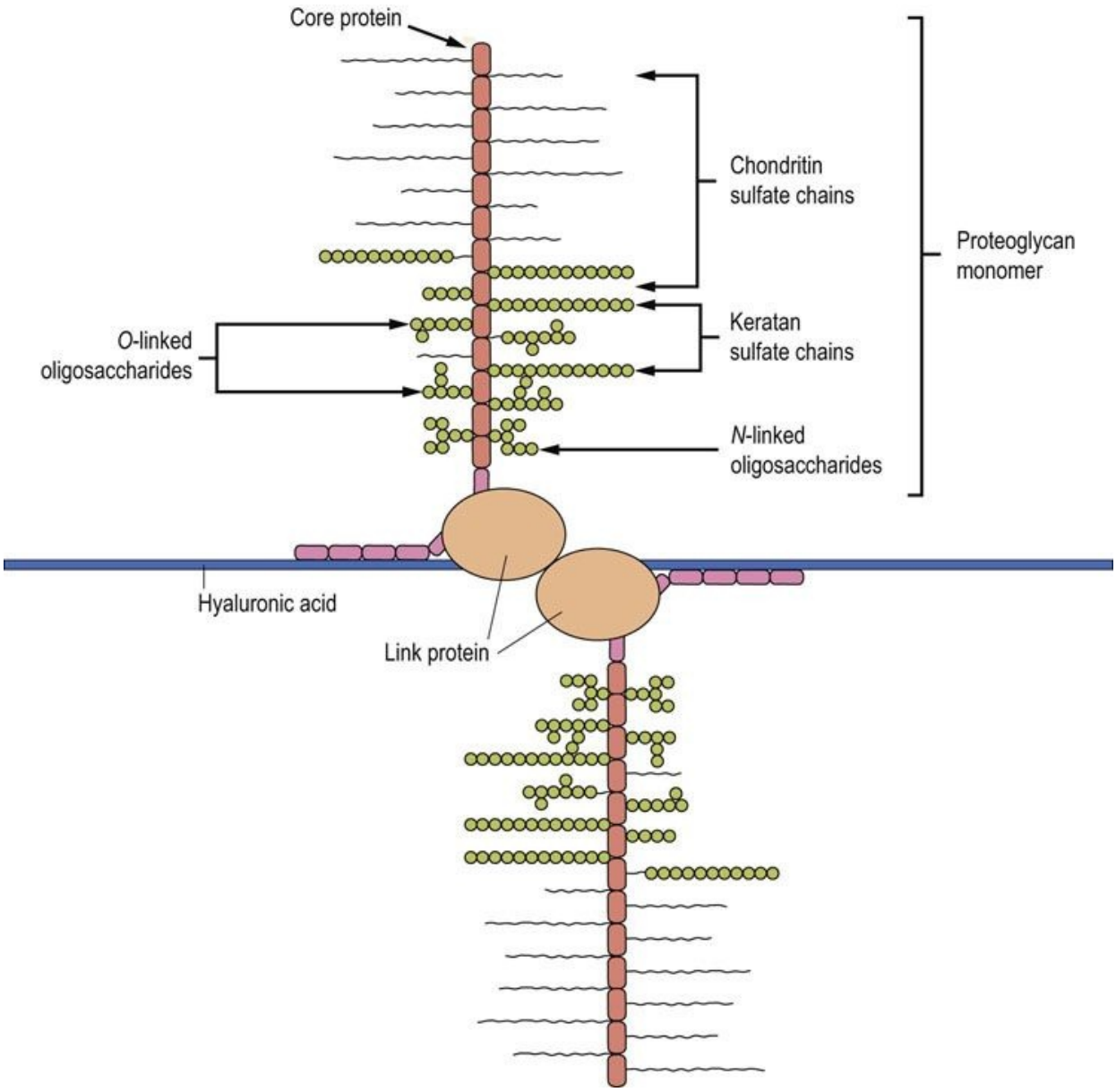


FIG. 29.8 Structure of aggrecan.
Associations between proteoglycans and hyaluronic acid form an **aggrecan** structure in the extracellular matrix (ECM). The extension of this structure yields a three-dimensional array of proteoglycans bound to hyaluronic acid, which creates a stiff matrix or 'bottlebrush' structure in which collagen and other ECM components are embedded.

The overall structure of cartilage can be likened to that of the vertical reinforced concrete slabs poured during the construction of large buildings, in which steel rods (collagen fibers) are embedded in an amorphous layer of cement (the proteoglycan aggregates). Collagen stabilizes the network of proteoglycans in cartilage in much the same way that the reinforcing rods in the

concrete provide structural strength for the cement walls. The structure of earthquake-resistant buildings, like the ECM, provides a balance between integrity and flexibility.

Although the amounts involved are low compared with those in skin and cartilage, organs such as the liver, brain, or kidney also contain a variety of proteoglycans:

- **Liver:** heparan sulfate is the principal GAG; it is present both intracellularly and on the cell surface of the hepatocyte, and the attachment of hepatocytes to their substratum in cell culture is mediated, in part, by this proteoglycan.
- **Kidney:** changes in both the collagen and proteoglycan content of the renal basement membrane are associated with diabetic renal disease. In this case, the change in structure and charge of the proteoglycan aggregate, known as perlecan, is associated with a change in the filtration selectivity of the glomerulus (Chapter 23).
- **Cornea:** two populations of proteoglycans have been identified in the cornea, one containing keratan sulfate and the other dermatan sulfate. These molecules have a much smaller hydrodynamic size than the large cartilage proteoglycans, which may be required for interaction of the corneal proteoglycans with the tightly packed and oriented collagen fibers in this transparent tissue. Corneal clouding in macular corneal dystrophy is associated with undersulfation of keratan sulfate I proteoglycan.

Some proteoglycans or GAGs, especially heparin and heparan sulfate, have important physiologic roles in binding proteins or other macromolecules. Heparin serves as an intracellular binding site for proteinases located in secretory granules of mast cells. Several proteoglycans are involved in binding of proteins and enzymes to the vascular wall. They may also function in the vascular wall to inhibit clot formation by activation of antithrombin III (Chapter 7).

Communication of cells with the extracellular matrix

Integrins are plasma membrane proteins that bind to and transmit mechanical signals between the ECM and intracellular proteins

Interactions between cells and the ECM regulate a wide variety of cellular processes including proliferation, migration, differentiation and even survival. Several cell surface receptors have been indentified that mediate these interactions including integrins, discoidin domain receptors, dystroglycan and others. Of these, the **integrins** appear to be the most ubiquitous form of ECM receptors. Integrins are widely expressed throughout the animal kingdom from sponges to humans. They are heterodimers of α and β chains that have been loosely grouped into subfamilies based upon the component β chain. To date, 18 α and eight β chains have been indentified in mammals. Through various combinations of α and β chains, over 20 different functional integrin heterodimers have been described. The specific combination of α and β chains dictates the specific ECM ligand for a particular integrin heterodimer. However, multiple integrin heterodimers can bind to some ECM components. For instance, $\alpha_4\beta_1$, $\alpha_5\beta_1$ and $\alpha_v\beta_3$ all interact with fibronectin. Adding to this complexity, several integrin heterodimers bind to multiple ECM components. For instance, $\alpha_v\beta_3$, which was originally described as a vitronectin receptor, can interact with not only vitronectin but also with fibronectin, fibrinogen and osteopontin as well.

In a functional integrin, both the α and β chains span the cell membrane (Fig. 29.9). Typically, each chain has a large extracellular domain, a single transmembrane domain and a short cytoplasmic tail. An exception to this is the β_4 protein, which has an exceptionally long cytoplasmic tail of over 1000 amino acids. The extracellular region of the integrin heterodimer interacts with ECM components in a divalent cation-dependent manner. The integrins are in an optimal position to transmit physical or mechanical signals from the ECM to the interior of the cell. These physical signals can be further distributed through the cell via the actin-containing cytoskeleton and ultimately modulate gene expression in the nucleus. This transmission of physical signals via the ECM–

integrin–cytoskeletal axis has been extensively investigated and termed '**tensegrity**'. Physical signals from the ECM can also be transduced into biochemical events in the cytoplasm of the cell via integrins. Unlike some other types of receptors, integrins do not themselves possess enzymatic activity. However, integrins associate with a number of cytoplasmic protein kinases including focal adhesion kinase (FAK) and Src. Activation of integrins initiates enzymatic cascades via these associated kinases that ultimately leads to changes in cell behavior and gene expression.

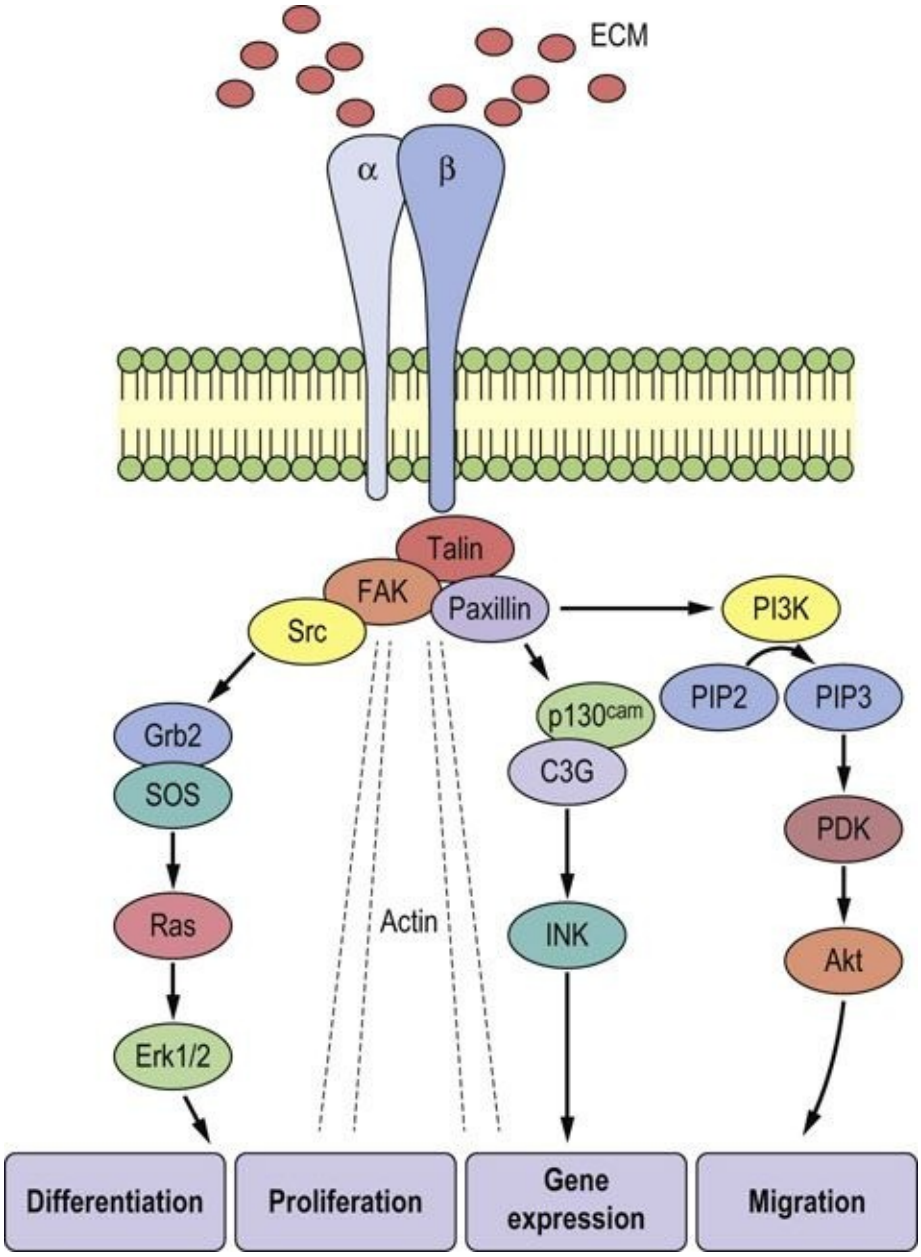


FIG. 29.9 Organization of integrins.
The α and β chains span the cell membrane, interacting with the ECM outside the cell and the cytoskeleton and signaling molecules inside. In this manner, integrins can transduce signals from the ECM into biochemical and mechanical events in the cytoplasm that ultimately lead to alterations in cell morphology and function. The ovals contain abbreviations for components of the complex signaling cascade that conveys information from the integrin molecule to the nucleus of the cell.

Advanced concept box Matrix

remodeling

The ECM is in a constant state of synthesis and degradation, repair and remodeling: for example, during cell migration, morphogenesis, angiogenesis, and in response to inflammation and injury. ECM turnover is mediated by a family of **matrix metalloproteinases (MMPs)**, about 30 zinc endoproteinases with specificity for different components of the matrix. The MMP family includes collagenases, stromelysins, matrilysins and elastases; these enzymes, with broad substrate specificities, catalyze degradation of collagen, aggrecan and accessory ECM proteins, such as fibronectin and laminin.

MMPs may be integral plasma membrane proteins, may be bound to the plasma membrane by a glycosylphosphatidylinositol (GPI) glycan anchor (Chapter 28), or secreted into the extracellular space; they exist as zymogens until activated locally by proteolytic cleavage in response to cellular signals or extracellular enzymes, such as thrombin and plasmin, activated during blood clotting and fibrinolysis. As with the cascade of protease reactions involved in blood coagulation, there are also tissue inhibitors of MMPs, known as TIMPs, a family of four proteins that inactivate MMPs and limit the spread of damage. The balance between the activation and inhibition of MMPs is critical to the integrity and function of the ECM; alterations in MMP activity are associated with skeletal dysplasias, coronary artery disease, arthritis and metastasis.

Advanced concept box

Extracellular matrix and tissue engineering

Over the past decade, the interest in producing replacement tissues through tissue engineering has grown considerably. The ultimate goal of tissue engineering is to combine appropriate cells and biomaterials to produce tissue equivalents that mimic normal

tissues and organs and can replace damaged or diseased tissues. As the biological and mechanical properties of tissues are determined in part by the heterogeneous composition and organization of the ECM, the successful generation of tissue equivalents will require the development of appropriate three-dimensional ECM scaffolds.

An attractive therapeutic approach is to combine undifferentiated stem cells with appropriate scaffolds and biochemical factors to promote differentiation of the cells along particular lines, depending upon the specific replacement tissue desired. Properties of the ECM scaffold, including ECM composition, porosity and mechanical properties, have important effects on stem cell differentiation. Culture of mesenchymal stem cells in scaffolds of relatively high stiffness tends to promote the formation of bone-like tissue and the formation of osteoblasts, while culture of the same stem cells in less stiff scaffolds results in formation of cartilage cells or chondroblasts. These and other studies illustrate that physical and mechanical cues from the ECM are important in regulating differentiation of stem cells. Advances in tissue engineering and production of replacement tissues will require a thorough understanding of the normal and pathologic ECM.

Summary

■ The ECM contains a complex array of fibrillar and network-forming collagens, elastin fibers, a stiff gelatinous matrix of proteoglycans, and a number of glycoproteins that mediate the interaction of these molecules with one another and with the cell surface.

■ Interactions between ECM components afford structure, stability, and elasticity to the ECM, and provide a route for communication between the intra- and extracellular environments in tissues.

■ The heterogeneity of both the protein and the carbohydrate components of the ECM provides for great diversity in the structure and function of the ECM in various tissues.

Active learning

1. Compare the structure of heparin, its mechanism of action, its route and frequency of administration to that of other common anticoagulants, such as aspirin and coumarin derivatives.
2. Discuss factors that promote the turnover of ECM components, as part of normal growth and development and in diseases such as rheumatoid arthritis.
3. Review the consequences of genetic defects in sulfation of proteoglycans.

Further reading

Couchman, JR, Pataki, CA. An introduction to proteoglycans and their localization. *J Histochem Cytochem*. 2012; 60:885–897.

Hanein, D, Horwitz, AR. The structure of cell-matrix adhesions: the new frontier. *Curr Opin Cell Biol*. 2012; 24:134–140.

Ingber, DE. Tensegrity-based mechanosensing from macro to micro. *Prog Biophys Mol Biol*. 2008; 97:163–179.

Kadler, KE, Baldock, C, Bella, J, Boot-Handford, RP. Collagens at a glance. *J Cell Science*. 2007; 120:1955–1958.

Kock, L, van Donkelaar, CC, Ito, K. Tissue engineering of functional articular cartilage: the current status. *Cell Tissue Res*. 2012; 347:613–627.

Murphy, G. Tissue inhibitors of metalloproteinases. *Genome Biol*. 2011; 12:233.

Stupack, DG. The biology of integrins. *Oncology*. 2007; 21:6–12.

Yamauchi, M, Sricholpech, M. Lysine post-translational modifications of collagen. *Essays Biochem*. 2012; 52:113–133.

Websites

Collagen in disease. www.dupuytrenfoundation.org.

Marfan's syndrome. www.marfan.org.

Mucopolysaccharidoses. www.ninds.nih.gov/health.

CHAPTER 30

Role of the Liver in Metabolism

Alan F. Jones

Learning objectives

After reading this chapter you should be able to:

- Discuss the participation of the liver in carbohydrate metabolism, and in particular its role in endogenous glucose production.
- Discuss the role of the liver in lipid metabolism.
- Outline changes in the hepatic protein synthesis that take place during the acute phase reaction.
- Describe ubiquitin-mediated mechanisms of proteolysis.
- Describe the pathway of heme synthesis.
- Describe the metabolism of bilirubin and the main types of jaundice.
- Comment on the mechanism of hepatotoxicity of drugs and alcohol.

Introduction

The liver has a central role in metabolism, because of both its anatomic placement and its many biochemical functions

The liver receives venous blood from the intestine: thus all of the products of digestion, ingested drugs and other xenobiotics reach it and may be further metabolized before entering the systemic circulation. The hepatic parenchymal cells, the **hepatocytes,** have an immensely broad range of synthetic and catabolic functions, which are summarized in Table 30.1.

Table 30.1

Functions of hepatic parenchymal cells and their disturbances in liver disease

Function	Markers of impairment in plasma
Heme catabolism	bilirubin
Carbohydrate metabolism	↓glucose
Protein synthesis	↓albumin
	prolonged prothrombin time
Protein catabolism	↑ammonia
	↓urea
Lipid metabolism	↑cholesterol
	↑triglycerides
Drug metabolism	biological half-time of a drug
Bile acid metabolism	↑bile acids

The liver plays important roles in carbohydrate, lipid and amino acid metabolism, in the synthesis and breakdown of plasma proteins, and in the storage of vitamins and metals. It also has the ability to metabolize, and so detoxify, an infinitely wide range of xenobiotics. The liver also has an excretory function, in which metabolic waste products are secreted into a branching system of ducts known as the **biliary tree**, which in turn drains into the duodenum; the biliary constituents are then excreted in feces.

The liver is the largest organ in the body and has a substantial reserve metabolic capacity

Mild disease may cause no symptoms and is detected only by biochemical changes in the blood. However, the patient with severe liver disease has a yellow

pigmentation of the skin (**jaundice**), bruises readily and may **bleed profusely**, has an abdomen distended with fluid (**ascites**), and may be confused or unconscious (**hepatic encephalopathy**) (Fig. 30.1). This chapter will describe the specialized metabolic functions of the liver and the abnormalities that occur in liver disease.

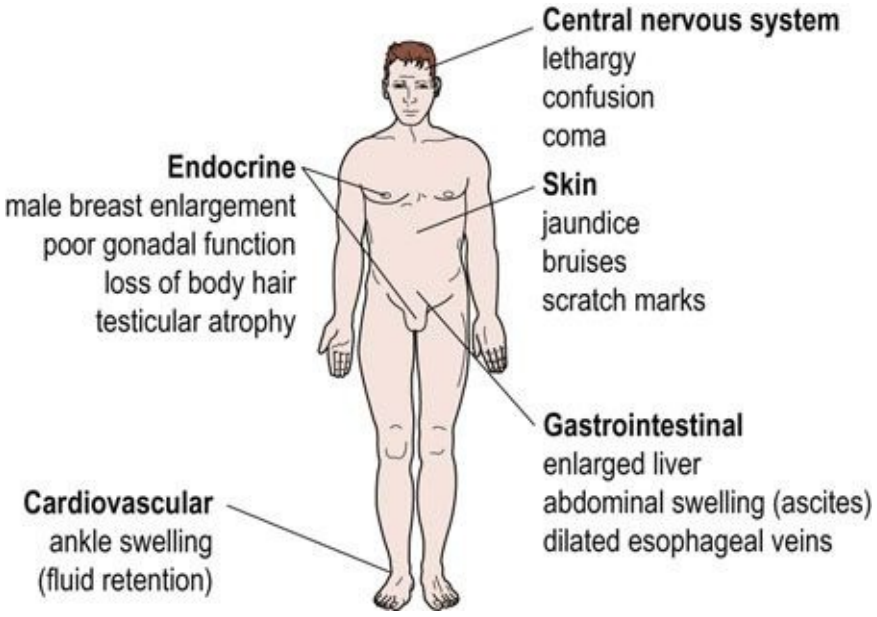


FIG. 30.1 Clinical features of severe liver disease.

Structure of the liver

The liver is the largest solid organ in the body and, in adults, weighs about 1500 g. Approximately 75% of its blood flow is supplied by the portal vein, which arises from the intestine. The hepatic artery supplies the remainder of its blood. Blood leaving the liver enters the systemic venous system in the hepatic vein. The biliary component of the liver comprises the gallbladder and bile ducts.

Structure of the liver facilitates exchange of metabolites between hepatocytes and plasma

Under the microscope, the substance of the liver is composed of a very large number of hepatocytes arranged in polyhedral lobules (Fig. 30.2). Portal tracts at the ‘corners’ of these polyhedrons contain branches of the portal vein, hepatic artery and interlobular bile ducts. Blood sinusoids arise from the terminal branches of the portal vein and interconnect and interweave through the hepatocytes before joining the central lobular vein, which in turn eventually flows into the hepatic vein.

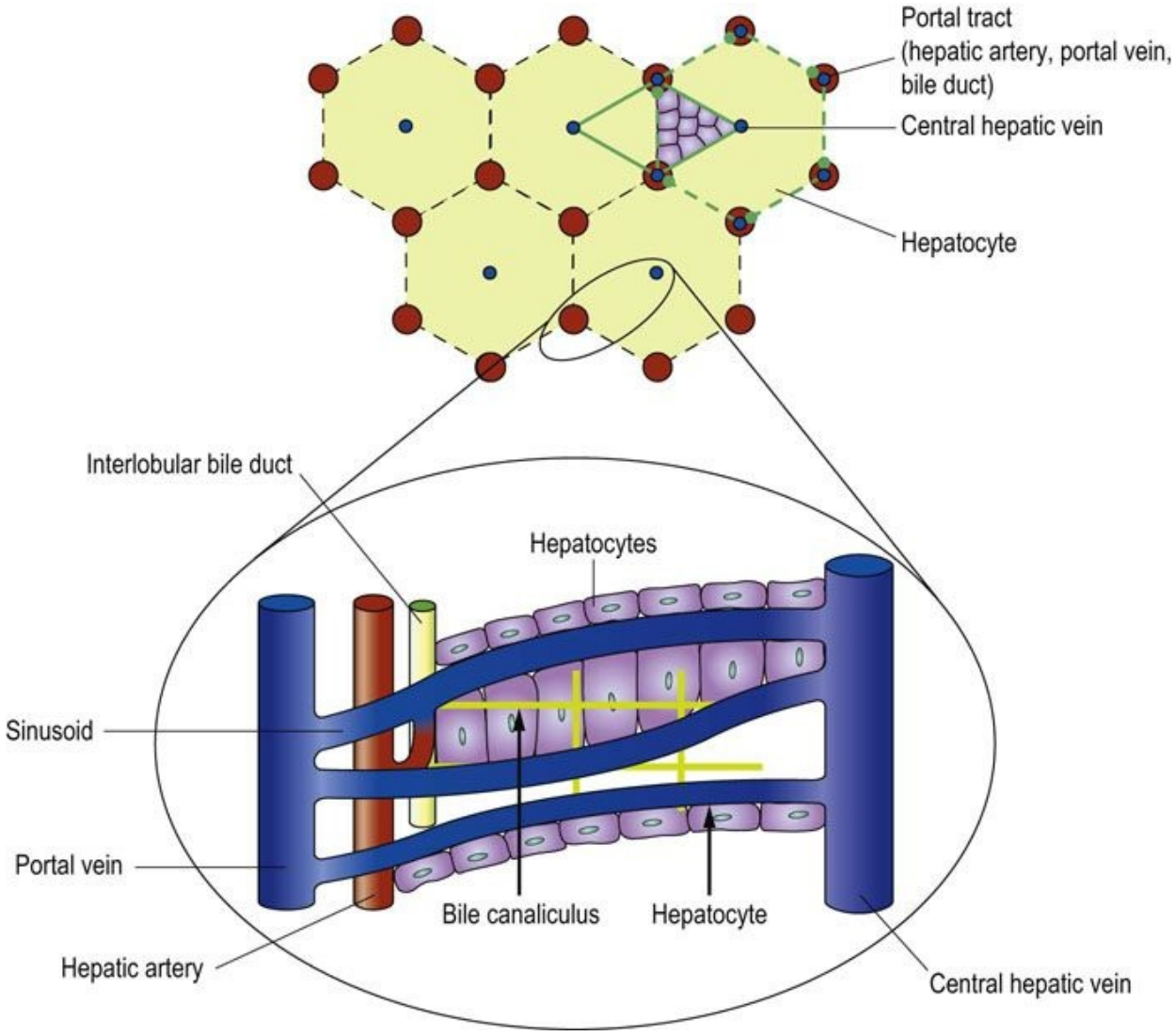


FIG. 30.2 Structure of the liver.

Sinusoids are lined by two cell types. The first are **vascular endothelial cells**, which are loosely connected one with another, leaving numerous gaps. There is no basement membrane between the endothelial cells and the hepatocytes. The second type of sinusoidal cells, known as **Kupffer cells**, are mononuclear phagocytes; they are generally found in the gaps between endothelial cells.

These anatomical arrangements facilitate the exchange of metabolites between hepatocytes and plasma, allow the hepatocytes to receive an arterial supply and permit the excretory products from hepatocyte metabolism destined for biliary excretion to enter the biliary ducts.

Liver and carbohydrate metabolism

The liver plays a central role in glucose metabolism, specifically in maintaining the circulating concentration of glucose (Chapters 13 and 21). This function depends upon its ability both to store a supply of glucose as glycogen, and to synthesize glucose from noncarbohydrate sources, other than lipid, through gluconeogenesis.

Depending on metabolic conditions, the liver can either take up or produce glucose

The liver possesses glucose-6-phosphatase, which permits the release of free glucose to the blood. Although muscle stores more glycogen than the liver, it has no glucose-6-phosphatase and therefore cannot directly contribute glucose to the blood. The kidney has both the ability to synthesize glucose-6-phosphate *de novo* by gluconeogenesis, and has glucose-6-phosphatase activity, but quantitatively it contributes much less than the liver (Chapter 23). Moreover, kidneys do not store glycogen.

Certain organs, notably the brain, are dependent upon glucose as a source of energy, and so the adult liver in the fasting state releases around 9 g of glucose each hour to the blood to maintain the blood glucose concentration. The substrates for gluconeogenesis are derived from lactate released by glycolysis in the peripheral tissues and from hepatic deamination of amino acids (mainly alanine) generated from the proteolysis of skeletal muscle (Chapter 21).

Liver and protein metabolism

Most plasma proteins are synthesized in the liver

Hepatocellular disease may alter protein synthesis both quantitatively and qualitatively. **Albumin** is the most abundant protein in blood and is synthesized exclusively by the liver (Chapter 4). Low plasma albumin concentration occurs commonly in liver disease. It is however, a poor index of hepatic synthetic function because in systemic illness (which often accompanies hepatic disease) there is an increased vascular endothelial permeability that allows the leakage of albumin into the interstitial space.

A better index of hepatocyte synthetic function is the production of the coagulation factors II, VII, IX, and X

All the coagulation factors undergo post-translational γ-carboxylation of specific glutamyl residues, allowing them to bind calcium. As a group, their functional concentration can be readily assessed in the hematology laboratory by measuring the prothrombin time (PT; Chapter 7).

The liver also synthesizes most of the plasma α-and β-globulins. Their plasma concentrations change in hepatic disease and in systemic illness; in the latter case, these changes form part of the **acute phase response**.

Response to an acute insult is associated with wide-ranging changes in liver protein synthesis

The ‘acute phase response’ is a term encompassing all the systemic changes which occur in response to infection or inflammation (Chapter 4). The liver synthesizes a number of ‘acute phase proteins’, which have been defined as those whose plasma concentrations change by more than 25% within a week of the inflammatory or infective insult. The production of these proteins is stimulated by pro-inflammatory cytokines released by macrophages, and of these interleukin-1 (IL-1), IL-6 and tumor necrosis factor (TNF) have a central role. The acute phase proteins have a number of different functions. **Binding proteins**, opsonins, such as C-reactive protein (CRP), bind to macromolecules released by damaged tissue or infective agents and promote their phagocytosis (Chapter 38). **Complement factors** also promote the phagocytosis of foreign

molecules. **Protease inhibitors**, such as α_1-antitrypsin and α_1-antichymotrypsin, inhibit proteolytic enzymes. The latter two also stimulate fibroblast growth and the production of connective tissue required for the repair and resolution of the injury.

A substantial supply of amino acids is required as substrates for this increase in hepatic protein synthesis and these are derived from the proteolysis of skeletal muscle. TNF and IL-1, again, are involved by stimulating the breakdown of specific intracellular proteins by the ubiquitin–proteasome system (see below).

Protein degradation by the ubiquitin–proteasome system

Ubiquitin marks intracellular proteins for proteasomal degradation

Hepatic protein turnover is highly regulated, which allows metabolic pathways to adapt to changing physiologic circumstances. Mammalian cells possess several proteolytic systems.

Plasma proteins and membrane receptors are endocytosed and then hydrolyzed by acid proteases within the lysosomes. Intracellular proteins, on the other hand, are degraded within structures known as proteasomes by the so-called ubiquitin–proteasome system (UPS) (Chapter 33). The discoverers of protein ubiquinylation were awarded the Nobel Prize in chemistry in 2003. The UPS is important in the activation of the NFκB pro-inflammatory pathway and the function of UPS is modified by reactive oxygen species (Chapter 37; see also Fig. 34.10).

Removal of nitrogen

The urea cycle is essential for the removal of nitrogen generated by amino acid metabolism

Catabolism of amino acids generates ammonia (NH_3) and ammonium ions (NH_4^+). Ammonia is toxic, particularly to the central nervous system (CNS). Most ammonia is detoxified at its site of formation, by amidation of glutamate to glutamine, which is mainly derived from muscle and is used as an energy source by the enterocytes. The remaining nitrogen enters the portal vein either as ammonia or as alanine, both of which are used by the liver for the synthesis of

urea (Chapter 19).

Impaired clearance of ammonia causes brain damage

The urea cycle is the major route by which waste nitrogen is excreted, and is described in Chapter 19. In neonates, inherited defects of any of the enzymes of the urea cycle lead to **hyperammonemia**, which impairs the function of the brain, causing encephalopathy. Such problems arise within the first 48 hours of life and inevitably are made worse by protein-rich foods such as milk (see Clinical Box on p. 243).

Heme synthesis

Heme is a constituent of hemoglobin, myoglobin and cytochromes

Heme is synthesized in most cells of the body. The liver is the main nonerythrocyte source of its synthesis. Heme is a porphyrin, a cyclic compound which contains four pyrrole rings linked together by methenyl bridges. It is synthesized from glycine and succinylcoenzyme A, which condense to form 5-aminolevulinate (5-ALA). This reaction is catalyzed by 5-ALA synthase, located in mitochondria, and is rate limiting in heme synthesis. Subsequently, in the cytosol, two molecules of 5-ALA condense to form a molecule containing a pyrrole ring, porphobilinogen (PBG). Then, four PBG molecules combine to form a linear tetrapyrrole compound, which cyclizes to yield uroporphyrinogen III and then coproporphyrinogen III. Final stages of the pathway occur again in the mitochondria where a series of decarboxylation and oxidation of side chains in uroporphyrinogen III yield protoporphyrin IX. At the final stage, iron (Fe^{2+}) is added by ferrochelatase to protoporphyrin IX to form heme. Heme controls the rate of its synthesis by feedback inhibition of 5-ALA synthase (Fig. 30.3).

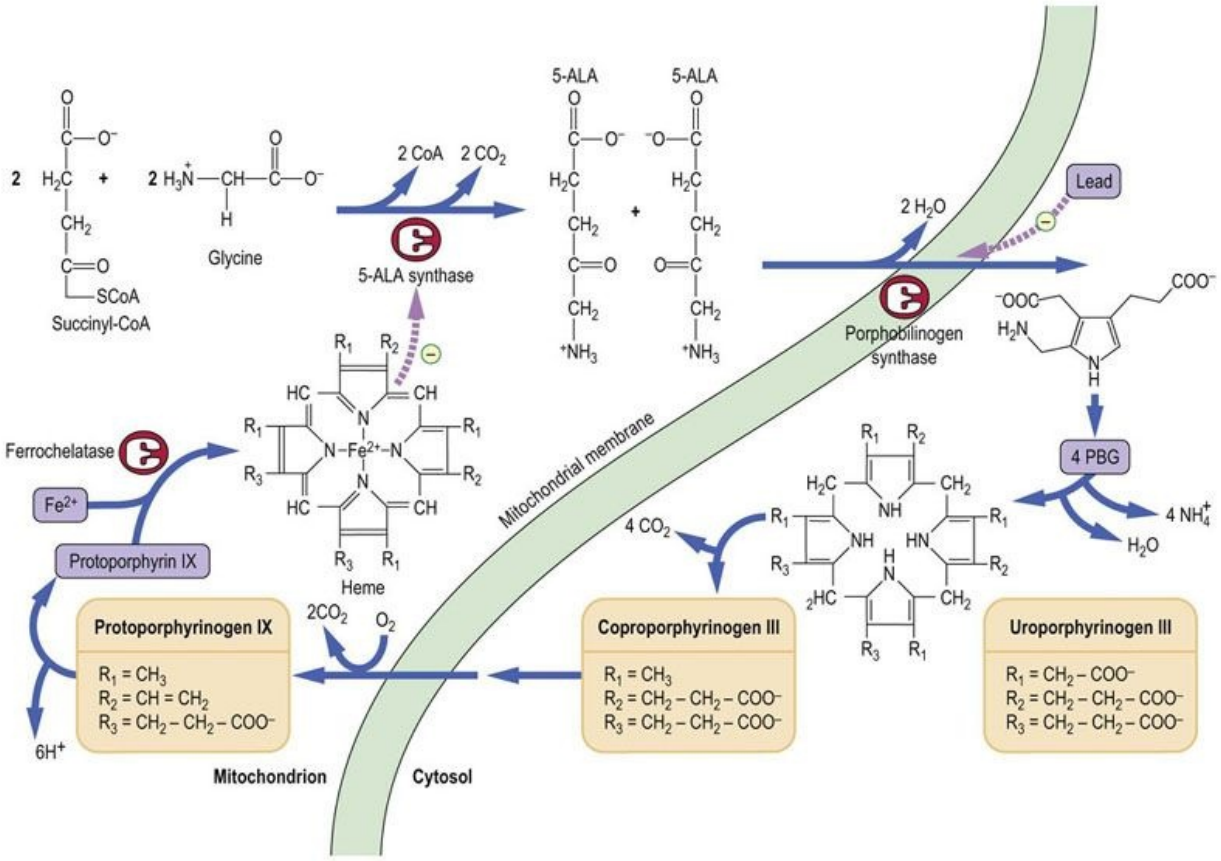


FIG. 30.3 The pathway of heme synthesis.
Part of the pathway is located in the mitochondria and part in the cytosol. ALA, 5-aminolevulinate; PBG, porphobilinogen. Hemoglobin is discussed in Chapter 5.

Advanced concept box The porphyrias

Defects in the heme synthetic pathway lead to rare disorders known as porphyrias. Different porphyrias are caused by deficiencies of different enzymes in the biosynthetic pathway, starting from 5-ALA synthase and ending with ferrochelatase. Porphyrias are classified as hepatic or erythropoietic, depending on the primary organ affected.

Three porphyrias are known as acute porphyrias and can be a cause of emergency admissions with abdominal pain (which needs to be differentiated from various surgical causes). They also cause neuropsychiatric symptoms. **Acute intermittent porphyria (AIC)**

is caused by the deficiency of hydroxymethylbilane synthase, an enzyme converting PBG to a linear tetrapyrrole; in this disorder the concentrations of 5-ALA and PBG increase in plasma and urine. **Hereditary coproporphyria** is due to a defect in the conversion of coproporphyrinogen III to protoporphyrinogen III (coprooxidase). The third acute porphyria is **the variegate porphyria**, the clinical manifestations of which are very similar to AIC.

Other porphyrias, such as porphyria cutanea tarda, present clinically as the sensitivity of skin to light (photosensitivity) which may cause disfiguration and scarring. Also, the pathway is inhibited by lead at the stage of porphobilinogen synthase.

Bilirubin metabolism

Excess bilirubin causes jaundice

Bilirubin is the catabolic product of heme. About 75% of all bilirubin is derived from the breakdown of hemoglobin from senescent red blood cells, which are phagocytosed by mononuclear cells of the spleen, bone marrow, and liver (reticuloendothelial cells). In normal adults the daily load of bilirubin is 250–350 mg. The ring structure of heme is oxidatively cleaved to biliverdin by heme oxygenase, a P-450 cytochrome (see below). Biliverdin is, in turn, enzymatically reduced to bilirubin (Fig. 30.4). The normal plasma concentration of bilirubin is less than 21 µmol/L (1.2 mg/dL). Increased concentrations (more than 50 µmol/L or 3 mg/dL) are readily recognized clinically, because at this concentration or more bilirubin imparts a yellow color to the skin and conjunctivae known clinically as jaundice, or icterus. Abnormalities in bilirubin metabolism are clinical important pointers to the presence of liver disease.

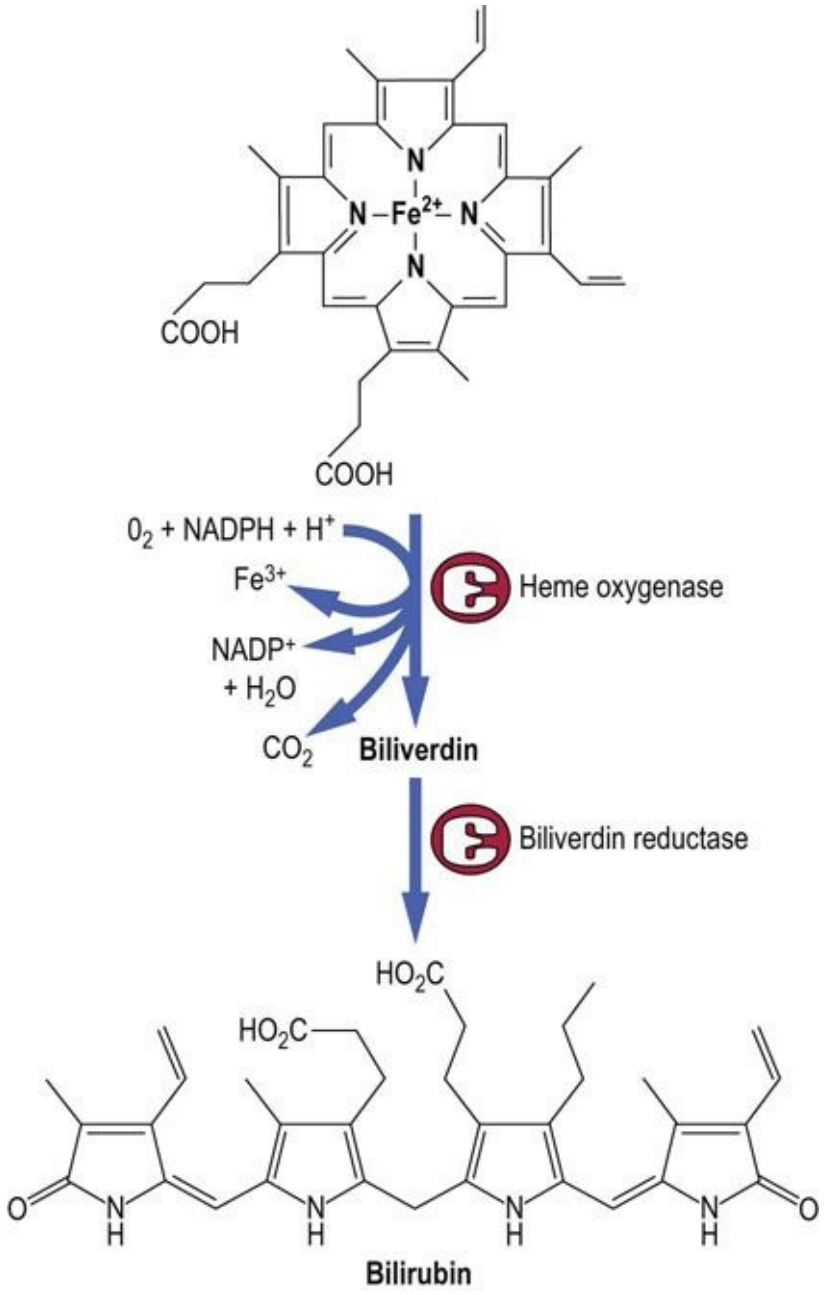


FIG. 30.4 Degradation of heme to bilirubin.

Bilirubin is metabolized by the hepatocytes and excreted in bile

Whereas biliverdin is water soluble, bilirubin, paradoxically, is not. Therefore, it must be further metabolized before excretion (Fig. 30.5). Bilirubin produced by the catabolism of heme in the reticuloendothelial cells is transported in plasma

bound to albumin. The hepatic uptake of bilirubin is mediated by a membrane carrier and may be competitively inhibited by other organic anions. The hydrophilicity of bilirubin is increased by esterification, usually known as conjugation, of one or both of its carboxylic acid side chains with glucuronic acid, xylose or ribose. The glucuronide diester is the major conjugate and its formation is catalyzed by the uridine diphosphate (UDP)-glucuronyl transferase. Conjugated bilirubin is water soluble and may be secreted by the hepatocyte into the biliary canaliculi. If this excretory process is impaired and the patient becomes jaundiced as a result, some conjugated bilirubin may be lost into the urine, imparting a characteristically dark color.

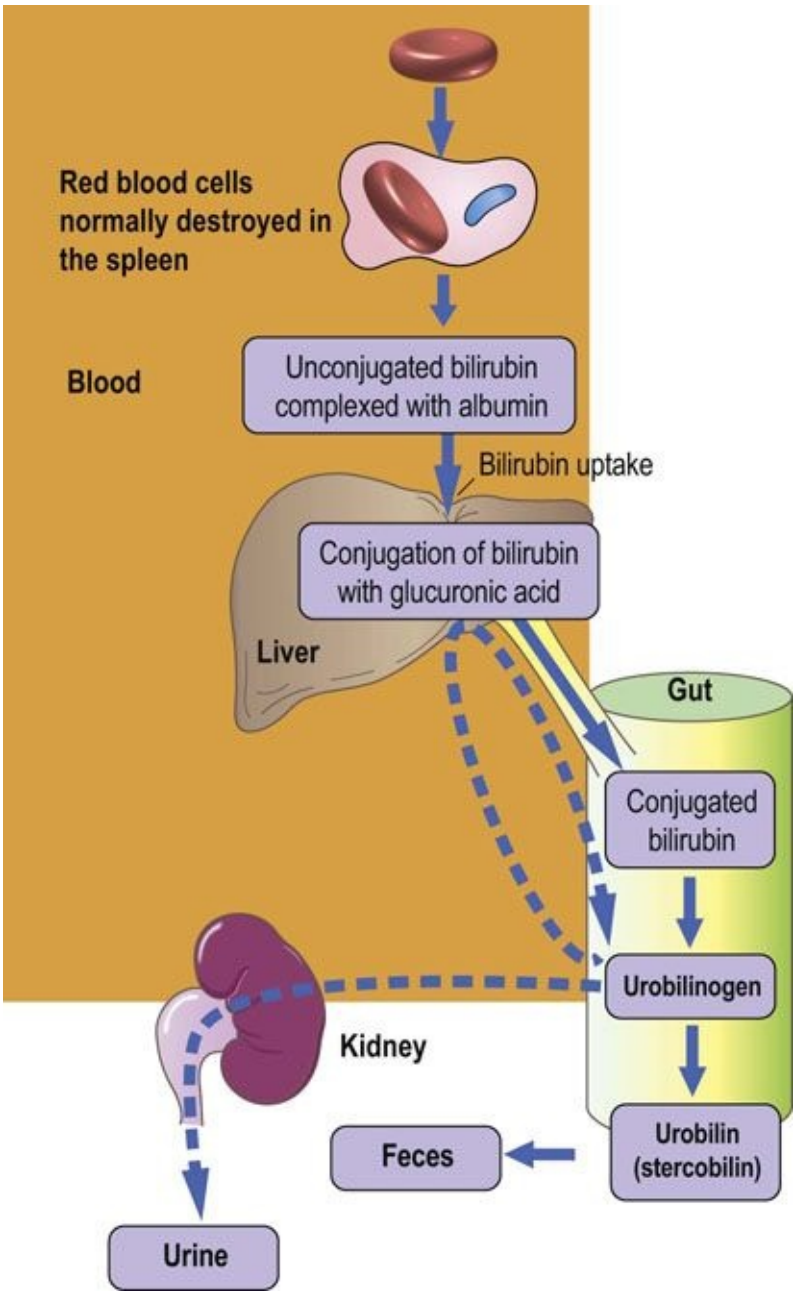


FIG. 30.5 Normal bilirubin metabolism.

Conjugated bilirubin in the gut is catabolized by bacteria to form stercobilinogen, also known as fecal urobilinogen, which is colorless. On oxidation, however, stercobilinogen forms stercobilin (otherwise known as fecal urobilin), which is colored; most stercobilin is responsible for the color of feces. Some stercobilin may be reabsorbed from the gut and can then be re-excreted by either the liver or the kidneys.

Bile acid and cholesterol metabolism

Bile acids are key elements in fat metabolism

Bile acids are synthesized in the hepatocytes and have a detergent-like effect, solubilizing biliary lipids and emulsifying dietary fat in the gut to facilitate its digestion. Their metabolism is described in Chapters 10 and 17. Bile is also the only route of cholesterol excretion from the body (Chapters 17 and 18).

Drug metabolism

Low substrate specificity of some hepatic enzymes produces a wide-ranging capability for drug metabolism

Most drugs are metabolized in the liver. Among other effects, this hepatic metabolism usually increases the hydrophilicity of drugs and therefore their ability to be excreted through the kidneys or in bile. Generally, drug metabolites are less pharmacologically active than the parent drugs; however, some drugs are inactive when administered (prodrugs) but are converted to their active forms as a result of liver processing. The hepatic drug-metabolizing systems need to act on an infinite range of molecules that can be encountered in the environment; this is achieved by the enzymes involved having low substrate specificity.

Drug metabolism proceeds in two phases

Phase I is the addition of the polar group: the polarity of the drug is increased by oxidation or hydroxylation catalyzed by a family of microsomal cytochrome P-450 oxidases.
Phase II is conjugation: cytoplasmic enzymes conjugate the functional groups introduced in the first phase reactions, most often by glucuronidation or sulfation, and also acetylation and methylation.

Three of the 12 cytochrome P-450 gene families share the responsibility for drug metabolism

The cytochrome P-450 enzymes are heme-containing proteins that co-localize with NADPH:cytochrome P-450 reductase. They are present in the endoplasmic reticulum. Most metabolism associated with the cytochrome P-450 superfamily takes place in the liver but these enzymes are also present in the epithelium of the small intestine. The reaction sequence catalyzed by these enzymes is shown in Figure 30.6. There are 12 cytochrome P-450 gene families, of which three, designated *CYP1*, *CYP2* and *CYP3*, are responsible for most of the phase I drug metabolism. In fact, six enzymes, CYP1A2, CYP3A4, CYP2C9, CYP2C19, CYP2D6 and CYP2E1, are responsible for approximately 90% of drug metabolism. CYP3A4 is one of the most important cytochrome P-450 enzymes.

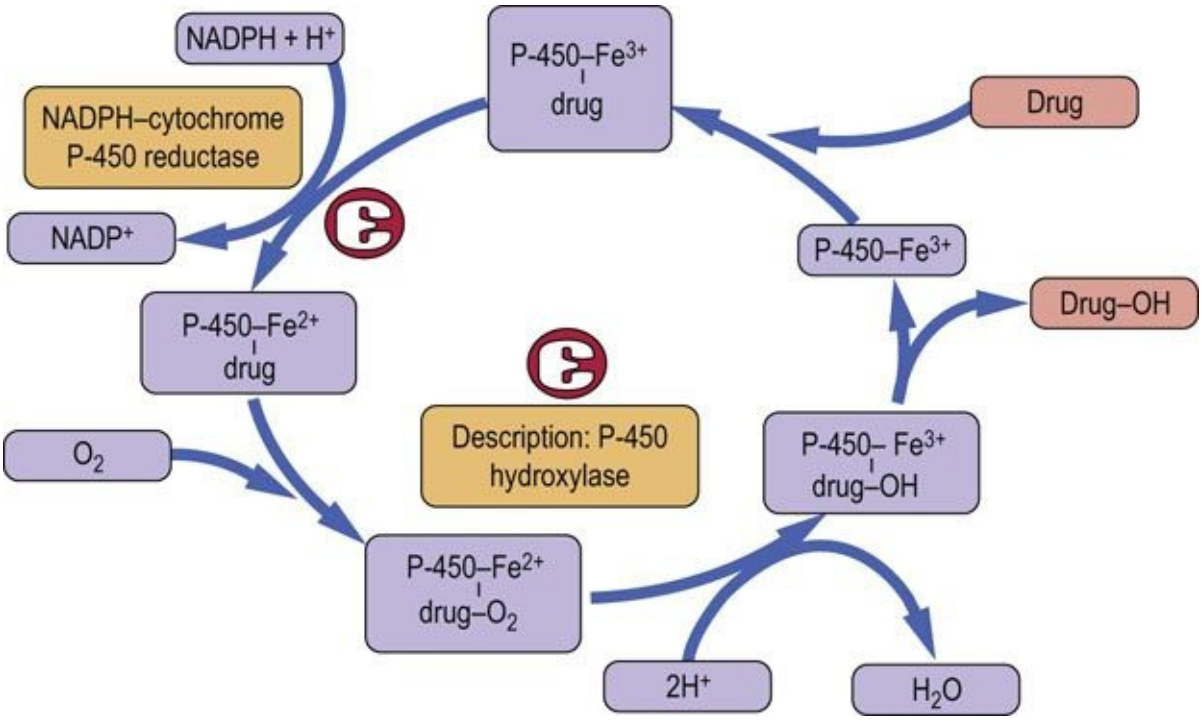


FIG. 30.6 The role of the cytochrome P-450 system in the metabolism of drugs.

Induction, and competitive inhibition of cytochrome P-450 enzymes are mechanisms of drug interactions

Hepatic synthesis of cytochromes P-450 is induced by certain drugs and other xenobiotic agents: this increases the rate of phase I reactions. On the other hand, drugs that form a relatively stable complex with a particular cytochrome P-450 inhibit the metabolism of other drugs that are normally substrates for that cytochrome. For instance, CYP1A2 metabolizes, among others, caffeine and theophylline. It can be inhibited by grapefruit juice, that contains a substance known as naringin, or by the antibiotic ciprofloxacin. When a person takes any of the inhibitory substances, normal substrates for CYP1A2 are metabolized more slowly and their plasma levels increase.

During immunosuppressive therapy the dose of the immunosuppressant cyclosporine may need to be reduced by up to 75% if the patient also takes the antifungal drug ketoconazole (see Wilkinson in Further Reading). Otherwise adverse reactions may occur.

The drugs that induce induction or repression of CYP3A enzymes often act through the nuclear receptor mechanism. They combine with nuclear receptors (i.e. in case of CYP3A4, the pregnane X receptor, PXR), which then form

heterodimers with retinoid X receptors (Chapter 17). Such complexes upregulate CYP3 synthesis by binding to response elements in the gene promoter.

Cytochrome P-450 gene polymorphisms determine response to many drugs

Allelic variation that affects the catalytic activity of a cytochrome P-450 also affects the pharmacologic activity of drugs. The best described example of such polymorphism is that of the P-450 cytochrome CYP2D6, which was recognized initially in the 5–10% of normal individuals who were noted to be slow to hydroxylate debrisoquine, a now little used blood pressure-lowering drug. However, CYP2D6 also metabolizes a significant number of other commonly used drugs, so that ‘debrisoquine polymorphism’ remains clinically significant.

Antiplatelet therapy with clopidogrel is the current standard of care for coronary artery disease patients undergoing revascularization. However, approximately 25% of patients experience a subtherapeutic antiplatelet response. Clopidogrel is a prodrug that undergoes hepatic biotransformation by CYP2C19 into its active metabolite. Several studies have reported that carriers of CYP2C19 variant allele exhibit a significantly lower capacity to transform clopidogrel to its active metabolite and are therefore at significantly higher risk of adverse cardiovascular events. Consequently, in the US, the FDA has recently changed clopidogrel's prescribing information to highlight the impact of CYP2C19 genotype on the clinical response to clopidogrel.

Genotyping of cytochromes P-450 to identify gene-relevant polymorphisms may become more common in an attempt to personalize an individual's response to a particular drug.

Drug hepatotoxicity

Drugs that exert their toxic effects on the liver may do so through the hepatic production of a toxic metabolite. Drug toxicity may occur in all individuals exposed to a sufficient concentration of a particular drug. However, a drug may be toxic in some individuals at concentrations normally tolerated by most other patients. This phenomenon is known as idiosyncratic drug toxicity and may be due to a genetic or immunologic cause.

The commonly prescribed drug acetaminophen (paracetamol) is hepatotoxic in excess

Acetaminophen is widely used as a painkiller. Taken in the usual therapeutic doses, it is conjugated with glucuronic acid or sulfate, which is then excreted by the kidneys. In overdose, however, the capacity of these conjugation pathways is overwhelmed and acetaminophen is oxidized by a liver P-450 cytochrome CYP3A4 to *N*-acetyl benzoquinoneimine (NABQI). NABQI can cause a free radical-mediated peroxidation of membrane lipids, and consequently hepatocellular damage. It may be detoxified by conjugation with glutathione, but in acetaminophen overdose glutathione stores also become exhausted, causing hepatotoxicity (Fig. 30.7). Therapeutically, a sulfhydryl compound, *N*-acetylcysteine (NAC), is routinely used as an antidote to acetaminophen poisoning. It promotes detoxification of NABQI by the glutathione pathway and also scavenges free radicals. The risk of hepatotoxicity can be reliably predicted from measurement of the plasma concentration of acetaminophen in relation to the time which has elapsed since the overdose, and NAC can be given to patients who are at risk of liver damage. Measurement of paracetamol is thus one of the emergency tests offered by clinical laboratories.

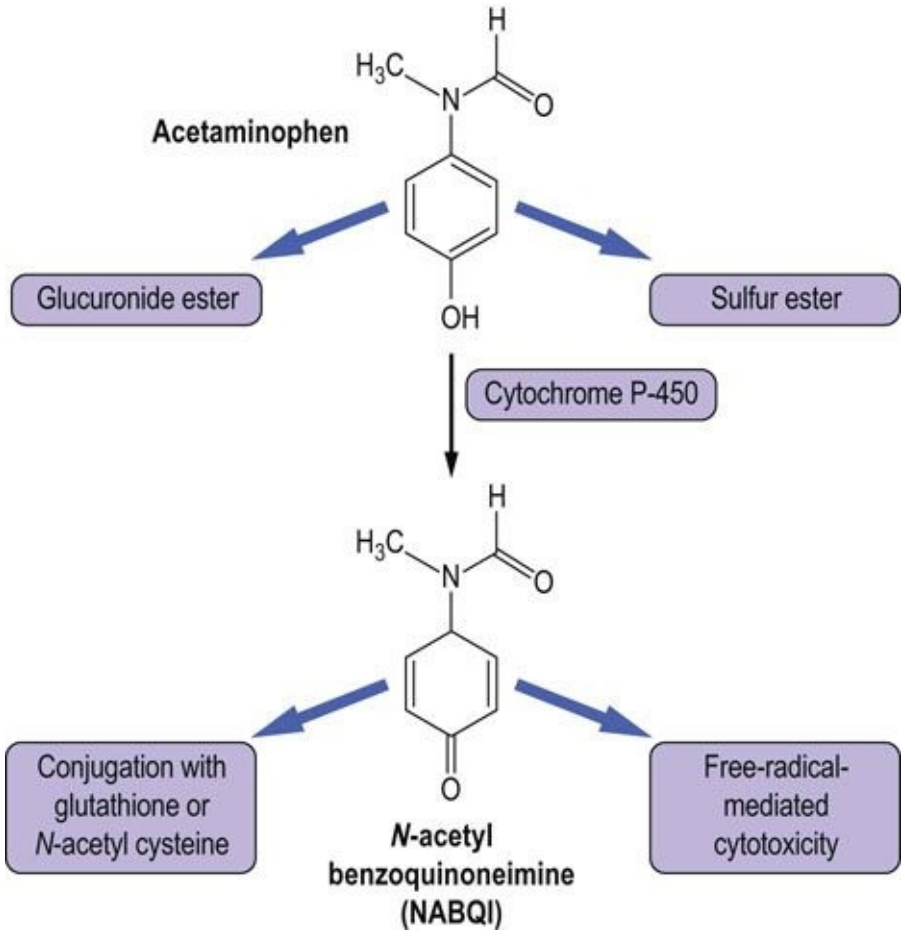


FIG. 30.7 Metabolism of acetaminophen (paracetamol).

Alcohol

Alcohol excess is a major cause of liver disease

Excess intake of ethyl alcohol (ethanol) remains the most common cause of liver disease in the Western world. Ethanol may cause excessive fat deposition in the liver (**alcoholic steatosis**), **hepatitis** or finally **fibrosis** (known as **cirrhosis**), which in turn leads to **liver failure**. There are over 25,000 deaths associated with liver disease in the US annually, and 40% of these are linked to alcoholic cirrhosis (see Donohue in Further Reading).

Ethanol is oxidized in the liver, mainly by alcohol dehydrogenase (ADH), to form acetaldehyde, which is in turn oxidized by aldehyde dehydrogenase (ALDH) to acetate. Nicotinamide adenine dinucleotide (NAD^+) is the cofactor

for both these oxidations, being reduced to NADH. A P-450 cytochrome, CYP2E1, also contributes to ethanol oxidation but is quantitatively less important than the ALD–ALDH pathway. Liver damage in patients who abuse alcohol may arise from the toxicity of acetaldehyde, which forms Schiff base adducts with other macromolecules.

Ethanol oxidation alters the redox potential of the hepatocyte

Ethanol oxidation results in the increased ratio of NADH to NAD^+ This inhibits oxidation of lactate to pyruvate (a step that requires NAD^+ as a cofactor) and also creates potential for the development of lactic acidosis. Because pyruvate is a substrate for hepatic gluconeogenesis, there is also a risk of hypoglycemia. The risk of hypoglycemia is further increased in alcoholics, when, because of poor nutrition, they have low hepatic glycogen stores. Also, the shift in the NADH/NAD^+ ratio inhibits β-oxidation of fatty acids and promotes triglyceride synthesis: excess of triglycerides is deposited in the liver and secreted into plasma as VLDL (see Clinical Box on p. 197). Hepatic steatosis can be readily diagnosed by ultrasonography of the liver when one sees a uniform increased echogenicity (Fig. 30.8). It is often associated with elevation in serum levels of the transaminase enzymes.

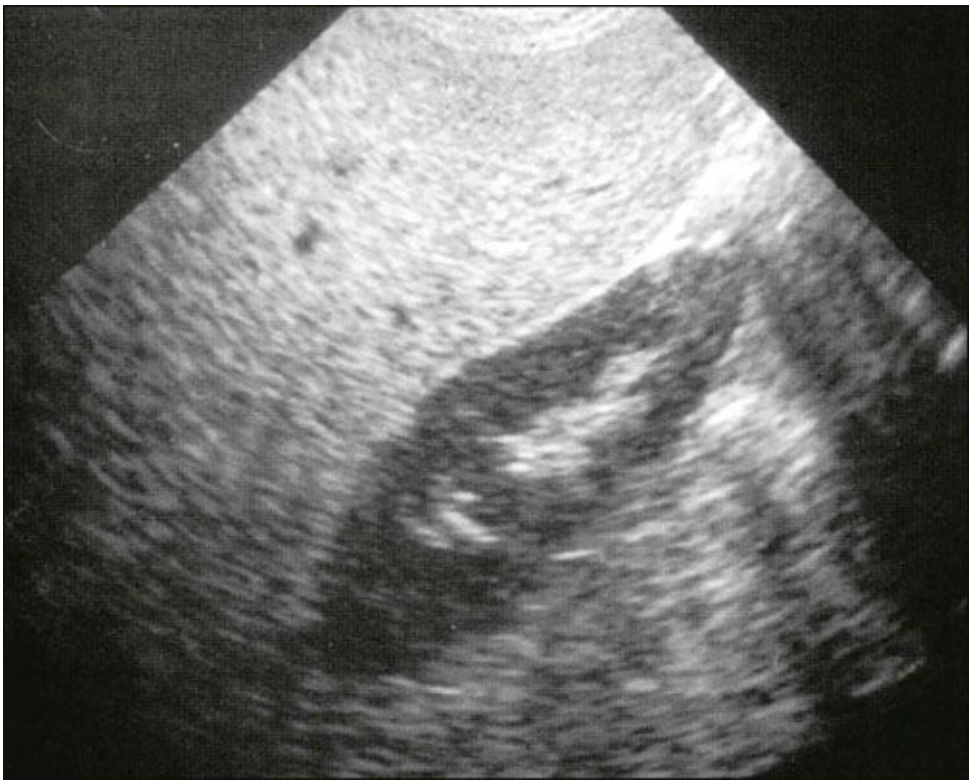


FIG. 30.8 An ultrasound scan of a liver showing steatosis. Courtesy Dr A Bannerjee, Birmingham Heartlands and Solihull NHS Trust, UK.

Ethanol consumption also affects the ubiquitin system of protein degradation. Chronic alcohol consumption decreases proteasome activity. This can deregulate the hepatocyte signaling system by inhibiting the Janus Kinase-Signal Transducer and Activator of Transcription (JAK-STAT) signalling pathway, which is involved in the acute phase response, antiviral defense and hepatic repair (Chapter 40). Inhibition of proteasome activity may also lead to increased apoptosis (Chapter 43), a feature of alcoholic liver disease (ALD). The ethanol-induced decrease in proteasome activity prevents the degradation of CYP2E1, which is involved in peroxidation reactions; this increases oxidative stress and may be another factor contributing to ALD.

Finally, alcohol-induced decrease in proteasome activity may lead to the accumulation of protein in the liver, which in turn causes liver enlargement (hepatomegaly; common in ALD). Other ethanol-induced phenomena include increased secretion of chemokines (including IL-8 and monocyte-chemoattractant protein-1 (MCP-1; Chapter 18)) by hepatocytes, leading to liver infiltration by neutrophils.

Clinical box A 22-year-old woman who took acetaminophen overdose

A 22-year-old woman was admitted to hospital in a semiconscious state. She had been found with a suicide note and empty acetaminophen containers. Tests revealed aspartate aminotransferase (AST) 5500 U/L, alkaline phosphatase (ALP) 125 U/L, bilirubin 70 µmol/L (4.1 mg/dL), prothrombin time 120 s (reference range 10–15 s), creatinine 350 µmol/L (4.0 mg/dL; reference range 44–80 µmol/L, 0.50–0.90 mg/dL), glucose 2.6 mmol/L (47 mg/dL; reference range 4.0–6.0 mmol/L, 72–109 mg/dL), and blood pH 7.1 (Reference range 7.35-7.45; This equals to 80 nmol/L H+ with reference range 35-45 nmol/L). No acetaminophen was found in her plasma.

Comment.

The patient had acute hepatic failure, most probably caused by acetaminophen poisoning. Blood acetaminophen may be undetectable if the patient first comes to medical attention more than 24 h after an overdose. The hepatocellular damage worsens over the first 72 h but may improve spontaneously after that, as a result of regeneration of hepatocytes. However, in patients with a metabolic acidosis (pH <7.35 or H^+ above 45 mnmol/L, after fluid resuscitation), markedly increased prothrombin time (>100 s) or serum creatinine >300 µmol/L (3.4 mg/dL), mortality is of the order of 90% and liver transplantation may be necessary. For reference ranges (Chapter 40), see Table 30.2 and Appendix 1.

Table 30.2

Laboratory tests used in differential diagnosis of jaundice

Test	Prehepatic	Intrahepatic	Posthepatic
Conjugated bilirubin	Absent	Increased	Increased
AST and ALT	Normal	Increased	Normal
ALP	Normal	Normal	Increased
Urine bilirubin	Absent	Present	Present
Urine urobilinogen	Present	Present	Absent

Reference ranges for liver function tests:

AST (aspartate aminotransferase), men 15–40 U/L, women 13–35 U/L; ALT (alanine aminotransferase), men 10–40 U/L, women 7–35 U/L; ALP (alkaline phosphatase), 50–140 U/L: ALP is physiologically elevated in children and adolescents; bilirubin <21 µmol/L (<1.0 mg/dL); (γ-glutamyl transpeptidase:) GGT, men <90 U/L, women <50 U/L.

Symptoms of alcohol intolerance are exploited to reinforce abstinence

Both ADH and ALDH are subject to genetic polymorphisms, which have been investigated as a potential inherited basis of susceptibility to alcoholism and ALD. Possession of the ALDH2[2] allele, which encodes an enzyme with reduced catalytic activity, leads to increased plasma concentrations of acetaldehyde after the ingestion of alcohol. This causes the individual to experience unpleasant flushing and sweating, which discourages alcohol abuse. Disulfiram, a drug that inhibits ALDH, also causes these symptoms when alcohol is taken, and may be given to reinforce abstinence from alcohol.

Pharmacogenomics

The response to any particular drug is influenced by drug's kinetic properties (pharmacokinetics) and its effects (pharmacodynamics)

An individual's response to a drug can be influenced by genes that code for drug-metabolizing enzymes, receptors and transporters. Any variability in these genes may lead to inter-individual differences in response to that drug.

The effectiveness and safety of drug therapy, particularly in elderly patients or those with renal or hepatic diseases as co-morbidities, and in patients whose metabolic capacity is diminished, is currently a major problem. Approximately 3% of hospital admissions in the US are linked to drug–drug interactions and a Dutch study reported values as high as 8.4%. In the US, there are 2 million cases of adverse drug reactions annually, including 100,000 deaths. Combined with the fact that most drugs are effective in only 25–60% of patients to whom they are prescribed, this makes research into individual response to drugs absolutely essential.

Pharmacogenomics studies the effects of genetic heterogeneity on drug responsiveness

Since the liver plays a central role in drug metabolism, the pharmacogenomics of some hepatic drug-metabolizing enzymes, specifically the cytochrome P-450 oxidases, is clinically very relevant. CYP2D6 is responsible for the metabolism of more than 100 pharmaceuticals, and a polymorphism of this enzyme is responsible for the long-established variation in the metabolism of debrisoquine, mentioned above. Patients are classified as ultra-rapid, extensive, intermediate and poor metabolizers of debrisoquine. There is one CYP2D6 genetic locus, and individuals may have two, one or no functional alleles, corresponding to extensive, intermediate and poor metabolizers, respectively: gene multiplication can lead to three functional alleles and the ultra-rapid metabolizer phenotype. Seventy-five CYP2D6 alleleic variants have been identified and pharmacogenetic techniques can identify the metabolizer phenotype, thereby predicting clinical response to treatment. Although debrisoquine is now obsolete, the CYP2D6 polymorphism is relevant for some drugs used in cardiac and

psychiatric practice. For instance, poor metabolizers are more likely than other individuals to experience drug toxicity, and less likely to gain benefit from the analgesic codeine, a prodrug which is metabolized by CYP2D6 to morphine, the active drug. A polymorphism of CYP2C19, again leading to extensive and poor metabolizer phenotypes, affects the metabolism of the proton pump inhibitor drugs used in gastroesophageal reflux disease and the effectiveness of treatment (Chapter 8).

Biochemical tests of liver function

Clinical laboratories offer a panel of measurements on plasma or serum specimens (Table 30.2). This group of tests is usually, and incorrectly, described as liver 'function' tests. While plasma activities of liver enzymes are markers of liver disease, they do not exactly reflect the function of the liver. The rate of prothrombin synthesis, assessed by prothrombin time (PT), is a better indicator of liver synthetic function.

The tests commonly include the measurements of:

- bilirubin
- albumin
- aspartate aminotransferase (AST) and alanine aminotransferase (ALT)
- alkaline phosphatase (ALP)
- γ-glutamyl transferase (γGT).

Transaminases

Aspartate aminotransferase (AST) and alanine amino-transaminase (ALT) are involved in the interconversion of amino and ketoacids, and are required for metabolism of proteins and carbohydrates (Chapter 19). Both are located in the mitochondria; ALT is also found in the cytoplasm. Serum activity of ALT and AST increases in liver disease (ALT is the more sensitive measurement).

Prothrombin time

In liver disease, the synthetic functions of the hepatocytes are likely to be affected, and so the patient would be expected to have a prolonged prothrombin time and low serum albumin concentration.

Alkaline phosphatase

ALP is synthesized both by the biliary tract and by bone, and in pregnancy by the placenta, but these tissues contain different ALP isoenzymes. The origin of the ALP may be determined from the isoenzyme pattern. Alternatively, the plasma activity of another enzyme, such as γ-glutamyl transpeptidase (GGT), which also originates in the biliary tract, may be measured and used to confirm the hepatic origin for a raised serum ALP activity.

Clinical box A seemingly healthy 45-year-old man with abnormal transaminases

A 45-year-old businessman had a routine medical examination, at which he was found to have a slightly enlarged liver. Tests revealed bilirubin 15 µmol/L (0.9 mg/dL), AST 434 U/L, ALT 198 U/L, ALP 300 U/L, GGT 950 U/L, and albumin 40 g/L (4 g/dL). He seemed perfectly well.

Comment.

The patient has asymptomatic liver disease. The biochemical tests show evidence of hepatocellular damage. This may be due to excess alcohol intake, in which case there may also be enlarged red blood cells (macrocytosis) and an increased serum uric acid concentration. Patients may deny alcohol abuse. However **nonalcoholic fatty liver disease** (NAFLD) is increasingly recognized as a cause of isolated abnormalities in serum transaminase concentrations. NAFLD occurs in 40% of patients with the so-called metabolic syndrome, in which central overweight due to the accumulation of visceral fat leads to insulin resistance, hypertension, dyslipidemia and hepatic steatosis. The later may lead on to cirrhosis, as may alcohol-induced liver disease. Needle biopsy of the liver may be necessary for diagnosis. In this patient, microscopic examination of the tissue revealed characteristic changes of alcoholic steatosis, hepatitis and central fibrosis. This may be the forerunner of cirrhosis and the patient should abstain from alcohol. Other causes, such as chronic viral infection of the liver or an autoimmune active chronic hepatitis, can be detected by blood tests. For reference ranges, see Table 30.2.

Classification of liver disorders

Hepatocellular disease

Inflammatory disease of the liver is termed **hepatitis** and may be of short (acute) or long (chronic) duration. Viral infections, particularly hepatitis A and E, are common infectious causes of acute hepatitis, whereas alcohol and acetaminophen are the most common toxicologic causes. Chronic hepatitis, defined as inflammation persisting for more than 6 months, may also be due to the hepatitis B and C viruses, alcohol, and immunologic diseases, in which the body produces antibodies against its own tissues (autoimmune diseases, Chapter 38). **Cirrhosis** is the result of chronic hepatitis and is characterized microscopically by fibrosis of the hepatic lobules. The term '**hepatic failure**' denotes a clinical condition in which the biochemical function of the liver is severely, and potentially fatally, compromised.

Cholestatic disease

Cholestasis is the clinical term for **biliary obstruction**, which may occur in the small bile ducts in the liver itself or in the larger extrahepatic ducts. Biochemical tests cannot distinguish between these two possibilities, which generally have radically different causes; imaging techniques such as ultrasound are more helpful.

Jaundice

Jaundice can be pre-, post-or intrahepatic

Jaundice is clinically obvious when plasma bilirubin concentrations exceed 50 µmol/L (3 mg/dL). Hyperbilirubinemia is the result of an imbalance between its production and excretion. The causes of jaundice (Table 30.3) are conventionally classified as:

Table 30.3

Causes of jaundice

Type	Cause	Clinical example	Frequency
Prehepatic	Hemolysis	Autoimmune	Uncommon
		Abnormal hemoglobin	Depends on region
Intrahepatic	Infection	Hepatitis A, B, C	Common/very common
	Chemical/drug	Acetaminophen	Common
		Alcohol	Common
	Genetic errors: bilirubin metabolism	Gilbert's syndrome	1 in 20
		Crigler–Najjar syndrome	Very rare
		Dubin–Johnson syndrome	Very rare
		Rotor's syndrome	Very rare
	Genetic errors: synthesis of specific proteins	Wilson's disease	1 in 200,000
		α_1-antitrypsin	1 in 1000 with genotype
	Autoimmune	Chronic active hepatitis	Uncommon/rare
	Neonatal	Physiologic	Very common
Posthepatic	Intrahepatic bile ducts blockage	Drugs	Common
		Primary bilary cirrhosis	Uncommon
		Cholangitis	Common
	Extrahepatic bile ducts blockage	Gallstones	Very common
		Pancreatic tumor	Uncommon
		Cholangiocarcinoma	Rare

■ **Prehepatic:** increased production or impaired hepatic uptake of bilirubin (Fig. 30.9).

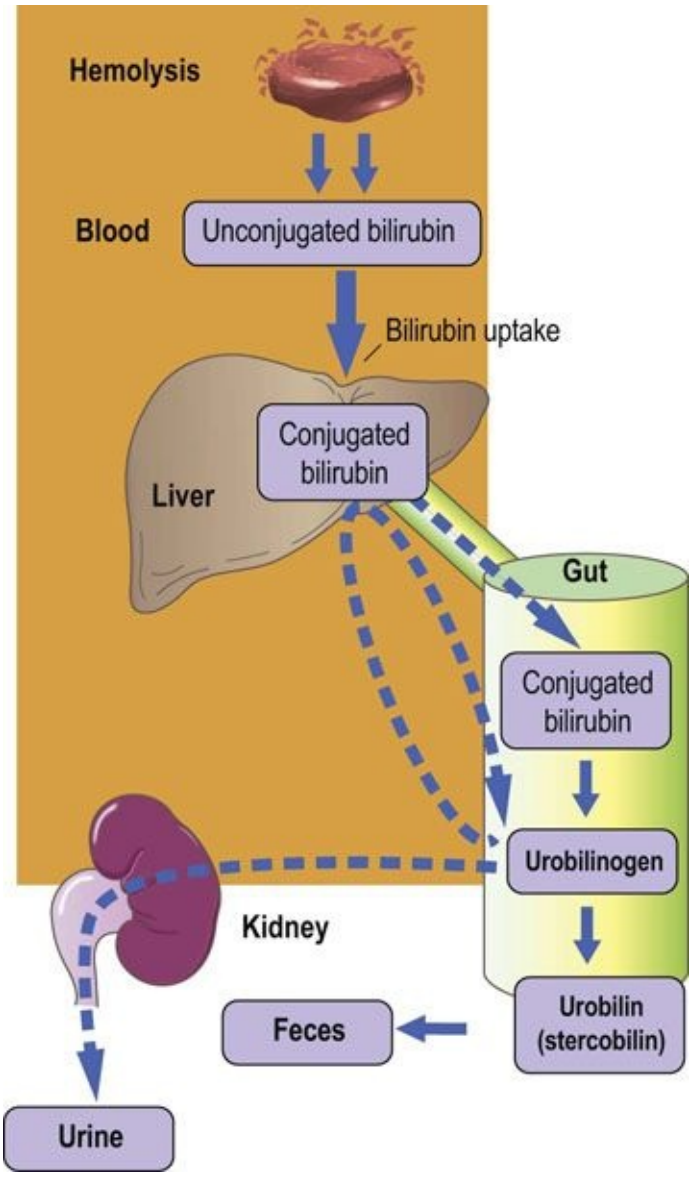


FIG. 30.9 Prehepatic (hemolytic) jaundice.
There is an increased concentration of plasma total bilirubin due to excess of the unconjugated fraction (see also Table 30.2).

■ **Intrahepatic:** impaired hepatic metabolism or secretion of bilirubin (Fig. 30.10).

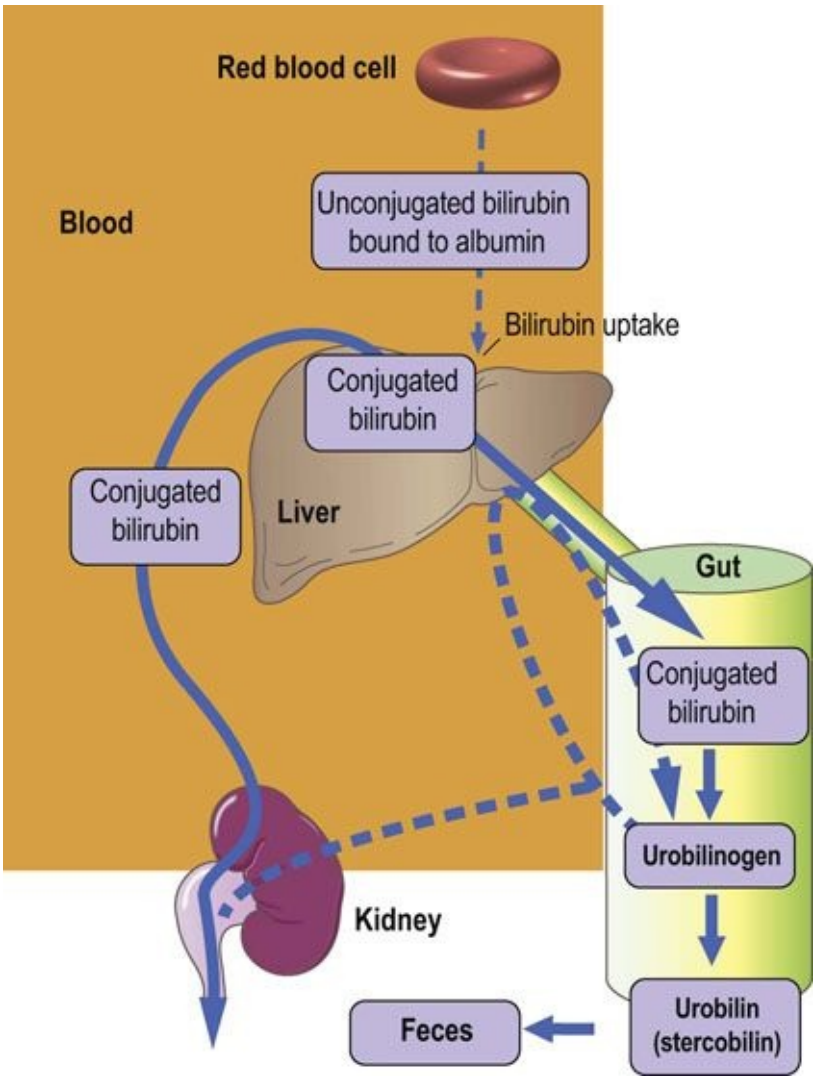


FIG. 30.10 Intrahepatic jaundice.
Bilirubin in plasma is increased due to an increase in the conjugated fraction. Increased serum enzyme activities signify hepatocyte damage (see also Table 30.2).

■ **Posthepatic:** obstruction to biliary excretion (Fig. 30.11 and Box on p. 210).

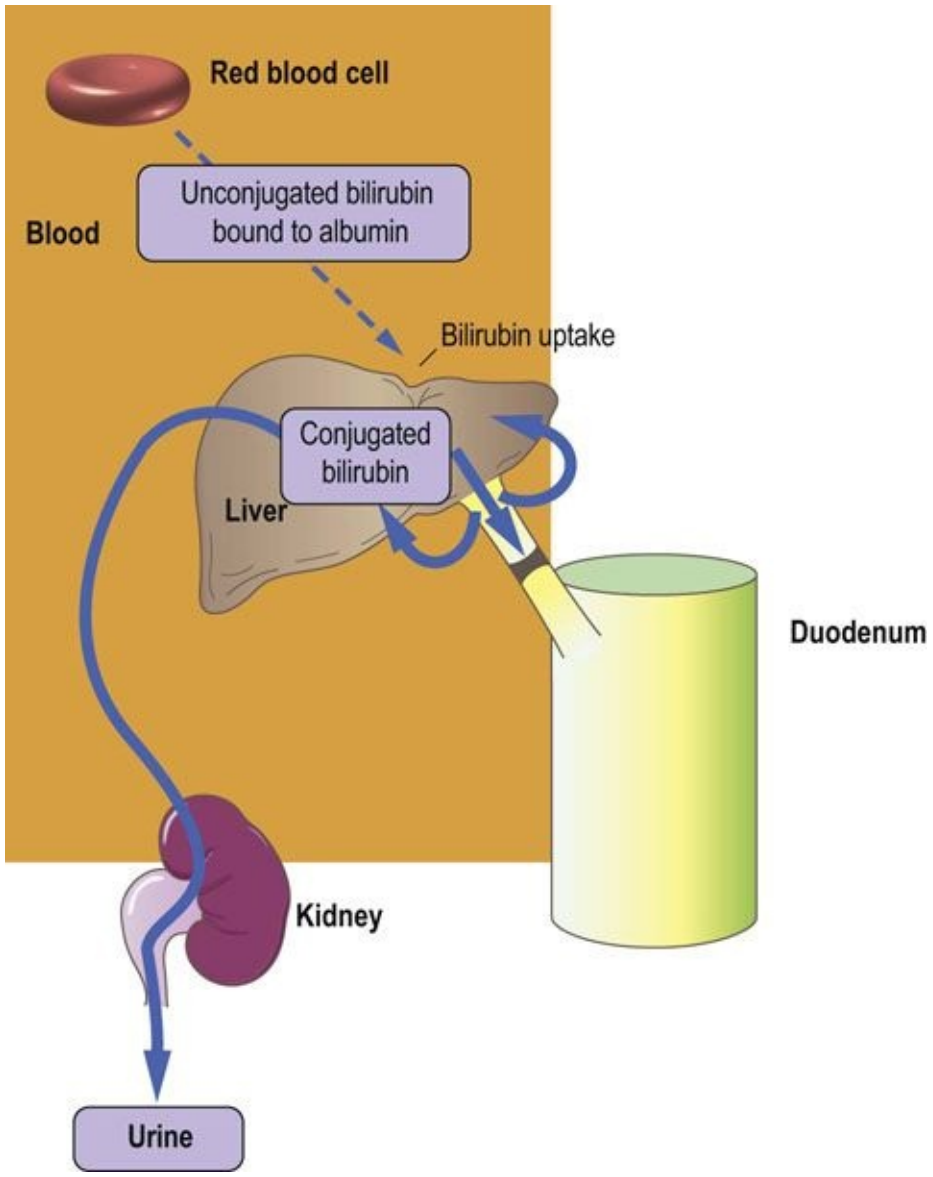


FIG. 30.11 Posthepatic jaundice.
Plasma bilirubin is elevated due to an increase in the conjugated fraction. Obstruction of the bile duct does not allow passage of bile to the gut. Stools are characteristically pale in color (see also Table 30.2). GGT, γ-glutamyl transpeptidase; ALP, alkaline phosphatase.

Prehepatic hyperbilirubinemia, results from excess production of bilirubin caused by hemolysis, or a genetic abnormality in the hepatic uptake of unconjugated bilirubin

Hemolysis is usually the result of immune disease, presence of structurally abnormal red cells or breakdown of extravasated blood. Intravascular hemolysis

results in the release of hemoglobin into the plasma, where it is either oxidized to methemoglobin (Chapter 5) or complexed with haptoglobin. More commonly, red cells are hemolyzed extravascularly, within phagocytes, and hemoglobin is converted to bilirubin, which is **unconjugated**. Unconjugated and conjugated bilirubin can be distinguished in the laboratory as so-called indirect and direct bilirubin.

Intrahepatic jaundice reflects a generalized hepatocyte dysfunction

In this condition, hyperbilirubinemia is usually accompanied by other abnormalities in biochemical markers of hepatocellular function.

In neonates, transient jaundice is common, particularly in premature infants, and is due to immaturity of the enzymes involved in bilirubin conjugation. Unconjugated bilirubin is toxic to the immature brain and causes a condition known as **kernicterus**. If plasma bilirubin concentrations are judged to be too high, phototherapy with blue-white light, which isomerizes bilirubin to more soluble pigments that might be excreted with bile, or exchange blood transfusion to remove the excess bilirubin, is necessary to avoid kernicterus.

Posthepatic jaundice is caused by obstruction of the biliary tree

In this condition the plasma bilirubin is conjugated and other biliary metabolites, such as bile acids, accumulate in the plasma. The clinical features are pale-colored stools, caused by the absence of fecal bilirubin and urobilin, and dark urine as a result of the presence of water-soluble conjugated bilirubin. In complete obstruction, urobilinogen and urobilin are absent from urine, as there can be no intestinal conversion of bilirubin to urobilinogen/urobilin, and hence no renal excretion of reabsorbed urobilinogen/urobilin.

Clinical box A 3-day-old neonate who developed jaundice

the significance of neonatal jaundice

A normal term baby developed jaundice on the third day of life, with a bilirubin concentration of 150 µmol/L (8.8 mg/dL), predominantly of the indirect form. The baby was otherwise well.

Comment.

About 50% of normal babies become jaundiced 48 h after birth. This physiologic jaundice is caused by temporary inefficiency in bilirubin conjugation, and resolves in the first 10 days. The hyperbilirubinemia is unconjugated in nature; if severe, it may require phototherapy (ultraviolet light to photoisomerize bilirubin into a nontoxic form) or exchange blood transfusion to prevent damage to the brain (kernicterus). Bruising from delivery, infection or poor fluid intake may exaggerate the hyperbilirubinemia. Jaundice in the first 24 h of life is abnormal and requires investigation to exclude hemolysis, as does jaundice that presents later or persists, after 10 days. This is always abnormal and is likely to indicate an inborn error of metabolism or structural defects of the bile ducts.

Clinical box A 65-year-old man with jaundice and no abdominal symptoms

the significance of adult jaundice

A 65-year-old man was admitted to hospital because of jaundice. There was no abdominal pain, but he had noticed dark urine and pale stools. The gallbladder appeared palpable but not tender. Liver function tests showed bilirubin 230 µmol/L (13.5 mg/dL), AST 32 U/L, and ALP 550 U/L. Dipstick urine testing revealed the presence of bilirubin, but no urobilin.

Comment.

The patient had a history typical of obstructive jaundice. The increased ALP and normal AST concentrations were consistent with this, and the absence of urobilin in the urine indicated that the biliary tract was obstructed. It was important to carry out liver

imaging tests to find the site of the obstruction; the absence of pain suggested that gall-stones were not the cause. Ultrasound showed dilatation of the bile duct and a computed tomography scan showed a mass in the head of the pancreas. During surgery, cancer of the pancreas was confirmed. For reference ranges, see Table 30.2.

Genomics of liver disease

Several hepatic diseases arise due to single gene disorders. Genetic techniques can identify individuals with a propensity to develop a disease or confirm the diagnosis in affected persons.

Hereditary hemochromatosis is a genetically determined disorder of iron metabolism

It is common in northern Europeans, with a population prevalence of 1 in 200 to 1 in 300. Patients absorb excessive amounts of iron from the gut, and tissue iron overload leads to multiorgan dysfunction, including cirrhosis of the liver. A transmembrane glycoprotein, known as HFE, modulates iron uptake. Mutations in the genetic locus which encodes this protein, the *HFE* gene, underlie hereditary hemochromatosis. Two point mutations, *C282Y* and *H63D*, are found in the majority of those with hereditary hemochromatosis and can be identified easily by PCR-based assays.

Wilson's disease is a condition associated with liver and CNS damage, results from genetic deficiency of ceruloplasmin

Ceruloplasmin is the major copper-containing protein of the liver and plasma, and functions as an iron oxidizing enzyme (ferroxidase): oxidation of Fe^{2+} to Fe^{3+} is necessary for the mobilization of stored iron, and nutritional copper deficiency produces anemia. Wilson's disease is a condition associated with damage to both the liver and the CNS.

Cirrhosis in this condition is due to excess copper deposition in the liver. Wilson's disease is also caused by a single gene defect but multiple mutations have been found in affected patients; no routine assay is available for diagnosis. The liver also synthesizes proteins responsible for storage (ferritin) and transport (transferrin) of iron (Chapters 11 and 22).

Deficiency of α_1-antitrypsin presents in infancy as liver disease, or in adulthood as lung disease

α_1-Antitrypsin is a member of the serpin family of serine protease inhibitors,

and, contrary to its name, its predominant target is macrophage-derived elastase. Genetic deficiency of α_1-antitrypsin presents in infancy as liver disease, or in adulthood as lung disease caused by elastase-mediated tissue destruction – early-onset lung disease and liver cirrhosis.

Several isoforms of α_1-antitrypsin exist as a result of allelic variation of the AA1T gene: the normal isoform is known as M and the two common defective isoforms as S and Z; the null allele produces no α_1-antitrypsin.

More than 90 allelic variants of the AA1T gene, at the so-called proteinase inhibitor (Pi) locus, have been described, the majority of which do not affect plasma levels or activity of AA1T. Phenotypic variants in AA1T were initially described by their relative mobility on electrophoresis, with the most common variant, M, having medium mobility. The Z and S variants are most frequently associated with AA1T deficiency, and are both due to point mutations which can also be detected by PCR assays.

Liver cancer is associated with particularly high plasma concentrations of α-fetoprotein

α-Fetoprotein (AFP) and albumin have considerable sequence homology, and appear to have evolved by reduplication of a single ancestral gene. In the fetus, AFP appears to serve physiologic functions similar to those performed by albumin in the adult; furthermore, by the end of the first year of life, AFP in the plasma is entirely replaced by albumin. During hepatic regeneration and proliferation, AFP is again synthesized; thus high plasma concentrations of AFP are observed in liver cancer.

There are a number of genetic disorders that impair bilirubin conjugation or secretion

Gilbert's syndrome affects up to 5% of the population, and causes a mild unconjugated hyperbilirubinemia that is harmless and asymptomatic. It is caused by a dinucleotide polymorphism in the TATA box promoter (Fig. 34.1) of the bilirubin UDP-glucuronyl transferase gene which impairs the hepatic uptake of unconjugated bilirubin.

Other inherited diseases of bilirubin metabolism are rare. **Crigler–Najjar syndrome**, which is the result of a complete absence or marked reduction in bilirubin conjugation, causes severe unconjugated hyperbilirubinemia that presents at birth; when the enzyme is completely absent, the condition is fatal.

The **Dubin–Johnson** and **Rotor's** syndromes impair biliary secretion of conjugated bilirubin and therefore cause conjugated hyperbilirubinemia, which is usually mild.

Summary

■ The liver plays a central role in human metabolism.
■ It is extensively involved in the synthesis and catabolism of carbohydrates, lipids and proteins.
■ It synthesizes an array of acute phase proteins in response to inflammation and infection, and laboratory measurements of such proteins are clinically useful in monitoring disease progress.
■ It is involved in the metabolism of bilirubin derived from the catabolism of heme.
■ Disease processes often cause the patient to present with jaundice due to hyperbilirubinemia.
■ The liver has a central role in the detoxification of drugs.
■ Its biochemical function is assessed in clinical practice using a panel of blood tests, called liver function tests, abnormalities of which can point to disease affecting the hepatocellular or biliary systems.

Active learning

1. Discuss how the anatomic position and structure of the liver allow it to absorb and metabolize lipids, proteins and carbohydrates, as well as xenobiotics from the intestine, before releasing such molecules or their derivatives to the systemic circulation.
2. Describe the function of the liver in protein synthesis and in the systemic response to inflammation.
3. Outline how the liver processes bilirubin, and describe the biochemical causes of hyperbilirubinemia (jaundice) and its classification.
4. How does the liver metabolize drugs?
5. Discuss biochemical tests used by the clinical laboratory in the investigation of liver disease.

Further reading

Bernal, W, Auzinger, G, Dhawan, A, *et al.* Acute liver failure. *Lancet*. 2010; 376:190–201.

Donohue, TM, Cederbaum, AI, French, SW. Role of the proteasome in ethanol-induced liver pathology. *Alcohol Clin Exp Res*. 2007; 31:1446–1459.

Hall, P, Cash, J, What is the real function of the liver 'function' tests? *Ulster Med J*. 2012;81(1):30–3. www.ums.ac.uk/umj081/081(1)030.pdf

Puy, H, Gouya, L, Deybach, J-C. Porphyrias. *Lancet*. 2010; 375:924–937.

Schuckit, MA. Alcohol-use disorders. *Lancet*. 2009; 373:492–501.

Wijnen, PAHM, Op den Buijsch, RAM, Drent, M, *et al.* Review article: the prevalence and clinical relevance of cytochrome P450 polymorphisms. *Aliment Pharmacol Ther*. 2007; 26(suppl2):211–219.

Wilkinson, GR. Drug metabolism and variability among patients in drug response. *N Engl J Med*. 2005; 352:2211–2221.

Websites

Liver diseases. MedlinePlus. www.nlm.nih.gov/medlineplus/liverdiseasesgeneral.html.

Liver disease. Lab Tests Online www.labtestsonline.org/understanding/conditions/liver_disease.html.

Liver function tests. Lab Tests Online www.labtestsonline.org.uk/understanding/analytes/liver-panel.

PharmGKB. The Pharmacogenomics Knowledgebase. www.pharmgkb.org/index.jsp.

CHAPTER 31

Biosynthesis and Degradation of Nucleotides

Alejandro Gugliucci and Robert Thornburg

Learning objectives

After reading this chapter you should be able to:

■ Compare and contrast the structure and biosynthesis of purines and pyrimidines, highlighting differences between de novo and salvage pathways.

■ Describe how cells meet their requirements for nucleotides at various stages in their cell cycle.

■ Explain the biochemical rationale for using fluorouracil and methotrexate in chemotherapy.

■ Describe the metabolic basis and therapy for classic disorders in nucleotide metabolism: gout, Lesch–Nyhan syndrome and SCIDS.

Introduction

Nucleotides, molecules composed of a pentose, a nitrogenous base and phosphate, are key elements in cell physiology since they are:

- precursors of deoxyribonucleic acid (DNA) and ribonucleic acid (RNA);
- components of coenzymes, e.g NAD(H), NADP(H), FMN(H_2), and CoA;
- energy currency, driving many metabolic processes, *e.g.* ATP and GTP;
- carriers in biosynthesis, *e.g.* UDP for carbohydrates and CDP for lipids;
- modulators of allosteric regulation of metabolism;
- second messengers, *e.g.* cAMP and cGMP.

We can synthesize ample amounts of purine and pyrimidine nucleotides from metabolic intermediates. In this way, although we ingest dietary nucleic acids and nucleotides, survival does not require their absorption and utilization. Because nucleotides are involved in so many levels of metabolism they are important targets for chemotherapeutic agents used in treatment of microbial and parasitic infections and cancer.

This chapter will describe the structure and metabolism of the two classes of nucleotides: purines and pyrimidines. The metabolic pathways are divided into four sections:

- De novo synthesis of nucleotides from basic metabolites, which is required in growing cells.
- Salvage pathways that recycle preformed bases and nucleosides and provide an adequate supply of nucleotides for cells at rest.
- Catabolic pathways for excretion of nucleotide degradation products, a process that is essential to limit the accumulation of toxic levels of nucleotides within cells: impaired elimination or increased production of uric acid may produce gout.
- Biosynthetic pathways for conversion of the ribonucleotides into the deoxyribonucleotides, providing precursors for DNA.

Purines and pyrimidines

Nucleotides are formed from three components: a nitrogenous base, a five-carbon sugar, and phosphate

The nitrogenous bases found in nucleic acids belong to one of two heterocyclic

groups, either purines or pyrimidines (Fig. 31.1). The major purines of both DNA and RNA are guanine and adenine. In DNA, the major pyrimidines are thymine and cytosine, while in RNA, they are uracil and cytosine; thymine is unique to DNA and uracil is unique to RNA.

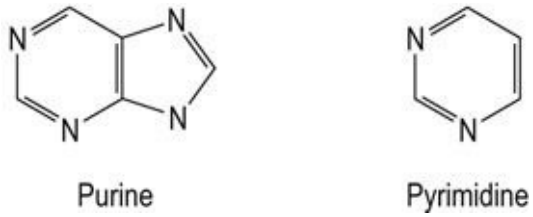
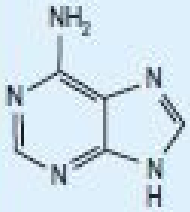
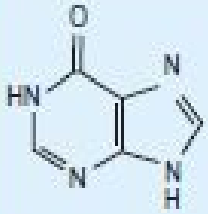
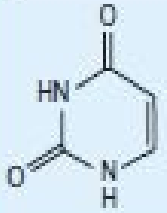
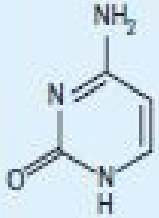
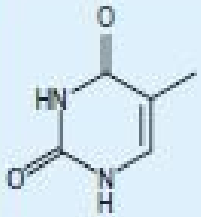


FIG. 31.1 Classification of nucleotides.
Basic structure of purines and pyrimidines.

When the nitrogenous bases are combined with a five-carbon sugar, they are known as nucleosides. When the nucleosides are phosphorylated, the compounds are known as nucleotides. The phosphate can be attached either at the 5′position or the 3′-position of ribose, or both. Table 31.1 gives the names and structures of the most important purines and pyrimidines.

Table 31.1

Names and structures of important purines and pyrimidines

Structure	Free base	Nucleoside	Nucleotide
	adenine	adenosine	AMP ADP ATP cAMP
	guanine	guanosine	GMP GDP GTP cGMP
	hypoxanthine	inosine	IMP
	uracil	uridine	UMP UDP UTP
	cytosine	cytidine	CMP CDP CTP
	thymine	thymidine	TMP TDP TTP

The designation NTP refers to the ribonucleotide. The prefix d, as in dATP, is used to identify

deoxyribonucleotides. dTTP is usually written as TTP, with the d-prefix implied.

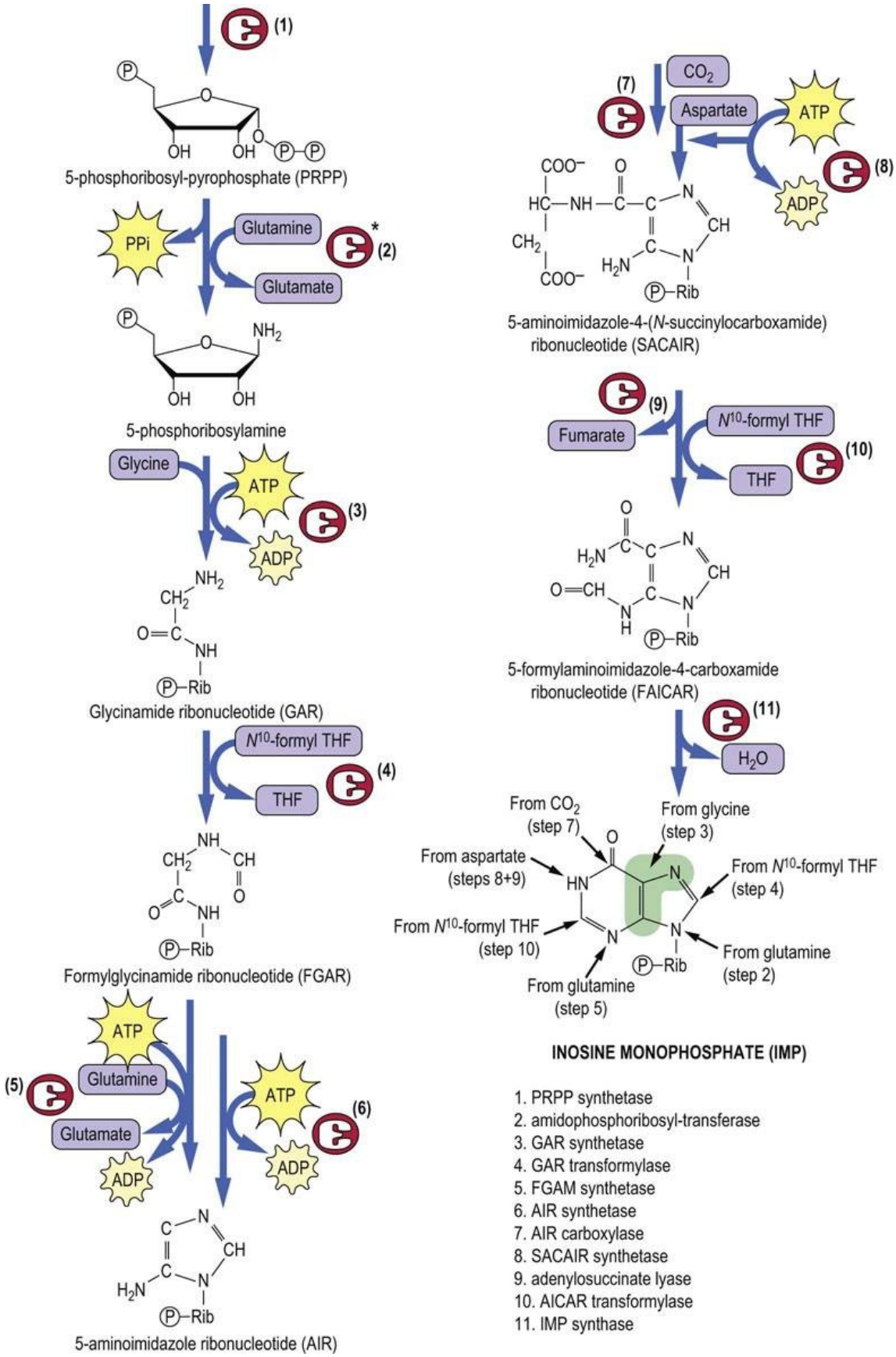
Purine metabolism

De novo synthesis of the purine ring: synthesis of inosine monophosphate (IMP)

Purines and pyrimidines are synthesized by both de novo and salvage pathways

The demand for nucleotide biosynthesis can vary greatly. It is high during the S-phase of the cell cycle, when cells are about to divide (Chapter 42). The process is therefore very active in growing tissues, actively proliferating cells like blood cells and cancer cells, and when tissues are regenerating. Purine and pyrimidine biosynthesis are energetically expensive processes that are subject to intracellular mechanisms that sense and effectively regulate the pool sizes of intermediates and products.

The raw materials for purine synthesis are: CO_2, nonessential amino acids (Asp, Glu, Gly), and folic acid derivatives which act as single carbon donors. Five molecules of ATP are needed for the synthesis of IMP, the first purine product and common precursor of AMP and GMP. The starting material for synthesis of IMP is ribose 5-phosphate, a product of the pentose phosphate pathway (Chapter 12). The first step, catalyzed by **ribose phosphate pyrophosphokinase (PRPP synthetase),** generates the activated form of the pentose phosphate by transferring a pyrophosphate group from ATP to form 5-phosphoribosyl-pyrophosphate (PRPP) (Fig. 31.2). In a series of 10 reactions, PRPP is converted to IMP. Most of the carbons and all of the nitrogens of the purine ring are derived from the amino acids; one carbon is derived from CO_2 and two from **N^{10}-formyltetrahydrofolate (THF)**, a derivative of **folic acid**. Folate deficiency can impair purine synthesis, which can produce disease or can be exploited clinically to kill rapidly dividing cells, which have a high demand for purine biosynthesis. The end product of this sequence of reactions is the ribonucleotide IMP; the nucleoside is inosine and the purine base is called hypoxanthine.



1. PRPP synthetase
2. amidophosphoribosyl-transferase
3. GAR synthetase
4. GAR transformylase
5. FGAM synthetase
6. AIR synthetase
7. AIR carboxylase
8. SACAIR synthetase
9. adenylosuccinate lyase
10. AICAR transformylase
11. IMP synthase

FIG. 31.2 Synthesis of IMP.
*The asterisk identifies the regulatory enzyme amidophosphoribosyl transferase (2).

Synthesis of ATP and GTP from IMP

IMP does not accumulate significantly within the cell. As shown in Figure 31.3, it is converted to both AMP and GMP. Two enzymatic reactions are required in each case (see Fig. 31.3). Distinct enzymes, adenylate kinase and guanylate kinase, use ATP to synthesize the nucleotide diphosphates from the nucleotide monophosphates. Finally, a single enzyme, termed **nucleotide diphosphokinase**, converts diphosphonucleotides into nucleotide triphosphates. This enzyme has activity towards all nucleotide diphosphates, including pyrimidines and purines and both ribo-and deoxyribonucleotides for synthesis of RNA and DNA, respectively.

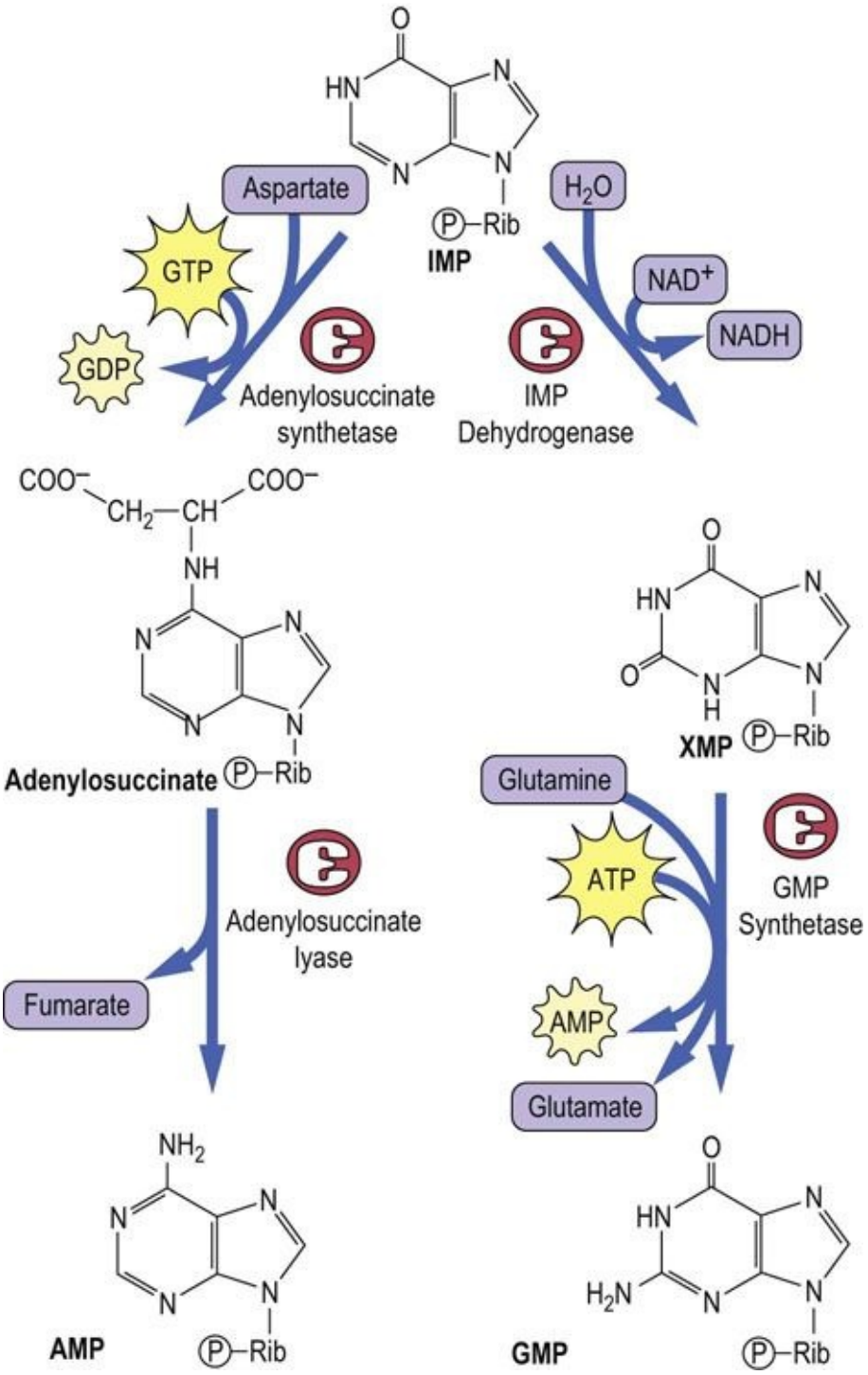


FIG. 31.3 Conversion of IMP into AMP and GMP.
Two enzymatic reactions are needed in each branch of the pathway. XMP, xanthosine monophosphate.

Salvage pathways for purine nucleotide biosynthesis

In addition to de novo synthesis, cells can use preformed nucleotides obtained from the diet or from the breakdown of endogenous nucleic acids through salvage pathways. In mammals, there are two enzymes in the purine salvage pathway. **Adenine phosphoribosyl transferase (APRT)** converts free adenine into AMP (Fig. 31.4A). **Hypoxanthine-guanine phosphoribosyl transferase (HGPRT)** catalyzes a similar reaction for both hypoxanthine (the purine base in IMP) and guanine (Fig. 31.4B). Purine nucleotides are synthesized preferentially by salvage pathways, so long as the free nucleobases are available. This preference is mediated by hypoxanthine inhibition of amidophosphoribosyl transferase, step 2 of the de novo pathway. Note that step 2 is the site of inhibition of purine biosynthesis, since PRPP is also used in other biosynthetic processes, including nucleotide salvage pathways.

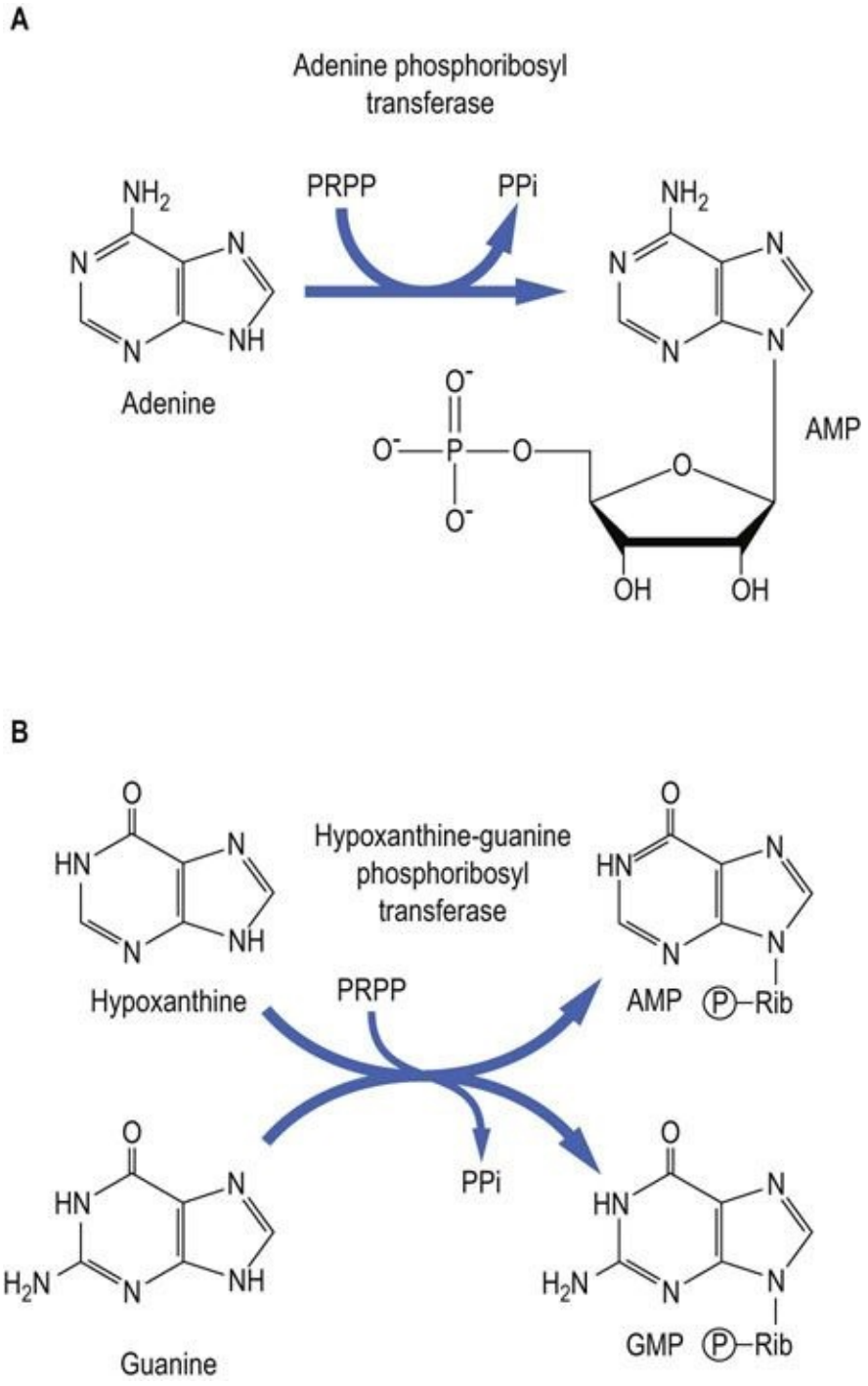


FIG. 31.4 The purine salvage pathways.
(A) adenine phosphoribosyl transferase. (B) Hypoxanthine-guanine phosphoribosyl transferase.

Purine and uric acid metabolism in humans

Sources and disposal of uric acid

Uric acid is the endproduct of purine catabolism in humans

Uric acid, the end product of purine catabolism in humans, is not metabolized and must be excreted. However, the complex renal handling of urate, described below, suggests an evolutionary advantage to having high circulating levels of urate. As noted in Chapter 37, uric acid is a circulating antioxidant. At pH 7.4, it is 98% ionized and therefore circulates as monosodium urate. This salt has poor solubility; the extracellular fluid becoming saturated at urate concentrations little above the upper limit of the reference range. Therefore, there is a tendency for monosodium urate to crystallize in subjects with **hyperuricemia.** The most obvious clinical manifestation of this process is gout, in which crystals form in cartilage, synovium and synovial fluid. This can be accompanied by **renal calculi** (urate stones) and **tophi** (accumulation of sodium urate in soft tissues). A sudden increase in urate production, for example during chemotherapy when many cells die rapidly, can lead to widespread crystallization of urate in joints, but mainly in urine, causing an acute **urate nephropathy**.

There are three sources of purines in man: de novo synthesis, salvage pathways, and diet. The body urate pool (and thus plasma uric acid concentration) is governed by the relative rates of urate formation and excretion. Over half of urate is excreted by the kidney, the rest by the intestines, where bacteria dispose of it. In the kidney, urate is filtered and almost totally reabsorbed in the proximal tubule. Distally, both secretion and absorption occur, so that overall urate clearance amounts to about 10% of the filtered load, *i.e.* 90% is retained in the body. Normally, urate excretion increases if the filtered load is increased. Because of the role of the kidney in urate metabolism, kidney diseases can lead to urate retention and urate precipitation in the kidney (stones) and urine. Dietary purines account for about 20% of excreted urate. Therefore, restricting purines in the diet (less meat and red wine) can reduce urate levels by only 10–20%.

Advanced concept box Salvage pathways are the principal source of nucleotides in lymphocytes

In humans, resting T lymphocytes, immune system cells produced in the thymus (Chapter 38), meet their routine metabolic

requirements for nucleotides through the salvage pathway, but de novo synthesis is required to support growth of rapidly dividing cells. The salvage of nucleotides is especially important in HIV-infected T lymphocytes. In asymptomatic patients, resting lymphocytes show a block in de novo pyrimidine biosynthesis, and correspondingly reduced pyrimidine pool sizes. Following activation of the T-lymphocyte population, these cells cannot synthesize sufficient new DNA. The activation process leads to cell death, contributing to the decline in the T-lymphocyte population during the late stages of HIV infection.

The salvage pathways are especially important for many parasites as well. These organisms prey metabolically on their host, using preformed metabolites, including nucleotides. Some parasites, such as *Mycoplasma*, *Borrelia*, and *Chlamydia*, have lost the genes required for the de novo synthesis of nucleotides; they obtain these important components from their host.

Clinical box Gout results from the excess of uric acid

Diagnosis.

The diagnosis of gout is primarily clinical and supported by the demonstration of hyperuricemia. About 90% of patients appear to excrete urate at an inappropriately low rate for the plasma concentration, while about 10% have excessive production. Gouty arthritis has a typically hyperacute onset (less than 24 h), with severe pain, swelling, redness and warmth in the joint(s), characteristically in the big toe. It is confirmed by the presence of tophi or sodium urate crystals in the synovial fluid. The crystals are needle-shaped, are seen inside neutrophils and have negative birefringence when viewed with polarized light.

Pathogenesis.

Urate crystals in joints are phagocytized by neutrophils (leukocytes in blood and tissues). The crystals damage lysosomal and cellular membranes, causing cellular disruption and death. Release of lysosomal enzymes in the joint precipitates an acute inflammatory reaction. Several cytokines enhance and perpetuate the inflammation and phagocytic cells, monocytes and macrophages aggravate the inflammation.

Management.

The acute attack is managed with antiinflammatory agents, including NSAIDs. Dietary changes (less meat, increased water intake, weight reduction) and changes in concurrent drug therapies, such as diuretics, may be helpful. Probenecid, a uricosuric drug, is commonly employed to reduce uricemia. Colchicine, a microtubule disruptor, may also be used during an acute attack to inhibit phagocytosis and inflammation. If the patient is already a hyperexcretor, or tophi or renal disease are present, then allopurinol is used. Allopurinol is an inhibitor of xanthine oxidase (Fig. 31.5). Allopurinol undergoes the first oxidation to yield alloxanthine, but cannot undergo the second oxidation. Alloxanthine remains bound to the enzyme, acting as a potent competitive inhibitor. This leads to reduced formation of uric acid and accumulation of xanthine and hypoxanthine, which are more soluble and are excreted in urine.

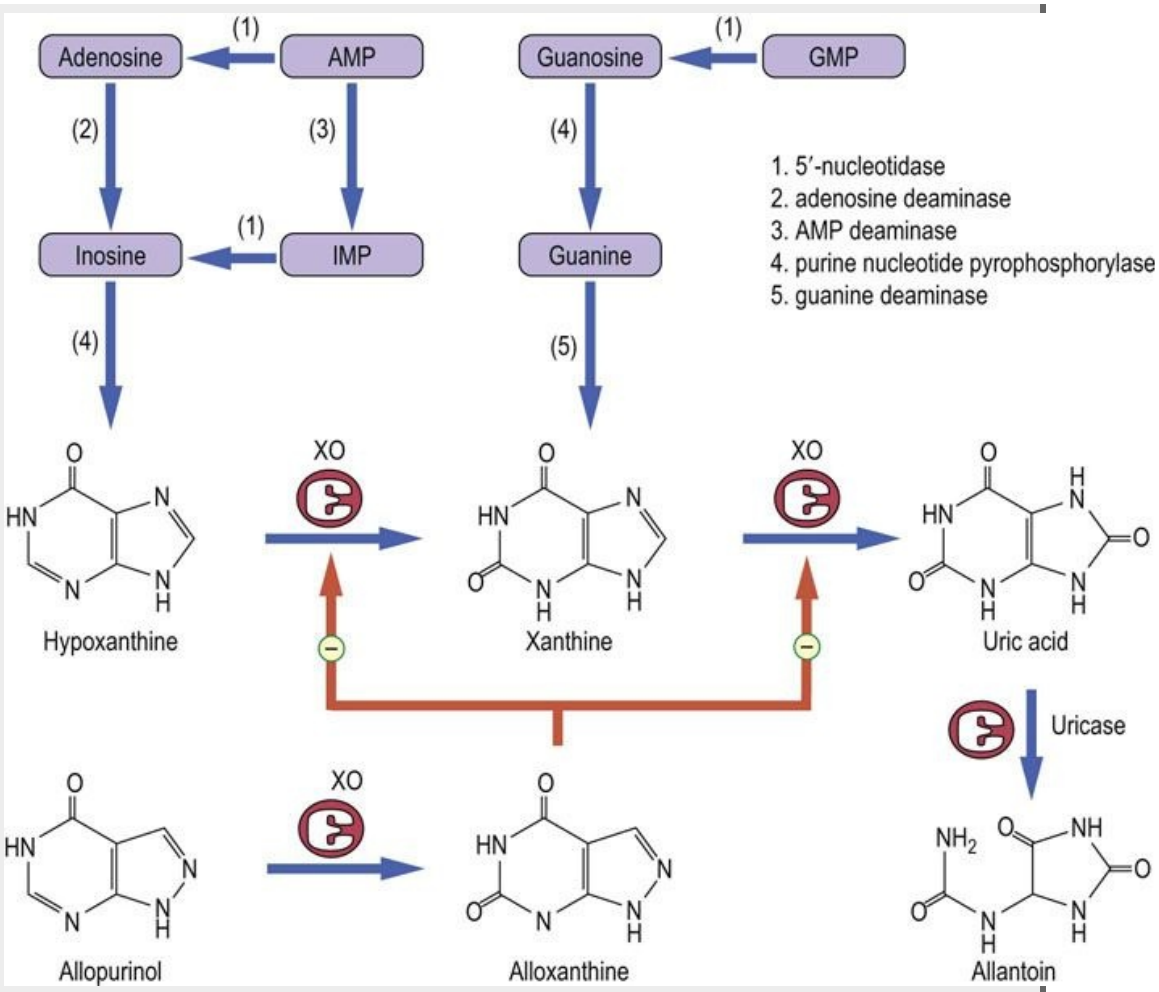


FIG. 31.5 Degradation of purines and biochemical basis of allopurinol treatment of gout.
Inhibition of xanthine oxidase (XO) by alloxanthine is the mechanism involved in **allopurinol** treatment of gout. The enzyme uricase is missing in primates (including humans) but is commonly used for measurement of serum uric acid levels in humans. (1) 5′-nucleotidase; (2) adenosine deaminase; (3) AMP deaminase; (4) purine nucleotide pyrophosphorylase; (5) guanine deaminase.

Endogenous formation of uric acid

Each of the purine monophosphates (IMP, GMP and AMP) can be converted into their corresponding nucleosides by 5′-nucleotidase. The enzyme purine nucleoside phosphorylase then converts the nucleosides inosine or guanosine

into the free purine bases hypoxanthine and guanine, and ribose-1-P. Hypoxanthine is oxidized and guanine is deaminated to yield xanthine (Fig. 31.5). Two other enzymes, **AMP deaminase** and **adenosine deaminase**, convert the amino group of AMP and adenosine into IMP and inosine, respectively, which are then converted to hypoxanthine. In effect, guanine is directly converted to xanthine, while inosine and adenine are converted to hypoxanthine, then to xanthine.

Xanthine oxidase (XO), the final enzyme in this pathway, catalyzes a two-step oxidation reaction, converting hypoxanthine to xanthine, then xanthine to uric acid. Uric acid is the final metabolic product of purine catabolism in primates, birds, reptiles, and many insects. Other organisms, including most mammals, fish, amphibians and invertebrates, metabolize uric acid to more soluble products, such as allantoin (see Fig. 31.5).

Hyperuricemia and gout

Most persons with hyperuricemia remain asymptomatic throughout life, but there is no gout without hyperuricemia

Plasma urate concentration is, on average, higher in men than in women, tends to rise with age, and is usually elevated in obese subjects and subjects in the higher socioeconomic groups. Higher levels of uric acid correlate with high sugar consumption. The risk of gout, a painful disease resulting from precipitation of sodium urate crystals in joints and dermis, increases with higher plasma urate concentrations (see Box on p. 414). Hyperuricemia can occur due to increased formation or decreased excretion of uric acid, or both. Decreased renal excretion of urate can result from a decrease in filtration and/or secretion. Many factors (including drugs and alcohol) also affect tubular handling of urates and can cause or increase hyperuricemia.

Pyrimidine metabolism

As with the purines, the pyrimidines (uracil, cytosine and thymine) are also synthesized through a complex series of reactions using raw materials readily available in cells. One important difference is that the pyrimidine base is made first and the sugar added late (Fig. 31.6), whereas purines are assembled on a ribose-5-P scaffold (Fig. 31.2). Uridine monophosphate (UMP) is the precursor of all pyrimidine nucleotides. The de novo pathway produces UMP, which is then converted to cytidine triphosphate (CTP) and thymidine triphosphate (TTP). Salvage pathways also recover preformed pyrimidines.

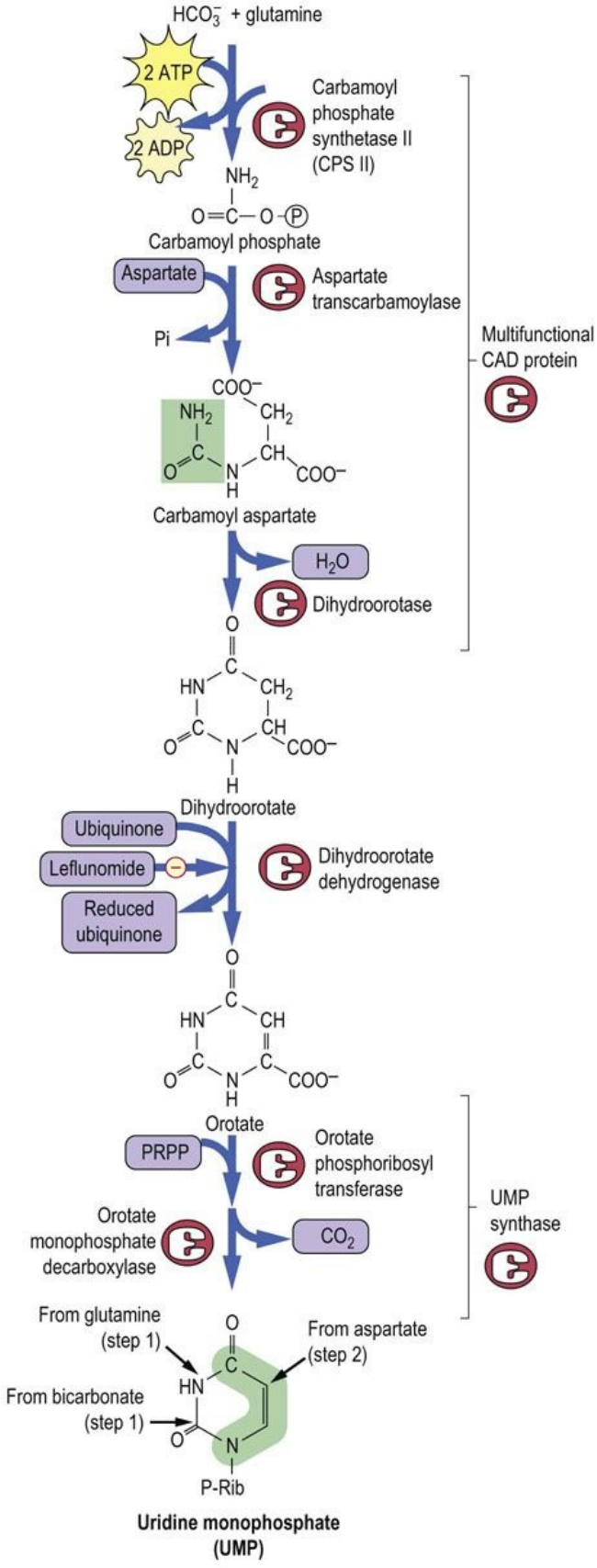

HCO_3^- + glutamine
2 ATP
2 ADP
Carbamoyl phosphate synthetase II (CPS II)
Carbamoyl phosphate
Aspartate
Aspartate transcarbamoylase
Pi
Multifunctional CAD protein
Carbamoyl aspartate
H_2O
Dihydroorotase
Dihydroorotate
Ubiquinone
Leflunomide
Reduced ubiquinone
Dihydroorotate dehydrogenase
Orotate
PRPP
Orotate phosphoribosyl transferase
Orotate monophosphate decarboxylase
CO_2
UMP synthase
From glutamine (step 1)
From bicarbonate (step 1)
From aspartate (step 2)
P-Rib
Uridine monophosphate (UMP)

FIG. 31.6 The metabolic pathway for synthesis of pyrimidines. Formation of orotic acid and UMP, the first pyrimidine nucleotide.

Clinical box Lesch–nyhan syndrome – HGPRT deficiency

The gene encoding HGPRT is located on the X-chromosome. Its deficiency results in a rare, X-linked recessive disorder, Lesch–Nyhan syndrome. The lack of HGPRT causes an overaccumulation of PRPP, which is also the substrate for the enzyme amidophosphoribosyl transferase. This stimulates purine biosynthesis by up to 200-fold. Because of increased purine synthesis, the degradation product, uric acid, also accumulates to high levels. Elevated uric acid leads to a crippling gouty arthritis and severe neuropathology, resulting in mental retardation, spasticity, aggressive behavior, and a compulsion towards self-mutilation by biting and scratching.

De novo pathway

Pyrimidine and purine nucleoside biosynthesis share several common precursors: CO_2, amino acids (Asp, Gln), and, for thymine, **N^5,N^{10}-methylene-THF**. The pathway for biosynthesis of UMP is outlined in Figure 31.6. The first step, catalyzed by the enzyme **carbamoyl phosphate synthetase II (CPS II)**, uses bicarbonate, glutamine, and 2 moles of ATP to form carbamoyl phosphate (CPS I is used in the synthesis of arginine in the urea cycle; Chapter 19). Most of the atoms required for formation of the pyrimidine ring are derived from aspartate, added in a single step by aspartate transcarbamoylase (ATCase). Carbamoyl aspartate is then cyclized to dihydroorotic acid by the action of the enzyme dihydroorotase. Dihydroorotic acid is oxidized to orotic acid by a mitochondrial enzyme, dihydroorotate dehydrogenase. **Leflunomide**, a specific inhibitor of this enzyme, is used for treatment of rheumatoid arthritis because blockage of this step inhibits lymphocyte activation and thereby limits inflammation. The ribosyl-5′-phosphate group from PRPP is then transferred

onto orotic acid to form orotidine monophosphate (OMP). Finally, OMP is decarboxylated to form UMP. UTP is synthesized in two enzymatic phosphorylation steps by the actions of UMP kinase and nucleotide diphosphokinase. CTP synthetase converts UTP into CTP by amination of UTP (Fig. 31.7). This completes the synthesis of the ribonucleotides for synthesis of RNA.

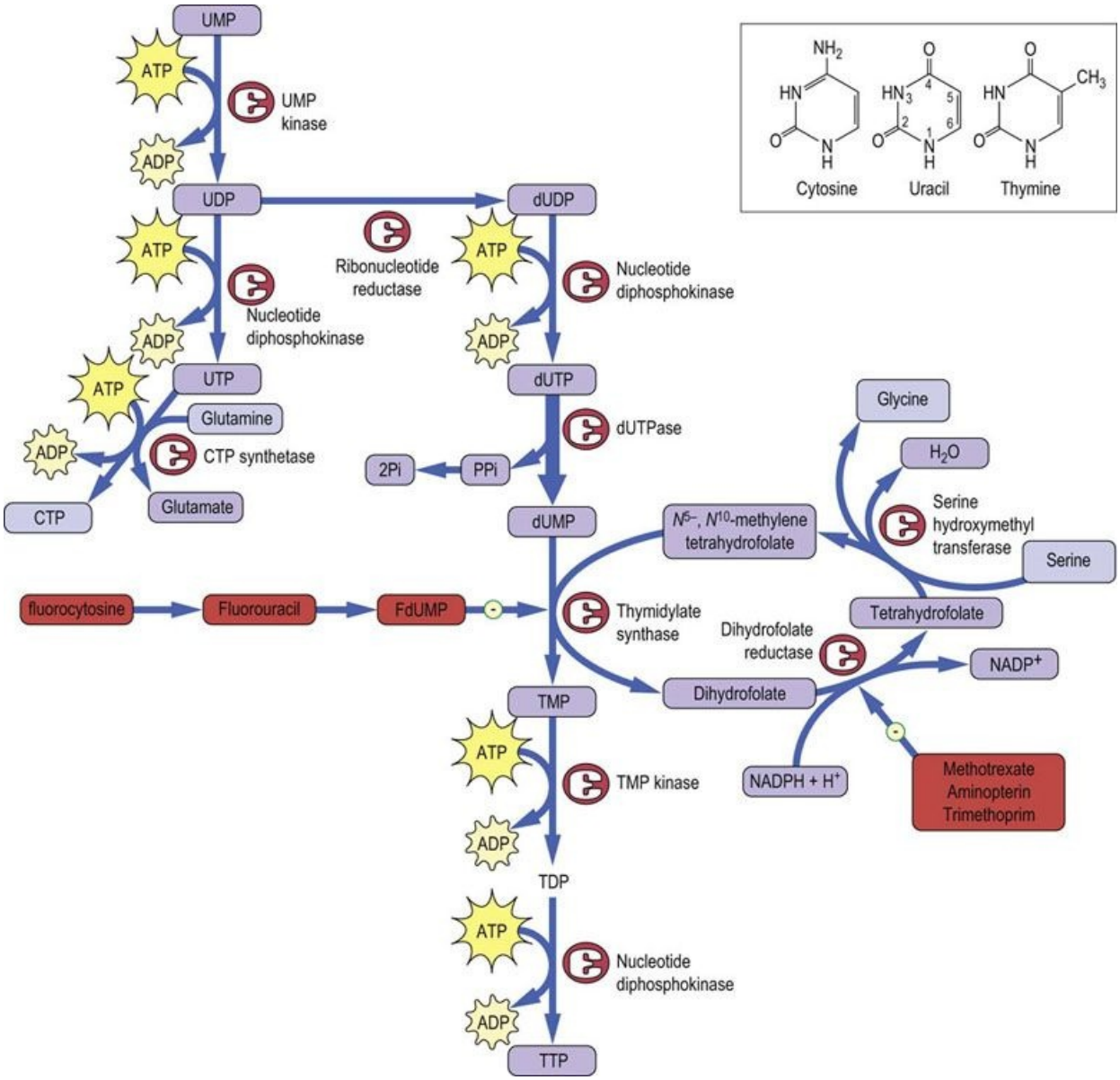


FIG. 31.7 Synthesis of pyrimidine triphosphates.
Synthesis of thymidine is inhibited by fluorodeoxyuridylate (FdUMP), methotrexate, aminopterin and trimethoprim at the indicated sites.

Metabolic channeling by multienzymes improves efficiency

In bacteria, the six enzymes of pyrimidine (UMP) biosynthesis exist as distinct proteins. However, during the evolution of mammals the first three enzymatic activities have been fused together into **CAD**, a single multifunctional polypeptide encoded by a single gene. The name of the enzyme derives from its three activities: **Carbamoyl phosphate synthetase, Aspartate transcarbamoylase, and Dihydroorotase**. The final two enzymatic activities of pyrimidine biosynthesis, orotate phosphoribosyl transferase and orotidylate decarboxylase, are also fused into a single enzyme, UMP synthase. As with the fatty acid synthase complex (Chaper 16), this fusion of sequential enzyme activities avoids the diffusion of the metabolic intermediates into the intracellular milieu, thereby improving the metabolic efficiency of the individual steps.

Pyrimidine salvage pathways

As with the purines, free pyrimidine bases, available from the diet or from the breakdown of nucleic acids, can be recovered by several salvage enzymes. Uracil phosphoribosyl transferase (UPRTase) is similar to the enzymes of the purine salvage pathways. This also activates some chemotherapeutic agents such as 5-fluorouracil (FU) or 5-fluorocytosine (FC). A uridine-cytidine kinase and a more specific thymidine kinase catalyze the phosphorylation of these nucleosides; nucleotide kinases and diphosphokinase complete the salvage process.

Formation of deoxynucleotides

Ribonucleotide reductase

Ribonucleotide reductase catalyzes reduction of ribose to deoxyribose in nucleotides for synthesis of DNA

Because DNA uses deoxyribonucleotides instead of the ribonucleotides found in RNA, cells require pathways to convert ribonucleotides into the deoxy forms. The adenine, guanine and uracil deoxyribonucleotides are synthesized from their corresponding ribonucleotide diphosphates by direct reduction of the 2′-hydroxyl by the enzyme ribonucleotide reductase, as shown for dUDP in Figure 31.7. The reduction of the 2′-hydroxyl of ribose uses a pair of protein-bound sulfhydryl groups (cysteine residues). The hydroxyl group is released as water, and the cysteines are oxidized to cystine during the reaction. To regenerate an active enzyme, the disulfide must be reduced back to the original sulfhydryl pair by disulfide exchange; this is accomplished by reaction with a small protein, **thioredoxin**. The thioredoxin, a highly conserved Fe-S protein, is in turn reduced by the flavoprotein, thioredoxin reductase.

A unique pathway to TTP

Thymine is synthesized by a complex reaction pathway, providing many opportunities for chemotherapy

The nucleotide deoxy-TMP, abbreviated as TMP because thymine is unique to DNA, is synthesized by a special pathway involving methylation of the deoxyribose form of uridylate, dUMP (Fig. 31.7). The TMP biosynthetic pathway leads from UMP to UDP, then, through ribonucleotide reductase, to dUDP. The dUDP is then phosphorylated to dUTP, which creates an unexpected biochemical problem. DNA polymerase does not effectively discriminate between the two deoxyribonucleotides, dUTP and TTP – the only difference is a methyl group at *C*-5. It incorporates dUTP into DNA in vitro; this reaction would lead to high rates of mutagenesis in vivo. Therefore, cells limit the concentration of dUTP by rapidly hydrolyzing dUTP to dUMP with the enzyme dUTPase. This enzyme cleaves a high-energy bond and releases pyrophosphate,

which is rapidly hydrolyzed to phosphate, shifting the equilibrium ever further towards the formation of dUMP. The dUMP is converted to TMP by **thymidylate synthase (TS)**, using N^5,N^{10}-methylene-THF as the methyl donor; the dihydrofolate product is recycled by action of the enzymes **dihydrofolate reductase** and serine hydroxymethyl transferase. Two rounds of phosphorylation of TMP yield TTP for synthesis of DNA.

The synthesis of TTP is a roundabout pathway, but provides opportunities for chemotherapy through inhibition of TMP biosynthesis (see Fig. 31.7). There is only one reaction in pyrimidine synthesis that requires a THF derivative: conversion of dUMP to TMP, catalyzed by thymidylate synthase. This reaction is often rate limiting for cell division. Indeed, folate deficiency impairs cell replication, especially the replication of rapidly dividing cells. Thus, folate deficiency is a frequent cause of anemia: bone marrow cells involved in erythropoiesis and hematopoiesis are among the most rapidly dividing cells in the body. As outlined in the Advanced Concept Box on the next page, inhibition of thymidylate synthase, either directly or by inhibition of THF recycling, provides a special opportunity for chemotherapy,

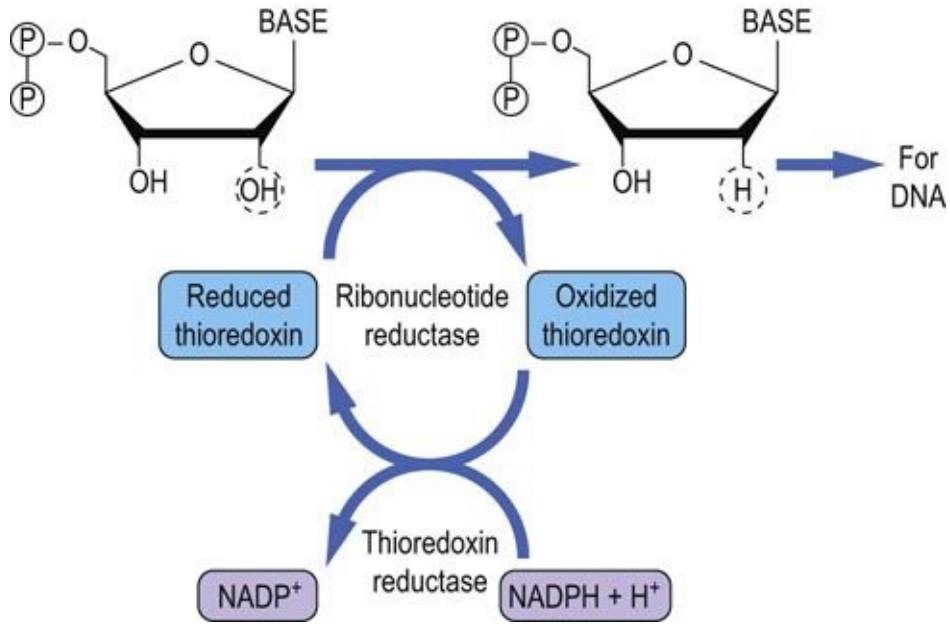
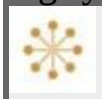


FIG. 31.8 Formation of deoxyribonucleotides, except TTP, by ribonucleotide reductase. Thioredoxin and NADPH (from the pentose phosphate pathway) are required for recycling of the enzyme.

targeting synthesis of DNA precursors in rapidly dividing cancer cells.



Advanced concept box
Chemotherapeutic targets: folate recycling and thymidylate synthase

Fluorodeoxyuridylate (FdUMP) is a specific, **suicide inhibitor** of thymidylate synthase. In FdUMP, a highly electronegative fluorine replaces the carbon-5 proton of uridine. This compound can begin the enzymatic conversion into dTMP by forming the enzyme–FdUMP covalent complex; however, the covalent intermediate cannot accept the donated methyl group from methylene THF, nor can it be broken down to release the active enzyme. The result is a suicide complex in which the substrate is covalently locked at the active site of thymidylate synthase. The drug is frequently administered as **flurouracil**, and the body's normal metabolism converts the fluorouridine into FdUMP. Fluorouracil is used against breast, colorectal, gastric, and uterine cancers.

Fluorocytosine is a potent antimicrobial agent. Its mechanism of action is similar to that of FdUMP; however, it must first be converted into fluorouracil by the action of **cytosine deaminase**. The fluorouracil is subsequently converted into FdUMP, which blocks thymidylate synthase as above. While cytosine deaminase is present in most fungi and bacteria, it is absent in animals and plants. Therefore, in humans, fluorocytosine is not converted into fluorouracil and is nontoxic, while in the microbes, metabolism of fluorocytosine results in cell death.

Aminopterin and **methotrexate** are folic acid analogues that bind about 1000-fold more tightly to dihydrofolate reductase (DHFR) than does dihydrofolate. In this manner, they competitively, almost irreversibly, block the synthesis of dTMP. These compounds are also competitive inhibitors of other THF-dependent enzyme reactions used in the biosynthesis of purines, histidine, and methionine. Trimethoprim binds to DHFR, and binds more tightly to bacterial DHFRs than it does to mammalian enzymes, making it an effective antibacterial agent. Folate analogues are relatively nonspecific chemotherapeutic agents.

> They poison rapidly dividing cells, not just cancer cells but also hair follicles and gut epithelial cells, causing the loss of hair and the gastrointestinal side effects of chemotherapy.

De novo nucleotide metabolism is highly regulated

Ribonucleotide reductase is the allosteric enzyme that coordinates a balanced supply of deoxynucleotides for synthesis of DNA

Because nucleotides are required for mammalian cells to proliferate, the enzymes involved in de novo synthesis of both purines and pyrimidines are induced during the S-phase of cell division. Covalent and allosteric regulation also plays an important role in control of nucleotide synthesis. The multimeric protein CAD is activated by phosphorylation by protein kinases in response to growth factors, increasing its affinity for PRPP and decreasing inhibition by UTP. Both of these changes favor biosynthesis of pyrimidines for cell division.

Mole per mole, pyrimidine biosynthesis parallels purine biosynthesis, suggesting the presence of a coordinated control. Among them, one of the key points is the PRPP synthase reaction. PRPP is a precursor for all the ribo-and deoxyribonucleotides. PRPP synthase is inhibited by both pyrimidine and purine nucleotides.

Ribonucleotide reductase coordinates the biosynthesis of all four deoxynucleotides

Because a single enzyme is responsible for the conversion of all ribonucleotides into deoxyribonucleotides, this enzyme is subject to a complex network of feedback regulation. Ribonucleotide reductase contains several allosteric sites for metabolic regulation. Levels of each of the dNTPs modulate the activity of the enzyme toward the other NDPs. By regulating the enzymatic activity of deoxyribonucleotide synthesis as a function of the concentration of the different dNTPs, often described as ‘cross-talk’ between the pathways, the cell insures

that the proper ratios of the different deoxyribonucleotides are produced for normal growth and cell division.

Catabolism of pyrimidine nucleotides

In contrast to the degradation of purines to uric acid, pyrimidines are degraded to readily soluble compounds, which are readily eliminated in urine and are not a frequent source of pathology. Orotic acidurias may occur in the rare cases when enzymes in the pyrimidine catabolic pathways are defective. The pyrimidine nucleotides and nucleosides are converted to the free bases and the heterocyclic ring is cleaved, yielding **β-aminoisobutyrate** as the main excretion product, plus some ammonia and CO_2.

Summary

■ Nucleotides are synthesized primarily from amino acid precursors and phosphoribosyl pyrophosphate by complex, metabolically expensive, multistep pathways.
■ De novo nucleotide metabolism is required for cell proliferation, but salvage pathways also play a prominent role in nucleotide metabolism.
■ Both classes of nucleotides (purines and pyrimidines) are synthesized as precursors (IMP, UMP), which are then converted into the DNA precursors (dATP, dGTP, dCTP, TTP).
■ With the exception of TTP, ribonucleotides are converted to deoxyribonucleotides by ribonucleotide reductase. TTP is synthesized from dUMP by a special pathway involving folates.
■ Salvage pathways have proven useful for the activation of pharmaceutical agents, while the uniqueness of the pathway for synthesis of TTP has provided a special target for chemotherapeutic inhibition of DNA synthesis and cell division in cancer cells.
■ High plasma concentrations of uric acid, the final product of purine catabolism in man, can lead to gout and kidney stones.

Clinical box Severe combined immunodeficiency syndromes (SCIDS) are caused by impaired purine salvage pathways

SCIDS are a group of fatal disorders resulting from defects in both cellular and humoral immune function. SCIDS patients cannot efficiently produce antibodies in response to an antigenic challenge. Approximately 50% of patients with the autosomal recessive form of SCIDS have a genetic deficiency in the purine salvage enzyme **adenosine deaminase**. The pathophysiology involves lymphocytes of both thymic and bone marrow origin (T and B lymphocytes), as well as ‘self-destruction’ of differentiated cells following antigen stimulation. The precise cause of cell death is not yet known, but may involve accumulation in lymphoid

tissues of adenosine, deoxyadenosine and dATP, accompanied by ATP depletion. dATP inhibits ribonucleotide reductase and therefore impedes DNA nucleotide synthesis. The finding that deficiency of the next enzyme in the purine salvage pathway, nucleoside phosphorylase, is also associated with an immune deficiency disorder suggests that integrity of the purine salvage pathway is critical for normal differentiation and function of immunocompetent cells in man.

Advanced concept box The yin and yang of xanthine oxidase – beyond uric acid production

Xanthine oxidase (XO) is a ubiquitous cytosolic flavoprotein which controls the rate-limiting step in purine catabolism. The oxidation of xanthine to uric acid produces $FADH_2$, and the reoxidation of $FADH_2$ produces reactive oxygen species (ROS; superoxide and hydrogen peroxide), which are toxic at high concentrations in the cell. An excess of ROS production is associated with many acute and chronic diseases, including ischemia-reperfusion injury (IRI), cardiovascular disease, microvascular syndromes, metabolic syndrome and cancer (Chapter 37). XO, among other enzymes, contributes to the generation of excess ROS in these conditions. In IRI, the ROS are produced during reperfusion of the tissue, *i.e.* during the recovery phase. Allopurinol, an XO inhibitor, is being evaluated as adjuvant therapy to limit ROS production during recovery from myocardial infarction and stroke. In contrast, XO is also being conjugated to antitumor antibodies in experimental studies to direct ROS production to the tumor environment in order to kill cancer cells. Thus, while XO is a lead actor in purine catabolism, it also plays other roles in pathology and therapy.

Clinical box Reverse evolution: uricase as a new treatment for refractory gout

Most of the drugs employed to treat patients with gout have been used for over 40 years. More recently, as our physiologic understanding of gout has improved, new innovative treatments have been developed, such as enzyme therapy by administration of uricase. Pegloticase, a recombinant uricase, is a novel treatment option for patients suffering from chronic and refractory gout. Human trials show that pegloticase maintains uric acid concentrations below 7 mg/dL in patients with chronic gout. These recombinant uricases may have a place in the treatment of gout, particularly in patients with severe and tophaceous gout to promote tophi dissolution. Pegloticase is an effective option for patients with symptomatic gout for whom existing hypouricemic agents are unsuccessful or contraindicated.

Active learning

1. Compare the roles of de novo synthesis and salvage pathways of nucleotide synthesis in various cell types, *e.g.* erythrocytes, lymphocytes, muscle and liver.
2. In addition to its activity as a xanthine oxidase inhibitor, what other activities of allopurinol might contribute to its efficacy for treatment of gout?
3. Discuss the use of thymidylate synthetase inhibitors and folate analogues for treatment of diseases other than cancer, *e.g.* arthritis, psoriasis.

Further reading

Agarwal, A, Banerjee, A, Banerjee, UC. Xanthine oxidoreductase: a journey from purine metabolism to cardiovascular excitation-contraction coupling. *Crit Rev Biotechnol*. 2011; 31:264–280.

Fang, J, Nakamura, H, Iyer, AK. Tumor-targeted induction of oxystress for cancer therapy. *J Drug Target*. 2007; 15:475–486.

Gangjee, A, Jain, HD, Kurup, S. Recent advances in classical and non-classical antifolates as antitumor and antiopportunistic infection agents. *Anticancer Agents Med Chem*. 2008; 8:205–231.

Garay, RP, El-Gewely, MR, Labaune, JP, Richette, P. Therapeutic perspectives on uricases for gout. *Joint Bone Spine*. 2012; 79:237–242.

Jordan, KM. Up-to-date management of gout. *Curr Opin Rheumatol*. 2012; 24:145–151.

Jurecka, A. Inborn errors of purine and pyrimidine metabolism. *J Inherit Metab Dis*. 2009; 32:247–263.

Lee, BE, Toledo, AH, Anaya-Prado, R, *et al.* Allopurinol, xanthine oxidase, and cardiac ischemia. *J Investig Med*. 2009; 57:902–909.

Nyhan, WL. The recognition of Lesch–Nyhan syndrome as an inborn error of purine metabolism. *J Inherited Metab Dis*. 1997; 20:171–178.

Shannon, JA, Cole, SW. Pegloticase: a novel agent for treatment-refractory gout. *Ann Pharmacother*. 2012; 46:368–376.

Sigoillot, FD, Berkowski, JA, Sigoillot, SM, *et al.* Cell cycle-dependent regulation of pyrimidine biosynthesis. *J Biol Chem*. 2003; 278:3403–3409.

Websites

Pictures of Gout. http://www.mediscan.co.uk/cfm/resultssearch.cfm?action=keyword&log=nk.

Gout. www.niams.nih.gov/Health_Info/Gout/default.asp.

Lesch–Nyhan. http://emedicine.medscape.com/article/1181356-overview.

Purine and pyrimidine metabolism disorders. http://www.merckmanuals.com/professional/pediatrics/inherited_disorders_of_metabolism/purine_and_p

SCIDS. www.scid.net, http://themedicalbiochemistrypage.org/nucleotide-metabolism.php.

CHAPTER 32

Deoxyribonucleic Acid

Robert Thornburg and Alejandro Gugliucci

Learning objectives

After reading this chapter you should be able to:

■ Describe the composition and structure of DNA, based on the Watson–Crick model, including the concepts of directionality and complementarity in DNA structure.

■ Describe the packaging of DNA in the nucleus.

■ Explain how replication of DNA is achieved with high fidelity.

■ Discuss the enzymes involved, the activities at replication forks, and the structures and intermediates participating in the replication process.

■ Outline the mechanism by which replication is controlled in the eukaryotic cell.

■ Describe the types of damage to DNA and the mechanisms involved in DNA repair.

■ Describe the mechanism of action of AZT for treatment of AIDS.

Introduction

Cellular nucleic acids exist in two forms, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA). Approximately 90% of the nucleic acid within cells is RNA and the remainder is DNA. DNA is the repository of genetic information. This chapter deals with the structure of DNA, the manner in which it is stored in chromosomes in the nucleus, and the mechanisms involved in its biosynthesis and repair.

Structure of deoxyribonucleic acid

DNA is an antiparallel dimer of nucleic acid strands

DNA is composed of nucleotides containing the sugar deoxyribose. Deoxyribose is missing the hydroxyl group at the 2′-position. The chains of DNA are polymerized through a phosphodiester linkage from the 3′-hydroxyl of one ribose to the 5′-hydroxyl of the next ribose (Fig. 32.1A). Thus, DNA is a duplex linear polymer of deoxyribose 3′,5′-phosphate, with purine and pyrimidine bases attached to carbon-1′ of the deoxyribose subunit.

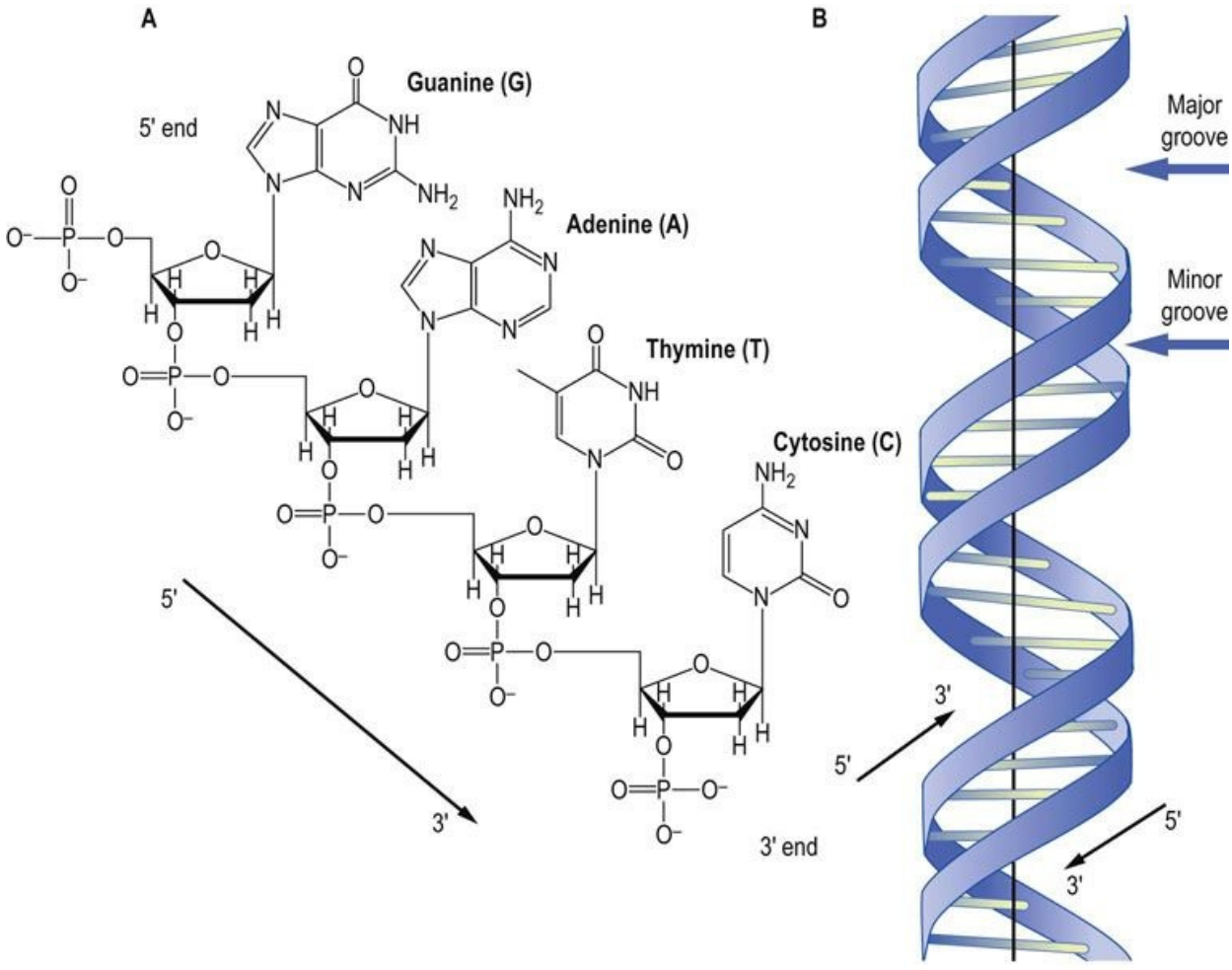


FIG. 32.1 Structure of DNA.
(A) A tetranucleotide sequence of DNA showing each of the nucleotides normally found

in DNA. The deoxyribose sugars are missing the 2′-hydroxyl that is present in the ribose sugars found in RNA. By convention, DNA is read from the 5′ to 3′ end, so the sequence of this tetranucleotide is 5′-GATC-3′. (B) A graphic representation of the structure of B-DNA, the major form of DNA in the cell. The base pairs in the middle are aligned nearly perpendicular to the helical axis. The major groove and the minor groove are shown. Note that the strands are antiparallel.

Using X-ray diffraction photographs of DNA taken by Rosalind Franklin, James Watson and Francis Crick proposed a structure for DNA in 1953. This model proposed that DNA was composed of **two intertwined complementary strands** with hydrogen bonds holding the strands together (Fig. 32.1B). The basic simplicity of this structure was consistent with the observation that in all DNA the molar content of A is equal to that of T, and the molar content of G is equal to that of C. While some of the details of the model have been modified, the Watson–Crick hypothesis was rapidly accepted and its essential elements have remained unchanged since originally proposed.

Watson and Crick model of DNA

As originally presented by Watson and Crick, DNA is composed of two strands, wound around each other in a right-handed, helical structure with the base pairs in the middle and the deoxyribosyl phosphate chains on the outside. The orientation of the DNA strands is **antiparallel**, *i.e.* the strands run in opposite directions. The nucleotide bases on each strand interact with the nucleotide bases on the other strand to form base pairs (Fig. 32.2). The base pairs are planar and are oriented nearly perpendicular to the axis of the helix. Each base pair is formed by hydrogen bonding between a purine and a pyrimidine. Consistent with this pairing, Guanine forms three hydrogen bonds with cytosine, and adenine forms two with thymine. Because of the specificity of this interaction between purines and pyrimidines on the opposite strands, the opposing strands of DNA are said to have complementary structures. The composite strength of the numerous hydrogen bonds formed between the bases of the opposite strands and the hydrophobic interactions among the bases is responsible for the extreme stability of the DNA double helix. While the hydrogen bonds between strands are affected by temperature and ionic strength, stable complementary structures can be formed at room temperature with as few as 6–8 nucleotides.

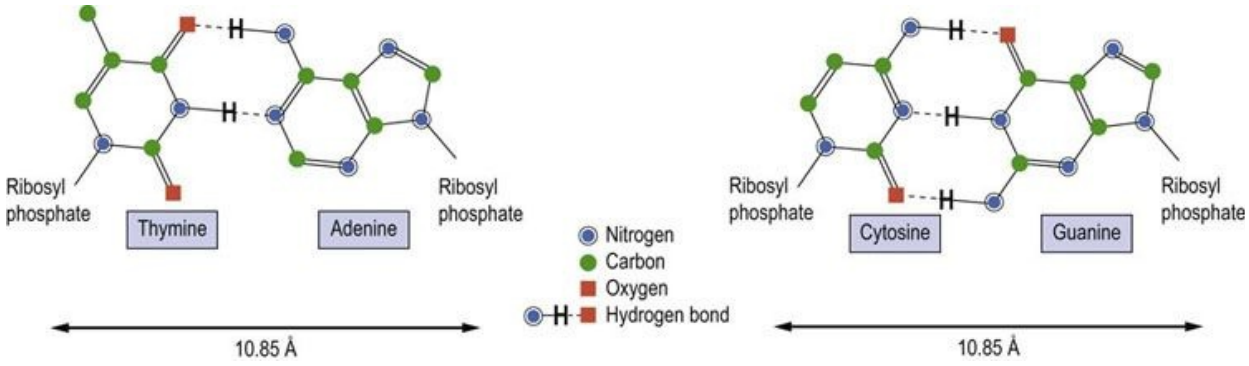


FIG. 32.2 Watson–Crick base pairing of nucleotides in DNA.
The AT base pairs form two hydrogen bonds and the GC base pairs form three hydrogen bonds. Thus, GC-rich regions are more stable than AT-rich regions.

Three-dimensional DNA

The three-dimensional structure of the DNA **double helix** is such that the deoxyribosyl phosphate backbones of the two strands are slightly offset from the center of the helix. Because of this, the grooves between the two strands are of different sizes. These grooves are termed the major groove and the minor groove (see Figs 32.1B). The major groove is more open and exposes the nucleotide base pairs. The minor groove is more constricted, being partially blocked by the deoxyribosyl moieties linking the base pairs. Binding of proteins to DNA occurs mostly in the major groove and is specific for the nucleotide sequence of DNA.

Alternative forms of DNA may help to regulate gene expression

Although the majority of DNA molecules in a cell exist in the B-form described above, alternative forms of DNA also exist. When the relative humidity of B-form DNA falls to less than 75%, the B-form undergoes a reversible transition into the A-form of DNA. In the A-form, the nucleotide base pairs are tilted 20° relative to the helical axis and the helix diameter is increased, compared to the B-form (Fig. 32.3). The A-form is observed when the DNA strands contain tracks of polypurine (and complementary polypyrimidine) residues. These regions do not efficiently bind histones and are therefore unable to form nucleosomes (see below), resulting in nucleosome-free (exposed) regions of DNA.

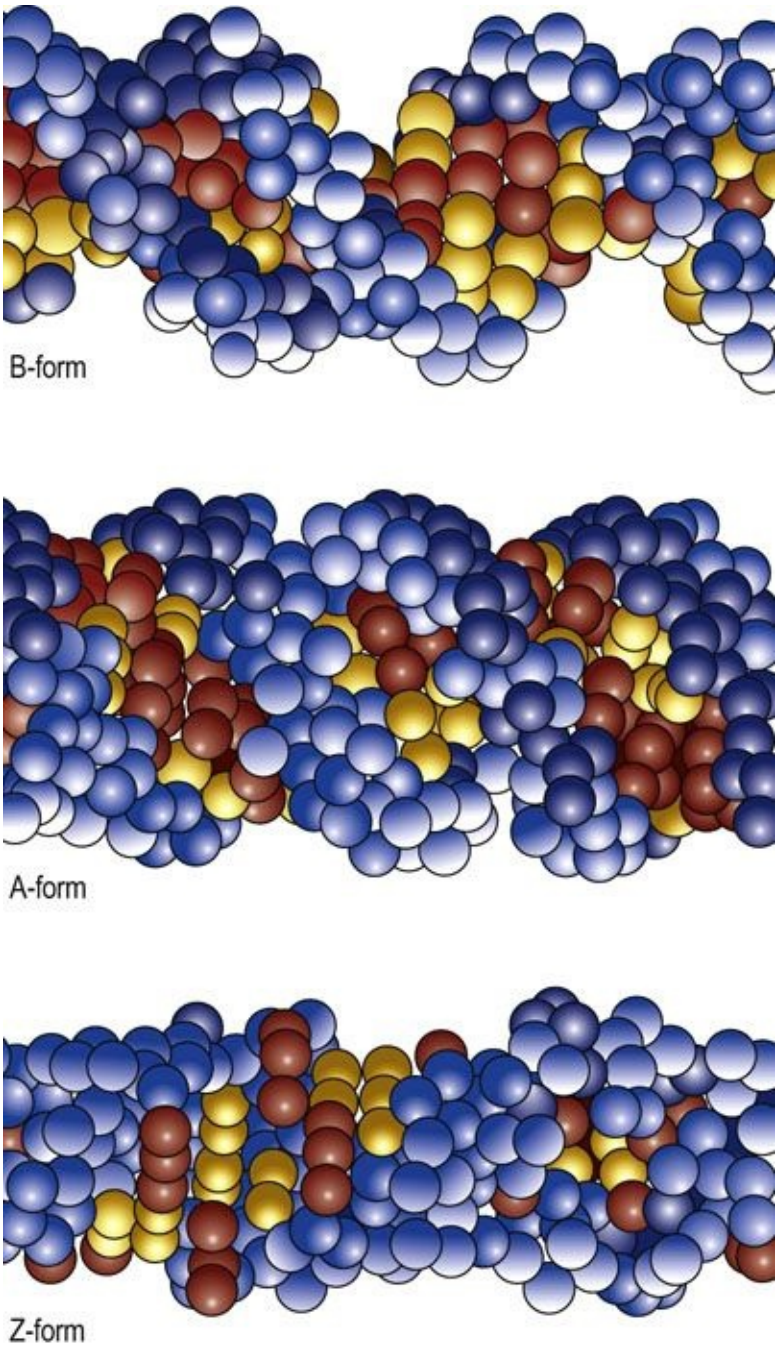


FIG. 32.3 The structures of different forms of DNA include the B-, A-and Z-forms. The sugar phosphate backbone of the DNA strands is colored blue. The nucleotide bases forming the internal base pairs are yellow for pyrimidines (thymine and cytosine) and red for purines (adenine and guanine).

Another unique form of DNA exists when the sequence of nucleotides consists of alternating purine/pyrimidine stretches. This form, termed Z-DNA, is also favored at high ionic concentrations. In Z-DNA, the base pairs flip 180°

relative to the sugar nucleotide bond. This results in a novel conformation of the base pairs relative to sugar phosphate backbones, yielding a form of DNA with a zigzag configuration (hence the name Z-DNA) along the sugar phosphate backbone. Surprisingly, this change in conformation leads to the formation of a left-handed DNA helix. While the Z-DNA form is favored at high ionic concentrations, it can also be induced at normal ionic concentrations by methylation of cytosine residues, a form of **epigenetic modification** of DNA (Chapter 35). Protein-binding interactions with these alternative forms of DNA, which are widely distributed in the genome, are involved in the regulation of gene expression.

The digital linear code (base pairing) in the DNA double helix has a significant component that acts by altering, along its length, the shape and stiffness of the molecule. In this way, one region of DNA is structurally differentiated from another, which provides another level of encoded information in three-dimensional space. These local shape and stiffness variations contribute superhelical structure and three-dimensional spatial interactions in DNA. It can be said that the superhelical density behaves as an analog regulatory mode as opposed to the more frequently accepted purely digital information content.

Separated DNA strands can reassociate to form duplex DNA

Complementary strands of DNA spontaneously hybridize to form helical structures

Because the DNA strands are complementary and are held together only by noncovalent forces, they can be separated into individual strands. This strand separation or denaturation of DNA is commonly induced by heating the solution. The dissociation is reversible and, on cooling, the complementary nucleotide sequences reassociate or reanneal to reform their original base pairs. This is the basis for one of the primary methods for DNA analysis, **Southern hybridization** (see Chapter 36). Because adenine and thymine interact through two hydrogen bonds and guanine and cytosine through three (see Fig. 32.2), AT-rich regions melt at lower temperatures than GC-rich regions in DNA. The denaturation of DNA can also be induced locally by enzymes or DNA-binding proteins. The promoter region of DNA contains a TATA sequence (the TATA box; see Fig.

35.1), an easily melted region of DNA that facilitates the unwinding of DNA during the early stages of gene expression (Chapter 35).

The human genome

The human genome contains 20,000–25,000 different protein coding genes spread over 23 chromosome pairs

Genes for specific proteins are unique DNA sequences that are present in single copies or at most only a few copies per genome. There are also several types of repeated DNA sequences within the genome. These are divided into two major classes: middle repetitive (<10 copies per genome) and highly repetitive (>10 copies per genome).

Some middle repetitive DNA consists of genes that specify transfer and ribosomal ribonucleic acids, which are involved in protein synthesis (Chapter 34), and histone proteins that are part of the nucleosome (below). Other middle repetitive DNA sequences have no known useful function but may participate in DNA strand association and chromosomal rearrangements during meiosis. The best-characterized highly repetitive sequence in humans is known as the Alu sequence. Between 300,000 and 500,000 Alu I repeats of about 300 base pairs are scattered throughout the human genome, comprising 3–6% of the total DNA. Individual repeats of the Alu sequence may vary by 10–20% in identity. Similar sequences are found in other mammals and in lower eukaryotes.

Satellite DNA

Satellite DNA was originally identified as a subfraction of DNA with a buoyant density slightly lower than that of genomic DNA because of its higher content of AT base pairs. It consists of clusters of short, species-specific, nearly identical sequences that are tandemly repeated hundreds of thousands of times. These clusters lack protein-coding genes and are found principally near the centromeres of chromosomes, suggesting that they may function to align the chromosomes during cell division to facilitate recombination. Because these repetitive sequences cover long stretches of chromosomes (100s to 1000s of kilobase pairs; kbp), determining the sequence of satellite DNA and sequencing the centromere region of DNA are major challenges to completing the non-coding sequence of eukaryotic genomes.

Mitochondrial DNA

The nucleus of eukaryotic cells contains the majority of the DNA in the cell – genomic DNA. However, DNA is also found in mitochondria and in plant chloroplasts, which is consistent with **endosymbiont theories** for the origins of these cellular organelles: namely, that they are parasites that adapted to intracellular life by symbiosis.

The mitochondrial genome is small in size, circular, and encodes relatively few proteins

In humans, the mitochondrial genome encodes 22 tRNAs, two rRNAs, and 13 mitochondrial proteins that are involved in the respiratory apparatus, including subunits of NADH dehydrogenase, cytochrome *b*, cytochrome oxidase and ATPase.

The remainder of the proteins that are found in mitochondria (about 1000) are produced from nuclear genes, synthesized in the cytoplasm on 'free' ribosomes (Chapter 34), then imported into the mitochondrion. This import process requires a special *N*-terminal **mitochondrial-import sequence** of about 25 amino acids in length that forms an amphipathic helix which interacts with transporter and chaperone proteins in the inner and outer mitochondrial membrane and matrix. Those few proteins that are encoded by the mitochondrial genome are synthesized in the mitochondrion, using machinery similar that used in the cytoplasm for synthesis of nonmitochondrial proteins (see Chapter 34).

DNA is compacted into chromosomes

Chromosomes are compact, highly organized forms of DNA

In eukaryotes, DNA is arranged in linear segments termed chromosomes. Each chromosome contains between 48 million and 240 million base pairs. The B-form of DNA has a contour length of 3.4 Å per base pair. Therefore, chromosomes have contour lengths of 1.6–8.2 cm, which is much larger than a cell. To fit within the nucleus, DNA is condensed >8000-fold into an organized structure. Interactions between DNA and mobile cations, such as Na^+, Mg^{2+}, and the polyamines such as spermidine, play an important role in the physical properties and biological function of DNA. Even in dilute solutions, approximately three out of four DNA charges are neutralized by a cation that is in some sense 'bound'. This neutralization facilitates compaction of DNA into

densely packaged chromatin and deformation of DNA by proteins.

Chromatin contains DNA, RNA and protein, plus inorganic and organic counterions

In the native chromosome, DNA is complexed with RNA and an approximately equal mass of protein. These DNA–RNA–protein complexes are termed chromatin. The majority of the proteins in chromatin are histones. Histones are a highly conserved family of proteins that are involved in the packing and folding of DNA within the nucleus. There are five classes of histones, termed H1, H2A, H2B, H3, and H4. They are all rich (>20%) in positively charged, basic amino acids (lysine and arginine). These positive charges interact with the negatively charged, acidic phosphate groups of the DNA strands to reduce electrostatic repulsion and permit tighter DNA packing.

Nucleosomes are the building-blocks of chromatin

The histone proteins associate into a complex termed a nucleosome (Fig. 32.4). Each of these complexes contains two molecules each of H2A, H2B, H3 and H4, and one molecule of H1. The nucleosome protein complex is encircled with about 200 base pairs of DNA that form two coils around the nucleosome core. The H1 protein associates with the outside of the nucleosome core to stabilize the complex. By forming nucleosomes, the packing density of DNA is increased about sevenfold.

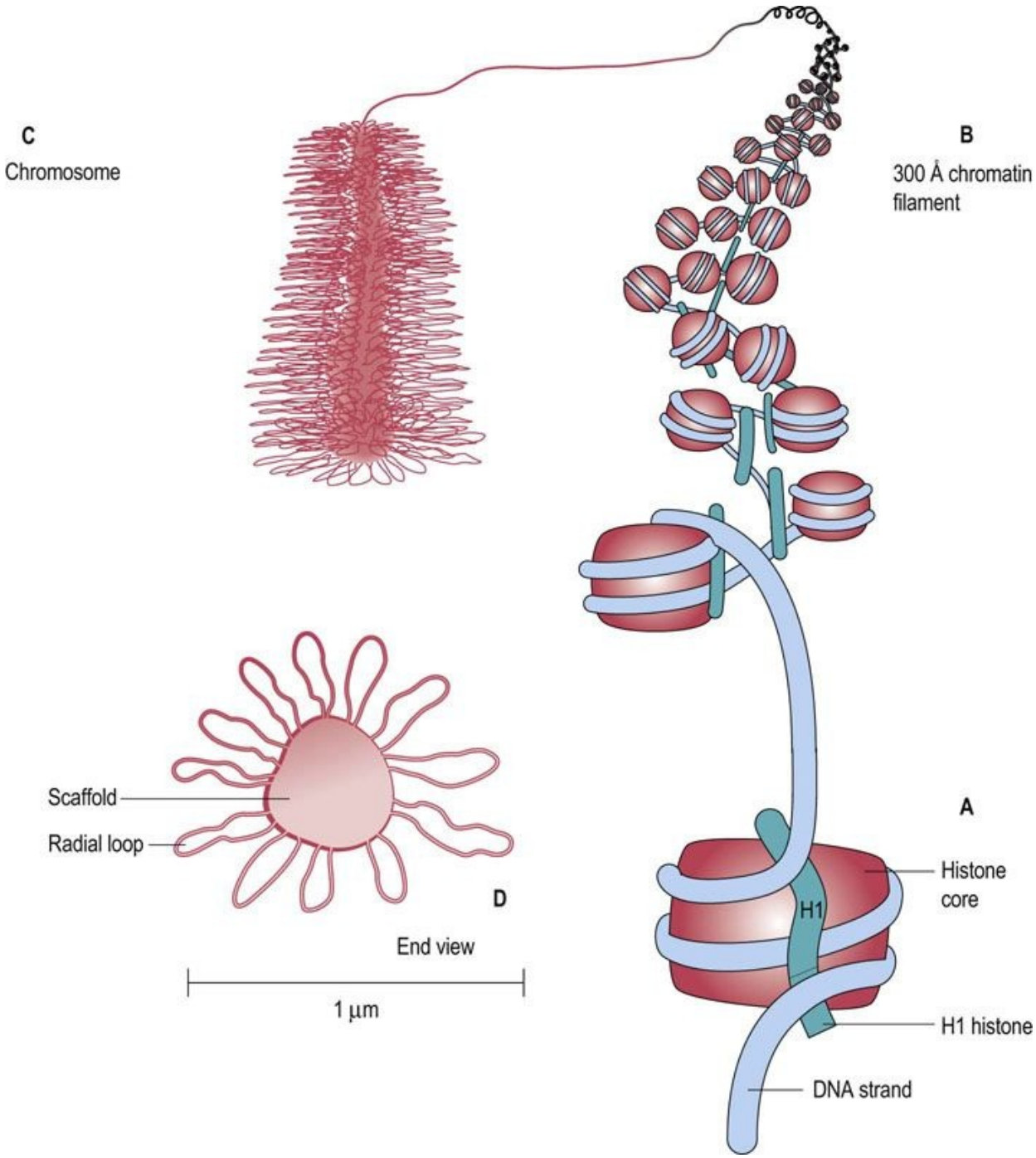


FIG. 32.4 Structures involved in chromosome packaging.
(A) The nucleosome core is composed of two subunits each of H2A, H2B, H3, and H4. The core is twice wrapped with DNA, and the H1 histone binds to the completed complex. (B) The 300 Å chromatin filament is formed by wrapping the nucleosomes into a spring-shaped solenoid. (C) The chromosome is composed of the 300 Å filaments, which bind to a nuclear scaffold, forming large loops of chromatin material. (D) The end view of a chromosome shows the central nuclear scaffold surrounded by the radial loops of chromatin. The diameter of a chromosome is about 1 µm.

The nucleosome particles themselves are also organized into other, more tightly packed structures, termed 300 Å chromatin filaments. These filaments are constructed by winding the nucleosome particles into a spring-shaped solenoid structure with about six nucleosomes per turn (see Fig. 32.4). The solenoid is stabilized by head-to-tail associations of the H1 histones. Finally, the chromatin filaments are compacted into the mature chromosome, using a nuclear scaffold. The scaffold is about 400 nm in diameter and forms the core of the chromosome. The filaments are dispersed around the scaffold to form radial loops about 300 nm in length. The final diameter of a chromosome is about 1 μm.

Telomeres

Telomeres are nucleoprotein complexes that cap the 3′ ends of the eukaryotic chromosomes. They are essential for cell viability. These structures consist of tandem repeats of short, G-rich, species-specific oligonucleotides. In humans, the repeated sequence is TTAGGG. Telomeres can contain as many as 1000 copies of this sequence. During the synthesis of telomeres, the enzyme **telomerase**, a ribonucleoprotein complex, adds preformed hexanucleotide repeats to the 3′-end of the chromosome, using its RNA as a template; there is no requirement for a DNA template. In human somatic cells telomeric DNA shortens in each cell division until it cannot exercise its end-protective functions, *e.g.* avoiding recognition of chromosome tips as double-stranded breaks. The shortening of telomeres after many cell replications has been linked to the development of cellular senescence. If telomeres turn out to be dysfunctional as a result of disproportionate shortening or defects in their intrinsic proteins, they trigger pathways that restrict proliferative life span. Telomere-based chromosome instability has been proposed as one driving force in oncogenesis.

The cell cycle in eukaryotes

Figure 32.5 shows the various phases of the growth and division of eukaryotic cells, known as the cell cycle. The G_1 phase is a period of cell growth that occurs prior to DNA replication. The phase during which DNA is synthesized or replicated is termed the S phase. A second growth phase, termed G_2, occurs after DNA replication but prior to cell division. The mitosis or M phase is the period of cell division. Following mitosis, the daughter cells either reenter the G_1 phase or enter a quiescent phase termed G_0, when growth and replication cease. The passage of cells through the cell cycle is tightly controlled by a variety of proteins termed cyclin-dependent kinases (see Chapter 42).

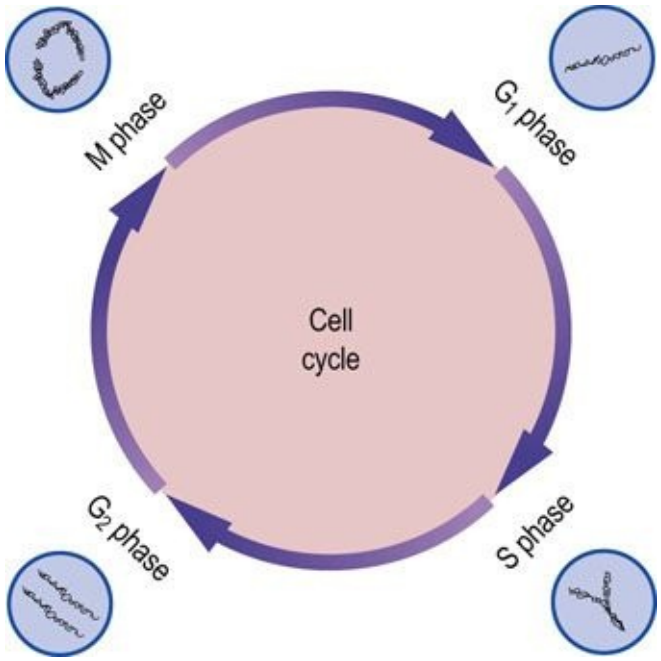


FIG. 32.5 Stages of the cell cycle.
G_1 and G_2 are growth phases that occur before and after DNA synthesis, respectively. DNA replication occurs during the S phase. Mitosis occurs during the M phase, producing new daughter cells that can reenter the G_1 phase (compare Fig. 43.1).

DNA Replication

DNA is replicated by separating and copying the strands

For cells to divide, their DNA must be duplicated during the S phase of the cell cycle. The structure of the DNA double helix and its complementarity suggested the mechanism for DNA replication – strand separation followed by strand copying. The separated parent strands serve as templates for the synthesis of the new daughter strands. This method of DNA replication is described as **semi-conservative** – each replicated duplex, daughter DNA molecule contains one parental strand and one newly synthesized strand.

DNA replication

The site at which DNA replication is initiated is termed the origin of replication

In prokaryotes, a DNA-binding protein termed DnaA binds to repeated nucleotide sequences located within the origin. Binding of 20–30 DnaA molecules to the origin of replication induces unwinding, which separates the strands in an AT-rich region adjacent to the DnaA-binding sites. Next, the hexameric protein DnaB binds to the separated DNA strands. DnaB has **helicase** activity that catalyzes ATP-mediated unwinding of the DNA helix. DNA **gyrase** also participates in separation of the strands. As this complex continues unwinding the DNA strands in both directions from the origin of replication, single-stranded DNA-binding proteins coat the separated strands to inhibit their reassociation.

Once the strands are sufficiently separated, another protein, termed DNA **primase**, is added, resulting in the formation of a **primosome complex** at the replication fork. The primosome synthesizes short ($n \leq 10$) RNA oligonucleotides complementary to each parental DNA strand. These oligonucleotides serve as primers for DNA synthesis. Once each RNA primer has been laid down, two **DNA polymerase III** complexes are assembled, one at each of the primed sites. Because of the unidirectional synthetic activity of the polymerase and the antiparallel nature of the two strands, the synthesis of DNA along the two strands is different (Fig. 32.6). The two daughter strands being

synthesized are termed the leading strand and the lagging strand. In addition to its polymerase activity, one of the subunits of DNA polymerase III has a **proofreading** exonuclease activity, which corrects mismatches and assures fidelity in replication of DNA.

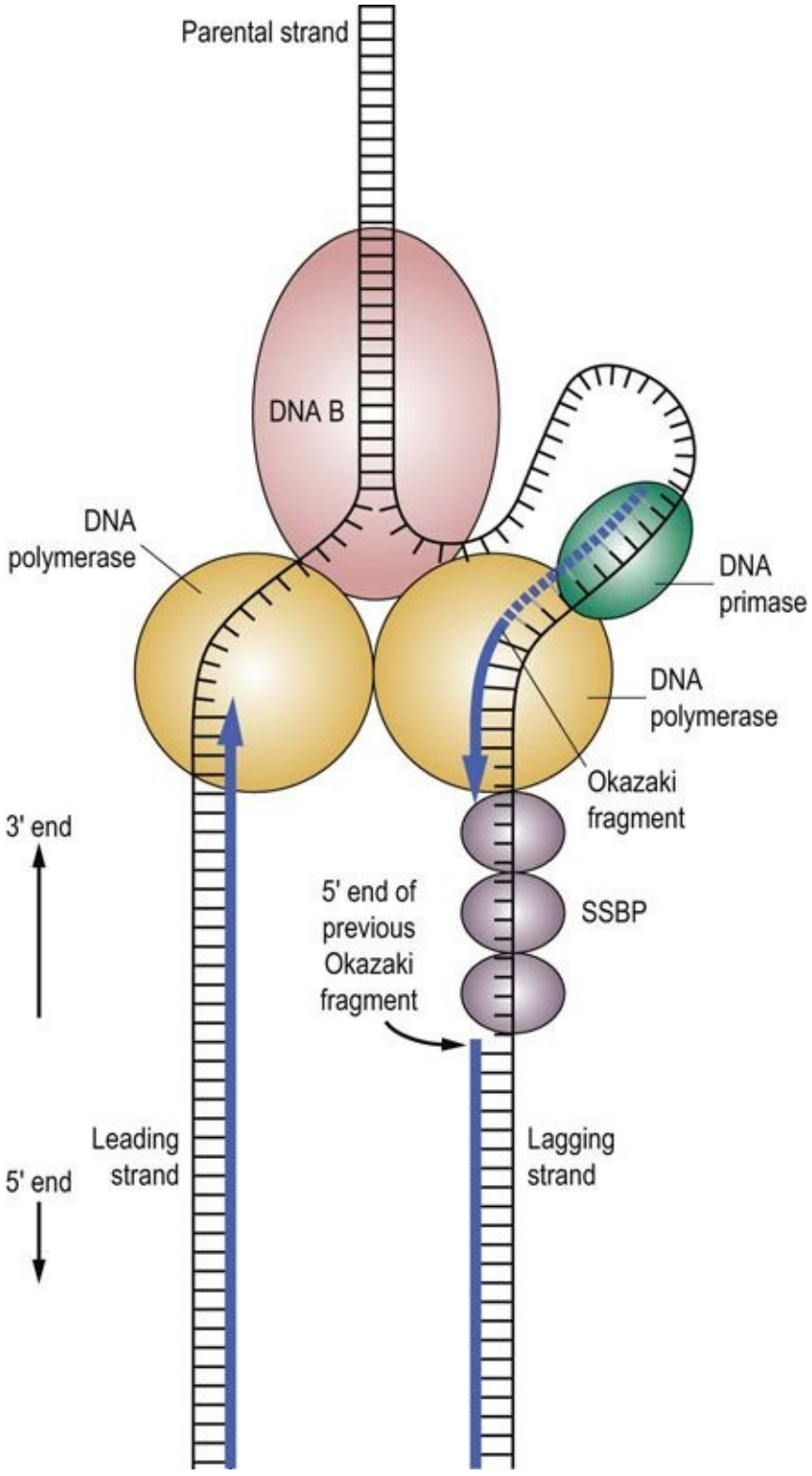


FIG. 32.6 DNA synthesis.
DNA synthesis occurs at a replication fork producing new strands termed the leading strand and the lagging strand. The 'railroad tracks' represent double-stranded DNA.

Some of the enzymes involved in DNA synthesis are shown: DNA B (helicase), DNA primase, DNA polymerase, and single-strand DNA binding protein (SSBP). The leading strand is replicated in a continuous fashion. However, for the lagging strand, RNA primers are periodically added by the DNA primase along the strand. DNA polymerase III elongates these RNA primers to form Okazaki fragments. When the Okazaki fragment is complete, the DNA polymerase III on the lagging strand will shift to the next RNA primer to initiate another Okazaki fragment. The exonuclease activity of DNA polymerase I removes the RNA primers and replaces them with DNA. DNA ligase seals the gaps in the DNA strands to complete synthesis of the lagging strand.

DNA synthesis proceeds in opposite directions along the leading and lagging strands of the template DNA

DNA synthesis proceeds along the leading strand in a 5′ to 3′ direction, producing a single, long, continuous strand. However, because DNA synthesis adds new nucleotides only at the 3′-end of the elongating DNA strand, DNA polymerase III cannot synthesize the lagging strand in one long continuous piece as it does for the leading strand. Instead, the lagging strand is synthesized in small fragments, 1000–5000 base pairs in length, termed **Okazaki fragments** (see Fig. 32.6). The primosome remains associated with the lagging strand and continues periodically to synthesize RNA primers complementary to the separated strand. As DNA polymerase III moves along the parental DNA strand, it initiates the synthesis of Okazaki fragments at the RNA primers, elongating different fragments from each primer.

When the 3′-end of the elongating Okazaki fragment reaches the 5′-end of the previously synthesized Okazaki fragment, DNA polymerase III releases the template and finds another RNA primer further back along the lagging strand, synthesizing another Okazaki fragment. Eventually, the Okazaki fragments are joined by **DNA polymerase I**. This enzyme, which also has a role in DNA repair, has an exonuclease activity that permits it to remove and replace a stretch of nucleotides as it proceeds along a DNA template. During DNA replication, DNA polymerase I removes the RNA primer and replaces it with DNA. Finally, DNA **ligase** joins the lagging-strand DNA fragments to form a continuous strand.

Eukaryotes stringently regulate DNA replication

Eukaryotic DNA synthesis is remarkably similar to prokaryotic DNA synthesis. However, eukaryotes have many more origins of replication. These are activated simultaneously during the S phase of the cell cycle, permitting rapid replication

of the entire chromosome. To insure that excess amounts of unfinished, replicating DNA do not accumulate, cells use a protein termed a **licensing factor** that is present in the nucleus prior to replication. Following each round of replication, this factor is inactivated or destroyed, preventing further replication until more licensing factor is synthesized later in the cell cycle.

Licensing factor is best understood in the yeast. In yeast there is a complex called the origin recognition complex (ORC), composed of six proteins. The ORC marks the origin of replication and remains bound to the origin throughout the cell cycle. It serves as a docking site for the other components that regulate DNA replication. Early events in DNA replication include the binding of several highly unstable proteins to ORC. These proteins include CDC6/18 and CDT1, which facilitate binding of a group of three additional proteins: MCM2, MCM3, and MCM5. Once the MCM proteins bind, the origin exists as a prereplication complex and is 'licensed' to enter S phase of the cell cycle. The initiation of DNA synthesis is triggered by the action of the CDC7 kinase together with other **cyclin-dependent kinases**. The activated MCM complex then participates in unwinding the replication origin and is thereby displaced from the origin. Following displacement, the origin forms a postinitiation complex and CDC6 is degraded, thereby preventing the reloading of the origin with additional licensing factor.

DNA repair

There are typically more than 10,000 modifications of DNA per cell per day

Because DNA is the reservoir of genetic information within the cell, it is extremely important to maintain the integrity of DNA. Therefore, the cell has developed multiple, highly efficient mechanisms for the repair of modified or damaged DNA.

Clinical box AZT therapy for HIV infection

Human immunodeficiency virus (HIV) infection results in a profound weakening of the immune system that makes the patient susceptible to a range of bacterial, fungal, protozoal and viral superinfections. **Kaposi's sarcoma** may also develop; it is a cancer-like disease of blood vessels caused by infection with human herpesvirus-8 (HHV-8).

Effective treatments of the HIV viral infection rely on detailed knowledge of the viral life cycle. For the AIDS virus, the viral genome is RNA. In the infected cell, it is copied into a DNA form by a viral enzyme termed **reverse transcriptase**. Reverse transcriptase is an error-prone enzyme that does not have the proofreading capabilities of DNA polymerase III. One therapeutic approach for treatment of AIDS takes advantage of the enzyme's lack of specificity in choice of complementary substrates. Several important antiviral drugs are nucleotide analogues that inhibit reverse transcriptase, including AZT (azido-2′,3′-dideoxythymidine; Retrovir, zidovudine), ddC (2′,3′-dideoxycytidine; Hivid, zalcitabine), and 3TC (2′,3′-dideoxy-3′-thiacytidine; Epivir, lamivudine) (Fig. 32.7). AZT, for example, is metabolized in the body into the thymine triphosphate (TTP) analogue azido-TTP. The HIV reverse transcriptase misincorporates azido-TTP into the reverse-transcribed viral genome. The incorporation of azido-TTP into DNA blocks further

chain elongation, because the 3′-azido group cannot form a phosphodiester bond with subsequent nucleoside triphosphates. The inability to synthesize DNA from the viral RNA template results in inhibition of viral replication The life cycle of HIV spans about 1.5 days from entry into a cell, replication, assembly, and release of new viral particles to infection of other cells. HIV lacks *proofreading* enzymes to correct errors occurring during conversion of its RNA into DNA via reverse transcription. Its short life cycle and high error rate cause the virus to mutate very hastily, causing large genetic variability of HIV. Most of the mutations are deleterious or bear no advantage, but some have a natural selection advantage to their parent and can allow them to evade the human immune system and antiretroviral drugs. The more active copies of the virus, the greater the possibility that one resistant to antiretroviral drugs will be made.

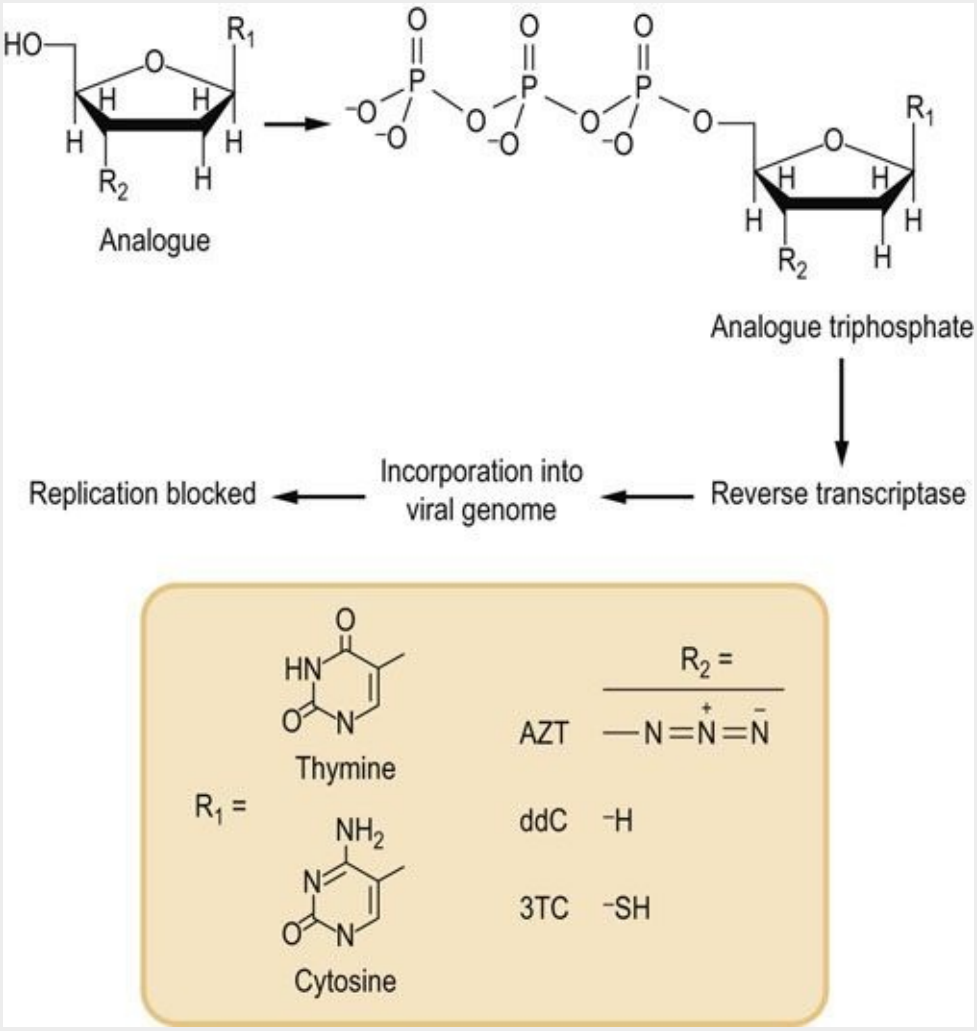


FIG. 32.7 Mechanism of action of antiretroviral chemotherapeutic agents.
This class of inhibitors includes several compounds with slightly different chemical structures in the nucleobase structure and in substitution at the 3′-carbon of the sugar ring. Structures of some of the most widely used drugs are shown. These compounds are metabolized to the triphosphate form via normal cellular metabolism (see Chapter 30). The triphosphate analogues are then incorporated into the viral genome by reverse transcriptase. This blocks viral DNA synthesis because the modified 3′ end R_2 of the viral DNA molecule is not a substrate for additional rounds of DNA synthesis. AZT, azido-2′,3′-dideoxythymidine; ddC, 2′,3′-dideoxycytidine; 3TC, 2′,3′-dideoxy-3′-thiacytidine.

If antiretroviral therapy is incorrectly employed, these multi-drug resistant strains can turn out to be the dominant genotypes very promptly. Improper serial use of the reverse transcriptase inhibitors such as zidovudine, didanosine, zalcitabine, stavudine,

and lamivudine may result in multi-drug resistant mutations.

DNA can be damaged by numerous types of endogenous and exogenous agents that cause nucleotide modifications, deletions, insertions, sequence inversions and transpositions. Some of this damage is secondary to chemical modification of DNA by alkylating agents (including many carcinogens), reactive oxygen species (Chapter 37) and ionizing radiation (ultraviolet or radioactive). Both the sugar and bases of DNA are subject to modification, yielding an estimated 10,000 to 100,000 modifications of DNA per cell per day. The nature of this damage is quite variable, including modification of single bases, single or double-strand breaks, and crosslinking between bases or bases and proteins. Oxidative damage is probably the most common form of DNA damage; it is increased in inflammation, by smoking, in aging and in age-related diseases, including atherosclerosis, diabetes and neurodegenerative diseases (Chapter 43). If not repaired, the accumulated damage will lead to permanent changes in the structure of DNA, setting the stage for loss of cellular functions, cell death or cancer.

Multiple enzymatic pathways repair a wide range of chemical modifications of DNA

Numerous chemical and environmental agents are known that produce specific chemical modification of the nucleotides in the DNA strand, leading to mismatches during DNA synthesis. After chromosomal replication, the resulting daughter strand contains a different DNA sequence (mutation) from the parent strand. Cells use excision repair to remove alkylated nucleotides and other unusual base analogues, thereby protecting the DNA sequence from mutations. The unmodified strand serves as the template for the repair process.

UV light produces thymine dimers: nucleotide excision repair

When short-wavelength ultraviolet (UV) light interacts with DNA, adjacent thymine bases undergo an unusual dimerization, producing a cyclobutylthymine dimer in the DNA strand (Fig. 32.8). The primary mechanism for repair of these

intrastrand thymine dimers is an excision repair mechanism. An endonuclease, which appears to be specific for this type of modification, cleaves the dimer-containing strand near the thymine dimer, and a small portion of that strand is removed. DNA polymerase I, the same enzyme that is involved in DNA biosynthesis, then recognizes and fills in the resulting gap. DNA ligase completes the repair by rejoining the DNA strands.

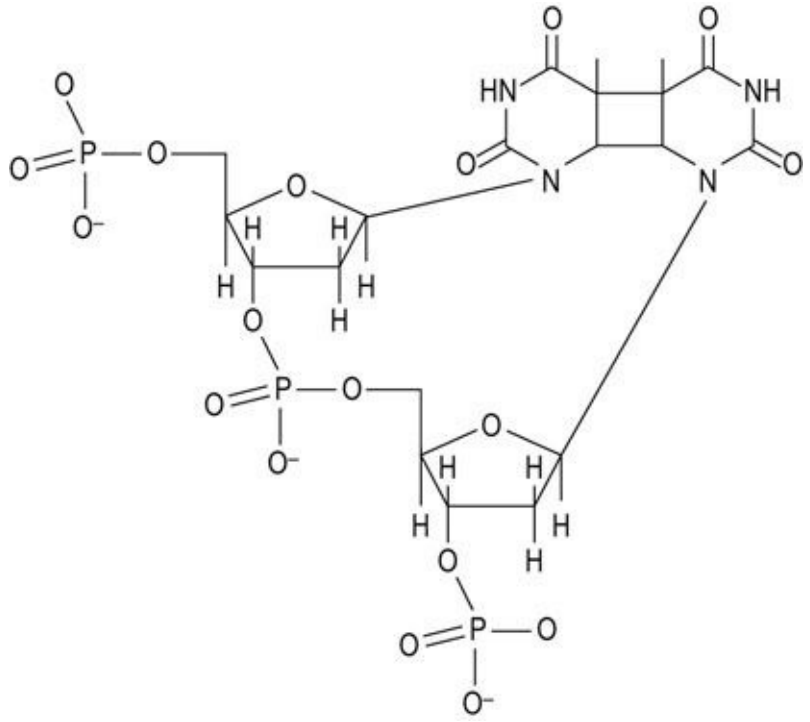


FIG. 32.8 Thymine dimer.
A thymine dimer consists of a cyclobutane ring joining a pair of adjacent thymine nucleotides.

Deamination: excision repair

Those nucleotides that contain amines, cytosine and adenosine, may spontaneously deaminate to form uracil or hypoxanthine, respectively. When these bases are found in DNA, specific **N-glycosylases** remove them. This produces base pair gaps that are recognized by specific apurinic or apyrimidinic endonucleases that cleave the DNA near the site of the defect. An exonuclease then removes the stretch of the DNA strand containing the defect. A repair DNA polymerase replaces the DNA and, finally, DNA ligase rejoins the DNA strand.

This repair mechanism is also referred to as excision repair.

Depurination

Single base pair alterations also include depurination. The purine-*N*-glycosidic bonds are especially labile, so that an estimated 3–7 purines are removed from DNA per min per cell. Specific enzymes recognize these depurinated sites, and the base is replaced without interruption of the phosphodiester backbone.

Strand breaks

Single-stranded breaks are frequently induced by ionizing radiation. These are repaired by direct ligation or by excision repair mechanisms. Double-stranded breaks are produced by ionizing radiation and some chemotherapeutic agents. Otherwise, double-stranded ends of DNA are rare in vivo; they are found at the end of chromosomes and in some specialized complexes involved in gene rearrangement. A specialized enzyme system is designed to recognize and rejoin these ends but if the ends drift away from one another, the damage is not readily repaired.

Clinical box Xeroderma pigmentosum

Xeroderma pigmentosum (XP) is a group of rare, life-threatening, autosomal recessive disorders (incidence = 1/250,000) that are marked by extreme sensitivity to sunlight. Upon exposure to sunlight or ultraviolet radiation, the skin of XP patients erupts into pigmented spots, resembling freckles. Multiple carcinomas and melanomas appear early in life, exacerbated by sun exposure, and the majority of patients succumb to cancer before reaching adulthood.

XP is the result of a defect in repair of UV-induced thymine dimers in DNA. There are at least eight polypeptides (genes) involved in recognition, unwinding and excision repair of UV-induced thymine dimers. Patients with XP must avoid direct sunlight, fluorescent light, halogen light or any other source of ultraviolet light. An experimental form of protein therapy, currently undergoing clinical evaluation, involves application of a skin lotion containing the missing protein or enzyme. Ideally, this

protein will enter the skin cells and stimulate the repair of UV-damaged DNA. However, protection occurs only where the lotion can be applied. For example, this treatment does not address the neurologic problems that affect about 20% of XP patients.

Mismatch repair

Errors that escape the proofreading activity of DNA polymerase III appear in newly synthesized DNA in the form of nucleotide mismatches. While readily repairable, the critical issue is identification of the strand to be repaired: which nucleotide strand is the daughter strand containing the error? In bacterial systems, mismatch repair is accomplished by post-replicative methylation of DNA at adenine residues in specific sequences spaced along the genome; methylation does not affect base pairing. Newly synthesized strands lack methylated adenine residues, so that the mismatch repair system enzymes scan the DNA, identify the mismatch, and then repair the unmethylated strand by excision repair. A similar approach is used to correct mismatches occurring during synthesis of mammalian DNA. Defects in mismatch repair are associated with hereditary nonpolyposis colon cancer, an autosomal dominant condition in humans.

8-Oxo-2′-deoxyguanosine

More than 20 different oxidative modifications of DNA have been characterized; the most studied is 8-oxo-2′-deoxyguanosine (8-oxoG) (Fig. 32.9). During the process of DNA replication, mismatches between the modified 8-oxoG nucleoside in the template strand and incoming nucleotide triphosphates result in G-to-T transversions, thereby introducing mutations into the DNA strand. Although excision repair mechanisms are effective, 8-oxoG, like other modified bases, may be reincorporated into DNA following excision.

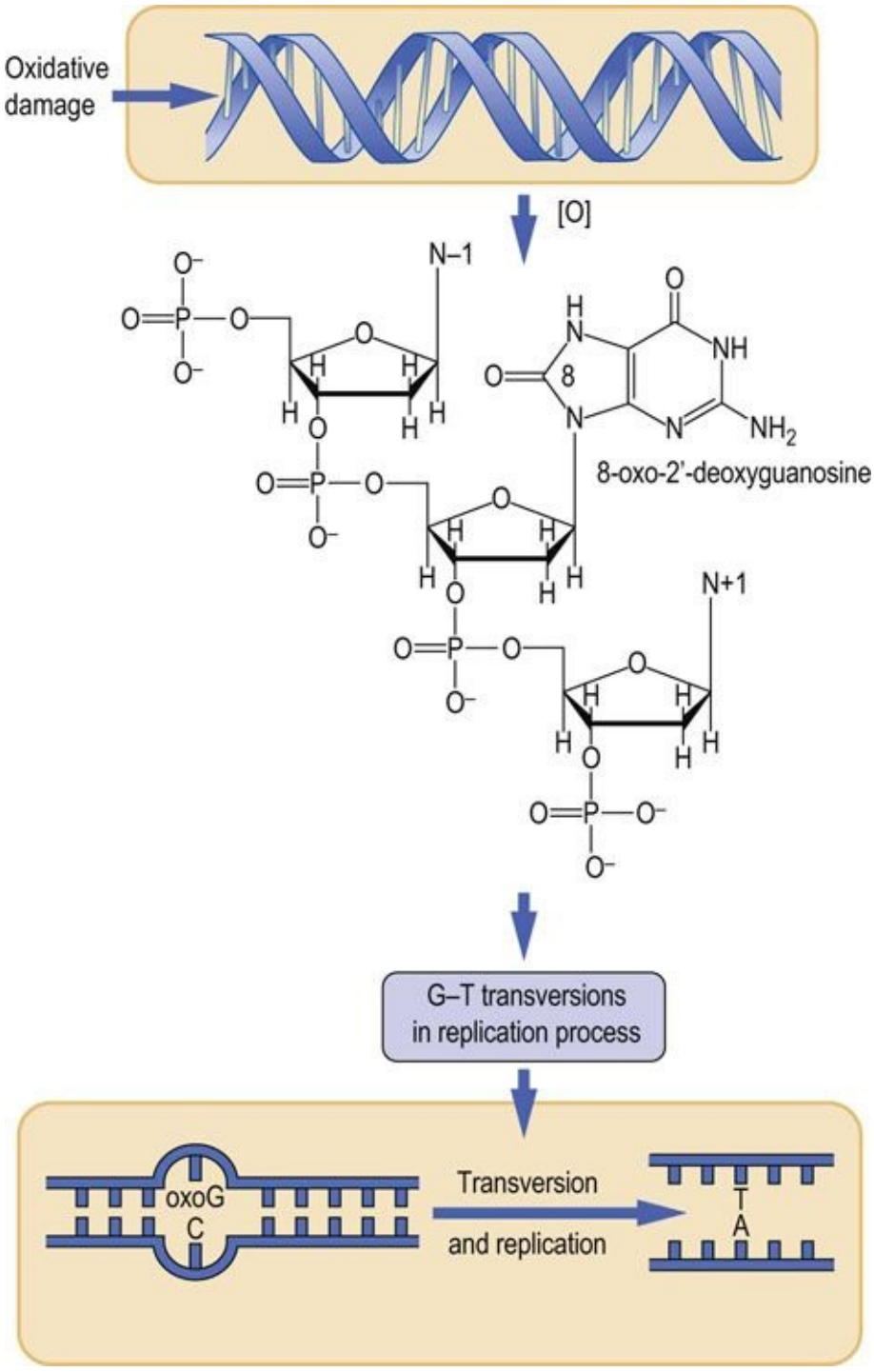


FIG. 32.9 Oxidative damage to DNA.
8-Oxo-2′-deoxyguanosine (oxoG) is an oxidative modification of DNA that causes mutations during replication of DNA. Replication of the strand containing oxoG frequently yields a pyrimidine A in the complementary strand, which, on further replication, yields an AT base pair, instead of the original GC base pair.

Recently, a mammalian protein, MTH1, was characterized that specifically degrades 8-oxo-dGTP, thereby preventing misincorporation of this altered nucleotide into DNA. Gene targeting was used to develop an MTH1 knockout mouse. Compared to the wild-type animal, the knockout showed a greater number of tumors in lung, liver, and stomach, illustrating the importance of this (and other) postrepair protection mechanisms.

In lung cells, inhalation of some particulate materials results in an increase in 8-oxoG levels. The inflammatory process may play a role in asbestos-induced formation of lung tumors. Smoking also induces oxidative damage and increases levels of DNA oxidation products in lungs, blood and urine. 8-Oxo-2′-deoxyguanosine is eliminated via renal filtration. Therefore, its urinary level is used as a sensitive biomarker for oxidative stress in many clinical studies (Chapter 43). A new competitive immunochromatography automatic analyzer that measures urinary 8-oxodG can be used as a point-of-care test for the assessment of oxidative stress.

Clinical test box Ames test for mutagens

Mutagens are chemical compounds that induce changes in the DNA sequence. A large number of natural and man-made chemicals are mutagenic. To evaluate the potential to mutate DNA, the American biochemist Bruce Ames developed a simple test, using special *Salmonella typhimurium* strains that cannot grow in the absence of histidine (His^- phenotype). These histidine auxotrophic strains contain nucleotide substitutions or deletions that prevent the production of histidine biosynthetic enzymes.

To test for mutagenesis, mutant bacteria are seeded on a culture medium lacking histidine; the suspected mutagen is added to the medium. The action of the mutagen occasionally results in the reversal of the histidine mutation, yielding a revertant strain that can now synthesize histidine and will grow in its absence. The mutagenicity of a compound is scored by counting the number of colonies that have grown, *i.e.* reverted to the His^+ phenotype. There is a good correlation between results of the Ames mutagenicity test and direct tests of carcinogenic activity in animals.

Some chemicals (**procarcinogens**) are not mutagenic per se, but

are activated to mutagenic compounds during metabolic processes, *e.g.* during drug detoxification in liver or kidney. Benzopyrene, for example, is not mutagenic but during its detoxification in liver, it is converted to diol epoxides which are potent mutagens and carcinogens. To provide sensitivity for detecting procarcinogens, the culture medium for the Ames test is supplemented with an extract of liver microsomes, a subfraction of tissue rich in smooth endoplasmic reticulum containing drug-metabolizing enzymes.

Clinical box Cancer treatment news: two new protein targets to counteract ‘relapse’ and ‘drug resistance’

Current cancer treatments such as ionizing radiation and chemotherapy target DNA. Their rationale is clear: these treatments disrupt the genome, and a balance is struck between preventing cancer cells from dividing and proliferating, while not irreversibly damaging healthy tissues. Nonetheless, cancer cells have a broad array of DNA repair mechanisms to limit injury. For this reason, DNA repair systems are targets of adjuvant therapy used to enhance sensitivity of cancer cells to DNA-targeted agents. Base excision repair and nucleotide excision repair are key mechanisms of DNA repair. There are two protein targets associated with the hallmark ‘relapse’ and ‘drug resistance’ phenomena seen during chemotherapy: Excision Repair Cross-Complementation Group 1 (ERCC1) and DNA polymerase beta. The former is a key player in nucleotide excision repair; the latter is the error-prone polymerase of base excision repair. Only a few ERCC1 inhibitors have been discovered, but more than 60 for DNA pol beta. The discovery of potent and tumor-specific inhibitors of these enzymes should improve current therapies where resistance develops, including bleomycin, alkylating agents

and cisplatin.

Summary

■ The human genome is composed of DNA, an antiparallel, double-stranded helical polymer of deoxyribonucleotides, stabilized by hydrogen binding between complementary bases.
■ DNA is packaged in the chromosome in a highly organized, condensed structure known as chromatin.
■ Genetic information is replicated by a semi-conservative mechanism in which the parental strands are separated and both act as templates for daughter DNA.
■ The replication of DNA is a complex, stringently regulated process. DNA is essentially the only polymer in the body that is repaired, rather than degraded, following chemical or biological modification. Repair mechanisms generally involve excision of modified bases and replacement, using the unmodified strand as a template.

Active learning

1. What are the possible functions of unusual forms of DNA?
2. Discuss the possible roles for middle repetitive and highly repetitive sequences in DNA.
3. Hereditary nonpolyposis colon cancer results from a defect in DNA mismatch repair. Why is this condition autosomal dominant?
4. What are the differences in replication and repair of nuclear vs mitochondrial DNA?

Further reading

Frias, C, Pampalona, J, Genesca, A, *et al.* Telomere dysfunction and genome instability. *Front Biosci*. 2012; 17:2181–2196.

Kaneko, K, Kimata, T, Tsuji, S, *et al.* Measurement of urinary 8-oxo-7,8-dihydro-2-deoxyguanosine in a novel point-of-care testing device to assess oxidative stress in children. *Clin Chim Acta*. 2012; 413:1822–1826.

Ruemmele, FM, Garnier-Lengline, H. Why are genetics important for nutrition? Lessons from epigenetic research. *Ann Nutr Metab*. 2012; 60(Suppl 3):38–43.

Slijepcevic, P. DNA damage response, telomere maintenance and ageing in light of the integrative model. *Mech Ageing Dev*. 2008; 129:11–16.

Spry, M, Scott, T, Pierce, H, *et al*. DNA repair pathways and hereditary cancer susceptibility syndromes. *Front Biosci*. 2007; 12:4191–4207.

Travers, AA, Muskhelishvili, G, Thompson, JM. DNA information: from digital code to analogue structure. *Philos Transact A Math Phys Eng Sci*. 2012; 370:2960–2986.

Venter, JC. Genome sequencing anniversary: the human genome at 10 – successes and challenges. *Science*. 2011; 331:546–547.

Venter, JC, Adams, MD, Myers, EW. The sequence of the human genome. *Science*. 2001; 291:1304–1351.

Watson, JD. *The double helix: a personal account of the discovery of the structure of DNA*. New York: WW Norton; 1980.

Watson, JD. *A passion for DNA. Genes, genomes and society*. Oxford: Oxford University Press; 2000.

Websites

DNA interactive. www.dnai.org. [This is the best introduction to the subject].

DNA replication. http://www.freesciencelectures.com/video/dna-replication-process/.

DNA workshop. www.pbs.org/wgbh/aso/tryit/dna.

Graphics. www.accessexcellence.org/AB/GG/structure.html.

Mitochondrial DNA. www.mitomap.org.

National Human Genome Research Institut. www.genome.gov

CHAPTER 33

Ribonucleic Acid

Gary A. Bannon and Robert Thornburg

Learning objectives

After reading this chapter you should be able to:

- Identify the major types of cellular RNA and the function of each.
- Describe the major steps in transcription of an RNA molecule.
- Explain the function of the different RNA polymerase enzymes.
- Describe the major differences between prokaryotic and eukaryotic mRNAs.
- Describe the different processing and splicing events that occur during synthesis of eukaryotic mRNAs.

Introduction

Transcription is defined as the synthesis of a ribonucleic acid (RNA) molecule using deoxyribonucleic acid (DNA) as a template

Transcription is a series of complicated enzymatic processes that result in the transfer of the genetic information stored in double-stranded DNA into a single-stranded RNA molecule that will be used by the cell to direct the synthesis of its proteins, a process known as translation. There are three general classes of RNA molecules found in prokaryotic and eukaryotic cells: ribosomal RNA (rRNA), transfer RNA (tRNA), and messenger RNA (mRNA). Each class has a distinctive size and function (Table 33.1), described by its sedimentation rate in an ultracentrifuge (S, Svedberg units) or its number of bases (nt, nucleotides, or kb, kilobases). Prokaryotes have the same three general classes of RNA as eukaryotes, but sizes and structural features differ:

Table 33.1

General classes of RNA

RNA	Size and length	Percent of total cellular RNA	Function
rRNA	28 S, 18 S, 5.8 S, 5 S (23 S, 16 S, 5 S)*	80	Interact to form ribosomes
tRNA	65–110 nt	15	Adapter
mRNA	0.5–6+ kb	5	Directs synthesis of cellular proteins

*Size of rRNA in prokaryotic cells. nt, nucleotides; kb, kilobases; S, Svedberg units.

■ **ribosomal RNA (rRNA)** from prokaryotes consists of three different sizes of RNA, while rRNA from eukaryotes consists of four different sizes of RNA. These RNAs interact with each other, and with proteins, to form a ribosome that

provides the basic machinery on which protein synthesis takes place.

■ **transfer RNAs (tRNAs)** consist of a class of RNAs that are 65–110 nt in length; they function as amino acid carriers and as recognition molecules that identify the mRNA nucleotide sequence and translate that sequence into the amino acid sequence of proteins.

■ **messenger RNAs (mRNAs)** represent the most heterogeneous class of RNAs found in cells. mRNAs generally range in size from 500 nt to ~6 kb (some rare but important mRNAs are >100 kb). mRNAs are carriers of genetic information, defining the sequence of all proteins in the cell; they are the 'working copy' of the genome.

In order to understand the complex series of events that lead to the production of these three classes of RNA, this chapter is divided into five parts. The first part deals with the molecular anatomy of the major types of RNAs found in prokaryotic and eukaryotic cells; by understanding the chemical nature of the final products of transcription, you will be better prepared to understand the steps involved in generating these molecules. The second part describes the main enzymes involved in transcription, and their specificities. The third part describes the three steps (initiation, elongation, and termination) required to produce a protein. In the fourth section, the modifications that are made to the primary products of transcription (post-transcriptional processing) are described. This information is expanded in Chapter 34. The final section describes briefly how cells regulate gene expression at the RNA level.

Molecular anatomy of ribonucleic acid molecules

In contrast to DNA, RNAs are mostly single stranded, and contain uracil instead of thymine

In general, the RNAs produced by prokaryotic and eukaryotic cells are single-stranded nucleic acid molecules that consist of adenine, guanine, cytosine, and uracil nucleotides joined to one another by phosphodiester linkages. The start of an RNA molecule is known as its 5′ end, and the termination of the RNA is its 3′ end. Even though most RNAs are single-stranded, they fold back on themselves. Therefore, they exhibit extensive secondary structure, including intramolecular double-stranded regions that are important to their function. These secondary structures, one of the most common of which is called a **hairpin loop** (Fig. 33.1), are the product of intramolecular base pairing that occurs between complementary nucleotides within a single RNA molecule.

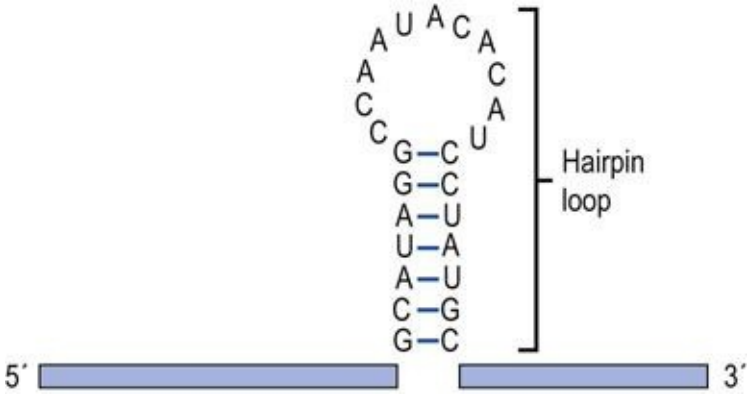


FIG. 33.1 RNA hairpin loop.
RNA can form secondary structures called hairpin loops. These structures form when complementary bases within an individual RNA share hydrogen bonds and form base pairs. Hairpin loops are known to be important in the regulation of transcription in both eukaryotic and prokaryotic cells.

rRNAs: the ribosomal RNAs

The eukaryotic rRNAs are initially synthesized as a single RNA transcript with a size of 45 S and about 13 kb long. This large primary transcript is processed into 28 S, 18 S, 5.8 S, and 5 S rRNAs (~3 kb, 1.5 kb, 160 and 120 nt, respectively). The 28 S, 5.8 S, and 5 S rRNAs associate with ribosomal proteins to form the large ribosomal subunit. The 18 S rRNA associates with different proteins to form the small ribosomal subunit. The large ribosomal subunit with its proteins and RNA has a characteristic size of 60 S; the small ribosomal subunit has a size of 40 S. These two subunits interact to form a functional 80 S ribosome (see Chapter 34). Prokaryotic rRNAs interact in a similar fashion to form these ribosomal subunits but have a slightly smaller size, reflecting the difference in prokaryotic and eukaryotic rRNA transcript size (Table 33.2).

Table 33.2

rRNAs and ribosomes

Cell type	rRNA	Subunit	Size	Intact ribosome
Prokaryotic	23 S, 5 S 16 S	Large Small	50 S 30 S	70 S
Eukaryotic	28 S, 5.8 S, 5 S 18 S	Large Small	60 S 40 S	80 S

tRNA: the molecular cloverleaf

Prokaryotic and eukaryotic tRNAs are similar in both size and structure. They exhibit extensive secondary structure and contain several **modified ribonucleotides** that are derived from the normal four ribonucleotides. All tRNAs have a similar fold, with four distinct loops that have been described as a **cloverleaf** (Fig. 33.2). The D loop contains several modified bases, including methylated cytosine and dihydrouridine (D), for which the loop is named. The **anticodon loop** is the structure responsible for recognition of the complementary codon of an mRNA molecule: specific interaction of an anticodon of the tRNA with the appropriate codon in the mRNA is due to base pairing between these two complementary trinucleotide sequences. A variable loop, 3–21 nt in length, exists in most tRNAs but its function is unknown. Finally, there is a TψC loop,

which contains a modified base, pseudouridine (ψ). Another prominent structure found in all tRNA molecules is the **acceptor stem**. This structure is formed by base pairing between the nucleotides at each end of the tRNA. The last three bases found at the extreme 3′ end remain unpaired, and always have the same sequence: 5′-CCA-3′. This 3′ end of the acceptor stem is the point at which an amino acid is attached via an ester bond between the 3′-hydroxyl group of the adenosine and the carboxyl group of an amino acid in preparation for protein synthesis (see Chapter 34).

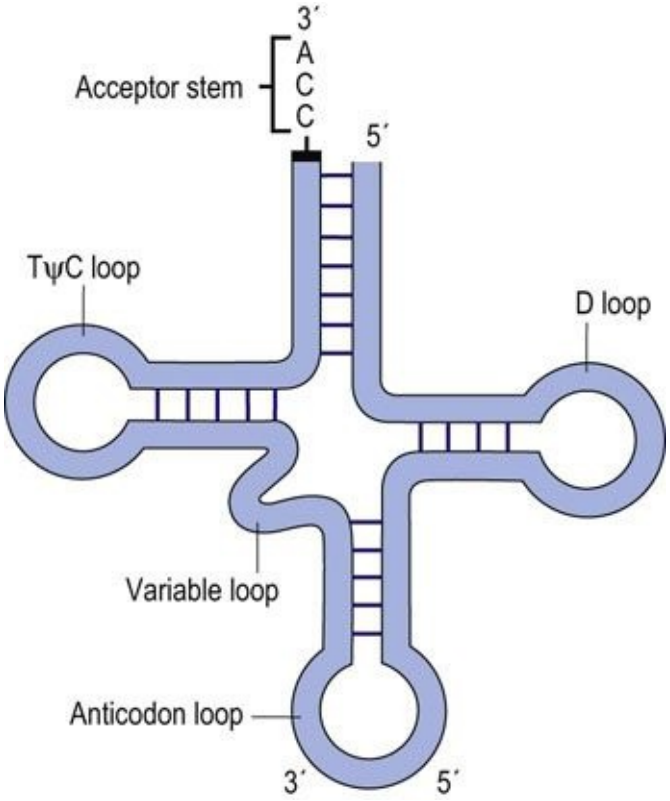


FIG. 33.2 The structure of a tRNA molecule.
A prototypical tRNA molecule is shown and the structures important to its function are indicated. The overall structure of the molecule is due to complementary base pairing between nucleotides within a single RNA. All tRNAs have this basic structure.

mRNA: prokaryotes and eukaryotes have dissimilar mRNAs

Prokaryotes and eukaryotes are very different kinds of organisms with

dramatically different life cycles. Therefore, it is not surprising that there are differences in the structures of their genes, in their mechanisms of transcription, and in the structures of their mRNAs. In fact, we can exploit these differences with novel antibiotics that target unique portions of the prokaryotic life cycle. There are a number of major differences between prokaryotic and eukaryotic mRNAs. These will be discussed in sections below. Briefly, these differences are:

■ Transcriptional units differ in structure: prokaryotic mRNAs are polycistronic, eukaryotic mRNAs are monocistronic (Fig. 33.3).

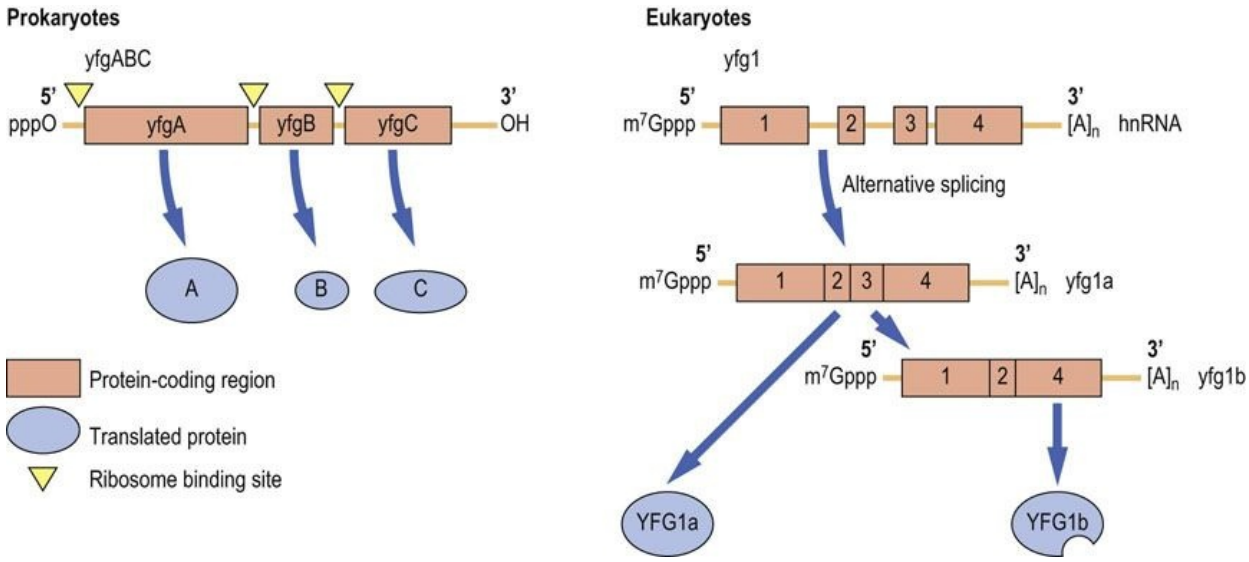


FIG. 33.3 Prototypical structures of prokaryotic (polycistronic) and eukaryotic mRNAs. Prokaryotic mRNAs have naked ends (triphosphate at the 5′ end and hydroxyl at the 3′ end). The boxes indicate those portions of the mRNA that encode a protein. Inverted triangles indicate location of ribosome-binding sites. The three genes in this cistron are translated into three different proteins. Other prokaryotic cistrons may encode up to 10 different proteins in a single mRNA. Nascent eukaryotic mRNA transcripts (**heterogeneous nuclear RNA (hnRNA)**) contain both exons (boxes) and introns (lines). Eukaryotic mRNAs are protected by a 7-methylguanine nucleotide cap (m^7Gppp) at the 5′ end and a polyA tail ($[A]_n$) at the 3′ end of the mRNA. After splicing, the mature mRNAs consist only of exons plus the 5′ and 3′ UTRs. Alternatively spliced mRNAs are translated into different protein isoforms. *yfg*: your favorite gene; UTR: untranslated region.

■ Compartmentalization of transcription and translation: prokaryotes synthesize RNA and protein in one compartment, the cellular cytoplasm, eukaryotes separate these events in the nucleus and cytoplasm.

■ Protection at their 5′ and 3′ ends: ends of prokaryotic mRNAs are naked, eukaryotic mRNAs have a 5′-cap and 3′-poly(A) tail.
■ Processing of mRNAs: prokaryotic mRNAs are not processed; eukaryotic mRNAs contain introns that are spliced out.

A major difference between prokaryotic and eukaryotic mRNAs relates to their transcriptional unit structure. In prokaryotes, transcriptional units generally contain multiple protein-coding regions (see Fig. 33.3), while in eukaryotes, each transcriptional unit generally codes for only a single protein. The **polycistronic mRNAs** of prokaryotes have individual start and stop codons at the beginning and end of each open reading frame, the sequence of mRNA that specifies the sequence of the polypeptide chain. Each stop codon is closely followed by another ribosome binding site and a translation start site that functions for the next open reading frame.

A second major difference between prokaryotic and eukaryotic mRNAs is the compartmentalization of the transcription and translation processes. Because prokaryotes lack a nucleus, transcription and translation are intimately coupled in the cytoplasm; prokaryotic translation is usually initiated before transcription is finished. By coupling these processes, prokaryotes increase the rate at which proteins are expressed, consistent with the relatively short life cycles of prokaryotes. In contrast, eukaryotic cells separate transcription in the nucleus from translation in the cytoplasm. Although this arrangement slows response time for protein production, it allows for much more subtle control of protein expression.

The post-transcriptional processing of mRNAs is also significantly different in prokaryotes and eukaryotes. Because of their importance, these differences will be detailed in a separate section below dealing with post-transcriptional processing of RNAs. Briefly, eukaryotes protect the 5′ and 3′ ends of mRNAs by addition of specific molecular structures (**5′ cap and polyA tail**) that function to reduce mRNA turnover. Also, eukaryotic genes contain introns, untranslated sequences that are present in nascent transcripts and must be spliced out to produce mature mRNAs.

Ribonucleic acid polymerases

RNA polymerases are large multimeric enzymes that transcribe defined segments of DNA into RNA with a high degree of selectivity and specificity

The enzymes responsible for the synthesis of RNA are called RNA polymerases (RNAPol). In contrast to DNA polymerases (Chapter 32), RNA polymerases do not require a primer to initiate RNA synthesis. The RNA polymerases generally consist of two high-molecular-weight subunits and several smaller subunits, all of which are necessary for accurate transcription. Prokaryotes contain a single RNA polymerase that synthesizes all RNAs; however, eukaryotes contain three different RNA polymerases, termed RNA polymerase I, II, and III.

Each polymerases specialize in transcription of one class of RNA

- RNAPol I transcribes ribosomal RNAs. The rRNAs are produced from a single transcriptional unit that is subsequently processed to produce the 18 S, 28 S, 5.8 S and 5 S rRNAs.
- RNAPol II transcribes most genes within a eukaryotic cell, including all protein-coding genes that yield mRNA. RNAPol II is exquisitely sensitive to α-amanitin, a potent and toxic transcription inhibitor found in some poisonous mushrooms.
- RNAPol III transcribes most of the small cellular RNAs, including the tRNAs.

Yeast RNAPol II consists of a 12-subunit core. It exists in two forms. The first is an open form that is shaped like a cupped hand with a cleft that binds the DNA molecule and associated transcription factors near the start point of transcription. After melting or dissociation of the bound DNA strands, the complex undergoes a large structural change that closes the cleft, forming a clamp around the **antisense** or template strand of the DNA. Then, a specific protein (*rbp4/7*) binds to the base of the clamp, locking the clamp in the closed state. The closed form is no longer competent for transcript initiation but is capable of transcript elongation. The yeast RNAPol II structure appears to be an excellent model for the human enzyme. In addition, it is also a good model for the function of RNAPol I and III because the core subunits are either shared or

are homologous between the various enzymes.

The bacterial RNA polymerase is similar to the eukaryotic enzyme complex, except the bacterial enzyme contains only four subunits ($\alpha_2\beta\beta'$) and, unlike the eukaryotic enzyme, requires only a single general transcription factor (σ-factor) to recognize the promoter and recruit the RNA polymerase to initiate transcription.

Messenger ribonucleic acid: transcription

Transcription is a dynamic process that involves the specific interaction of enzymes with DNA to produce RNA molecules

It is convenient to divide transcription into three separate stages: initiation, elongation and termination. Transcription proceeds along the antisense or template strand, producing a **complementary RNA** that is identical to the sense strand of the DNA (Fig. 33.4), except that the thymine residues in DNA are substituted by uracil residues in RNA.

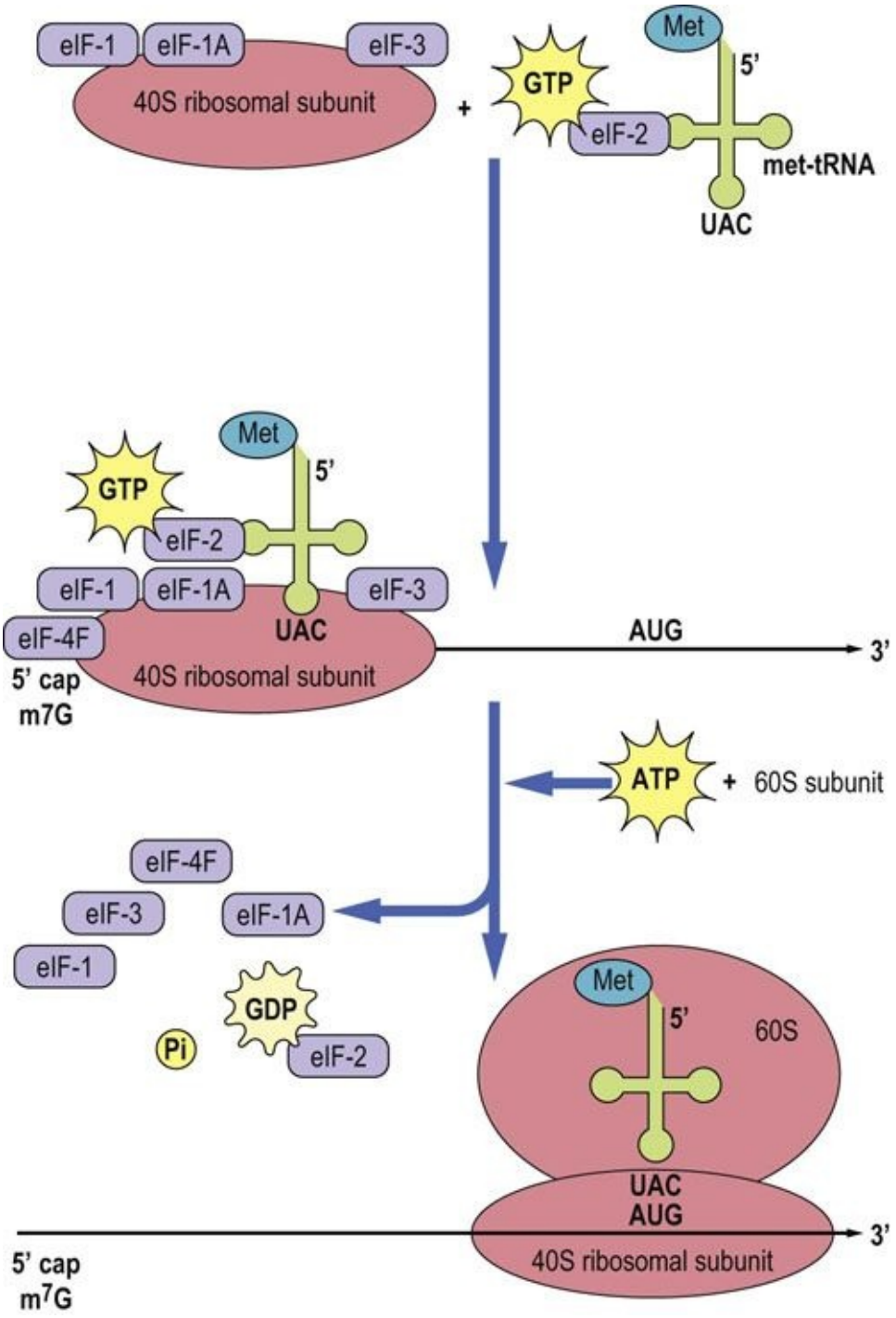


FIG. 33.4 Transcription.
Transcription involves the synthesis of an RNA by RNA polymerase using DNA as a template. The RNA polymerase holoenzyme uses the antisense strand of DNA to direct the synthesis of an RNA molecule that is complementary to this strand.

Clinical box Amanitin poisoning: picking the wrong mushroom

An otherwise healthy young woman presents herself to the emergency room in the early morning with severe nausea, abdominal cramping and copious diarrhea. The patient's vital signs show tachycardia and the skin has poor turgor, indicating dehydration. While giving her medical history, the patient explains that her symptoms began suddenly, about 6 h after she had eaten dinner. Suspecting food poisoning, the patient is asked to recall everything eaten over the past 24 h. The patient reports that she had eaten mushrooms for dinner and added that the mushrooms were picked on a recent hike through the woods. The patient is started aggressively on saline and electrolytes to replenish lost fluids and is given activated charcoal to absorb any residual or recirculating toxins in the gastrointestinal tract. The patient appears to stabilize over the next 24 h and is alert; however, she remains lethargic and the skin begins to take on a yellowish tinge. Blood work shows reduced blood glucose, elevated serum aminotransferase, and increased prothrombin time. Amylase and lipase levels are normal, indicating no pancreatic involvement, and urinalysis indicates no renal involvement. The doctor consults a gastroenterologist who advises increased monitoring of hepatorenal function and continued aggressive intravenous fluid and electrolyte treatment. After approximately 5 days, the patient recovers. What is the biochemical basis of this woman's illness?

Comment.

About 95% of all mushroom fatalities in North America are associated with ingestion of the species *Amanita*. These species produce a toxin, α-amanitin, that binds to RNAPol II and inhibits its function. The first cells that encounter the toxin are those lining the digestive tract. Cells incapable of synthesizing new mRNAs die, causing acute gastrointestinal distress. Liver failure is a serious complication of α-amanitin ingestion due to the induction of apoptosis in liver cells by amanitin. Jaundice and liver function

tests (transaminase, alkaline phosphatase, bilirubin, aminotransferase levels, and prothrombin time) indicate the level of hepatic involvement (see Chapter 30). Most accidental mushroom exposures occur in children younger than 6 years old, who because of their size absorb a larger toxin dose per kg of body weight. In an adult, the ingestion of a single *Amanita phalloides* mushroom can be fatal. Mortality rates range from 10% to 20% of all patients. No specific amatoxin antidote is available, although administration of high doses of penicillin G displace amanitin from circulating plasma proteins, thereby promoting its excretion.

Initiation

Initiation involves the site-specific interaction of the RNA polymerase with DNA

Because most genomic DNA does not encode proteins, identification of transcription start sites is crucial to obtain desired mRNAs. Special sequences termed **promoters** recruit the RNA polymerase to the transcription start site (Fig. 33.5). Promoters are usually located in front (upstream) of the gene that is to be transcribed. However, RNAPol polymerase III promoters are actually located within the gene.

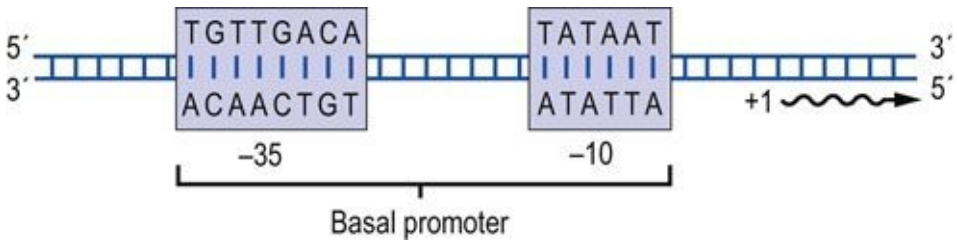


FIG. 33.5 Prokaryotic promoters.
Promoters in prokaryotic genes are located immediately upstream of the transcription start site. Two common conserved regions have been identified that are found at −35 and −10, respectively. The consensus sequences in these regions are shown. Position +1 indicates the first nucleotide that will be transcribed into RNA. The TATA box at –10 is a common AT-rich promoter element.

Prokaryotic genes generally contain simple promoters. Promoters are generally rich in adenine (A) and thymine (T). The presence of these nucleotides facilitates separation of the two DNA strands, because hydrogen bonding between A-T base pairs is weaker than between G-C base pairs. Comparisons of large numbers of prokaryotic promoters have identified two common conserved regions. These are located about 10 nucleotides and 35 nucleotides upstream (–10 and –35 sequences, respectively) from the transcription start site (see Fig. 33.5). The –10 sequence is known as the **TATA box**. This sequence binds the prokaryotic general transcription factor (σ-factor) that interacts and recruits the RNA polymerase to the promoter. Strong promoters tend to match the consensus sequence shown in Figure 33.5; sequences of weaker promoters differ from the consensus sequence and bind the σ-factor and the RNA polymerase less tightly.

In eukaryotic RNAPol II promoters, **regulatory elements** (specific short DNA sequences) termed upstream activation sequences (UASes), enhancers, repressors, CAATT and TATA box sequences are spread over several hundred to several thousand nucleotides. Individual transcription factors (activators or repressors) recognize and bind to these UASes. The control of initiation and the regulation of gene expression are outlined in detail in Chapter 34.

Elongation

Elongation is the process by which single nucleotides are added to the growing RNA chain

In prokaryotes, elongation is a relatively simple process. Ribonucleotides bind to an entry site on the RNA polymerase. If the incoming ribonucleotide matches the next base on the DNA template, the incoming ribonucleotide is transferred into the polymerase active site and a new phosphodiester bond is formed. If it does not match, the ribonucleotide is released and the process repeated until the correct ribonucleotide is found. After the formation of the phosphodiester bond, the RNA polymerase translocates along the template DNA strand. It is thought that the RNA polymerase accomplishes this by oscillating a small helical region of the RNA polymerase molecule between straight and bent conformations, permitting the polymerase to ratchet about 3 Å (=1 nucleotide step) along the antisense strand. After translocation, a new nucleotide is added.

In eukaryotes, after RNA polymerase II initiates transcription, a pair of

negative elongation regulatory factors (NELF and DSIF) trap the RNA polymerase in the starting position. An RNA–protein complex, termed P-TEFb, is a kinase that phosphorylates these two inhibitory molecules, thereby releasing the RNA polymerase to continue RNA synthesis.

Vesicular stomatitis virus (VSV) and HIV produce viral proteins that stabilize the RNA polymerase complex, either directly or by recruiting host factors. The HIV protein TAT (a trans-activating regulatory protein) is one of the better understood of these stabilization proteins. Upon interaction with RNA polymerase the TAT protein rapidly recruits the P-TEFb complex, resulting in increased transcription of the full-length viral RNA, at the expense of cellular RNAs.

Elongation can be a rapid process, occurring at the rate of ~40 nt per second. For elongation to occur, the double-stranded DNA must be continually unwound, so that the template strand is accessible to the RNA polymerase. DNA **topoisomerases** I and II, enzymes associated with the transcription complex, move along the template with the RNA polymerase, separating DNA strands so that they are accessible for RNA synthesis.

Termination

Termination of transcription is catalyzed by multiple mechanisms in both prokaryotes and eukaryotes

At the end of a transcriptional unit, the RNA polymerases terminate RNA synthesis at defined sites. Transcriptional termination mechanisms are much better understood in prokaryotes than in eukaryotes. In prokaryotes, termination occurs via one of two well-characterized mechanisms that both require the formation of hairpin loops in the RNA secondary structure (Fig. 33.6). In *rho*-**in**dependent, intrinsic termination, a hairpin loop is formed just upstream of a sequence of 6–8 uridine (U) residues located near the 3′ end of the transcript. The formation of this secondary structure dislodges the RNA polymerase from the DNA template, resulting in termination of RNA synthesis. In *rho*-dependent termination, the RNA transcript encodes a binding site for *rho* protein, an ATP-dependent helicase; *rho* binds to and travels along the RNA transcript, unwinding it from DNA. It ‘chases’ the RNA polymerase but is slower than RNAPol II. A *rho* termination site near the end of the transcriptional unit causes the RNAPol II to pause, allowing the *rho* protein to catch up and unwind the

RNA:DNA duplex, displacing the RNAPol II from the template, thereby stopping transcription.

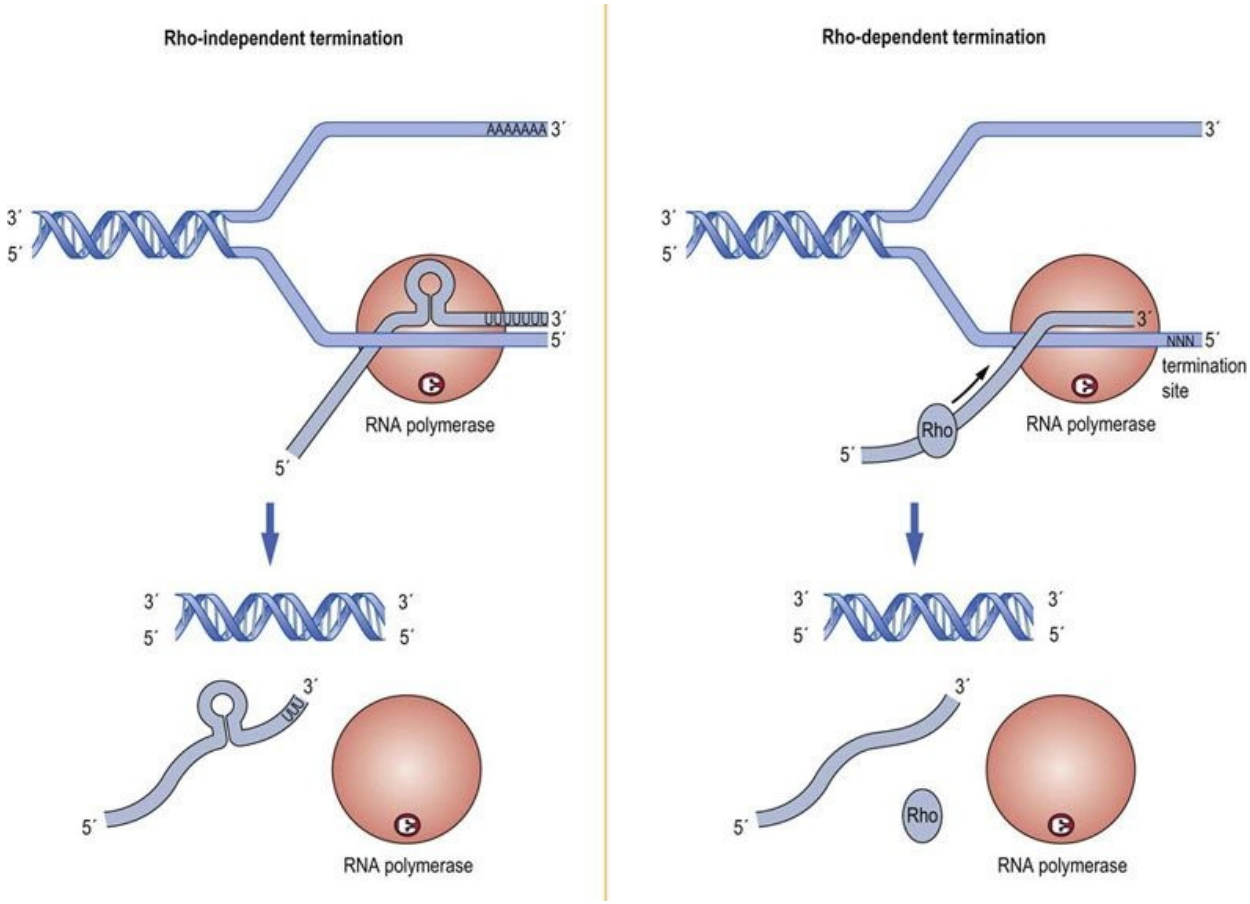


FIG. 33.6 Transcription termination in prokaryotes.
Two mechanisms of transcription termination in bacterial cells are known. *Rho*-independent termination relies on the formation of a secondary structure in the newly transcribed RNA to dislodge the RNA polymerase from the DNA template and stop transcription. *Rho*-dependent termination requires the action of the *rho* protein, an ATP-dependent helicase. This protein moves along the newly transcribed RNA, unwinding the RNA:DNA duplex, catching up with the RNA polymerase when it pauses at the termination site, and causing the polymerase to dissociate from the DNA template.

In eukaryotes, the three RNA polymerases employ different mechanisms to terminate transcription. RNAPol I uses a specific protein, transcript termination factor 1 (TTF1), that binds to an 18 nt terminator site located about 1000 nt downstream of the rRNA coding sequence. When RNAPol I encounters the TTF1 bound to DNA, a releasing factor catalyzes the release of the polymerase

from the rRNA gene. RNAPol III uses a mechanism that is similar to bacterial '*rho*-independent' termination; however, the length of the uridine (U) stretch is shorter and there is no requirement for a RNA secondary structure to dislodge the RNA polymerase. The mechanism of termination by RNAPol II, which transcribes most eukaryotic genes, is not well understood, in part because the RNAPol II products are immediately processed by removal of the nascent 3′ end and addition of a polyadenosine (polyA) tail.

Clinical box A stubborn microbe

A young charity worker reports to your clinic. He has just returned from a prolonged period of work overseas. He complains of fever, weight loss, fatigue and night sweats. He has a productive cough. When telling his medical history, he relates that a local physician treated him with rifampicin, a powerful inhibitor of bacterial RNA polymerase, but his condition has not improved. In fact, he complains that his condition has worsened. On physical exam, you note that he has abnormal breath sounds in the upper lobes of the lungs. A chest radiograph shows a nodular, patchy infiltrate in the upper pulmonary lobes. You suspect tuberculosis, so you order a tuberculin skin test and because culture of mycobacteria can take weeks, you also order a PCR-based test (Chapter 36) specific for the assay of *M. tuberculosis* DNA. You admit the young man to an isolation room. The next day, the PCR-based test is positive for *M. tuberculosis* and after 72 h, the tuberculin skin test shows an induration 10 mm in diameter, indicating a rather strong response. The patient asks you why the antibiotics have not worked.

Comment.

Antibiotics work by targeting specific functions in the cell. Rifampicin is a potent inhibitor of prokaryotic RNA polymerases, but not eukaryotic RNA polymerases. Because rifampicin is one of the indicated antibiotics for tuberculosis but the patient's condition has worsened, you suspect that this young man may have an antibiotic-resistant or multidrug resistant (MDR) form of tuberculosis. Antibiotic-resistant forms are increasingly common in tuberculosis as they are in many bacterial diseases. You order growth tests on the mycobacterium isolate to determine if the strain

is indeed multidrug resistant (MDR-TB) and you immediately start the patient on a multidrug regimen designed to combat MDR-TB. After 2 weeks of daily dosing, the patient can be switched to 2–3 times per week dosing. Vital staining of sputum with fluorescein diacetate (FDA) which indicates living bacilli after initial antibiotic treatment is recommended to confirm MDR-TB Continued treatment of MDR-TB can last for 18–24 months, and consultations should be made with an expert on MDR-TB. Relapse is high for multidrug resistant TB, from 20% to 65%.

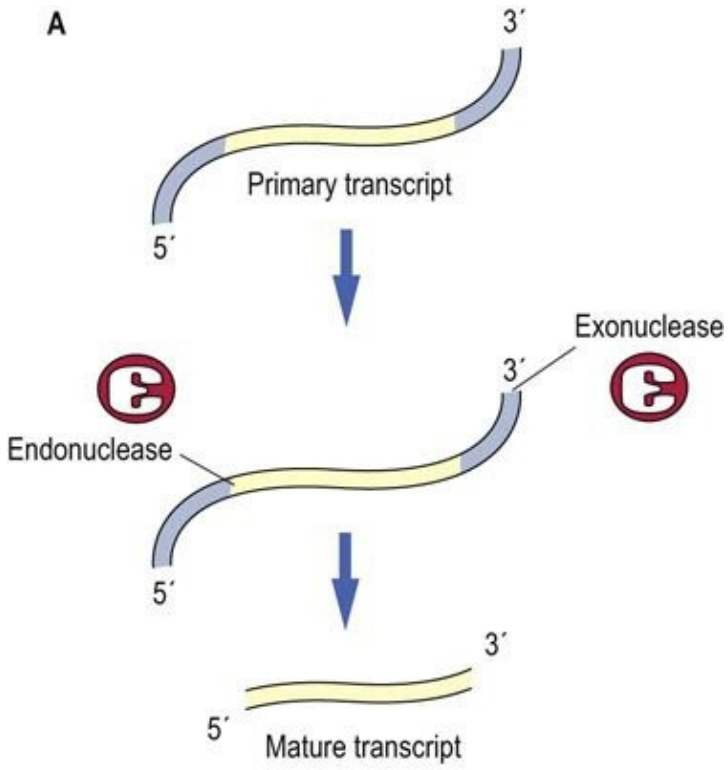
Post-transcriptional processing of ribonucleic acids

The prokaryotic life strategy is to replicate as rapidly as possible when conditions support growth. Eukaryotes have a more controlled life strategy that invests more fully in increased regulation to achieve stable growth but limits rapid reproduction. Both strategies work well for each type of organism as evidenced by the rich diversity of life on earth; mechanisms of RNA synthesis and processing have evolved to optimize each life strategy. This is especially true for mRNAs.

Pre-rRNA and pre-tRNA

tRNAs and rRNAs are synthesized as larger precursors (pre-RNAs) that must be processed to yield mature transcripts (Fig. 33.7)

In prokaryotes a single 30 S rRNA transcript (~6.5 kb) contains specific leader and trailer regions located at the 5′ and 3′ ends of the transcript as well as one copy each of the 23 S, 16 S and 5 S rRNAs. The rRNA genes also contain a number of tRNAs that are embedded in the pre-rRNA transcript. The rRNA transcript must be processed to liberate the functional RNAs. Including each of the ribosomal RNAs in a single transcript is clearly advantageous to maintain the ratio of large-to-small ribosomal subunits.


A
3′
Primary transcript
5′
Exonuclease
3′
Endonuclease
5′
3′
5′
Mature transcript

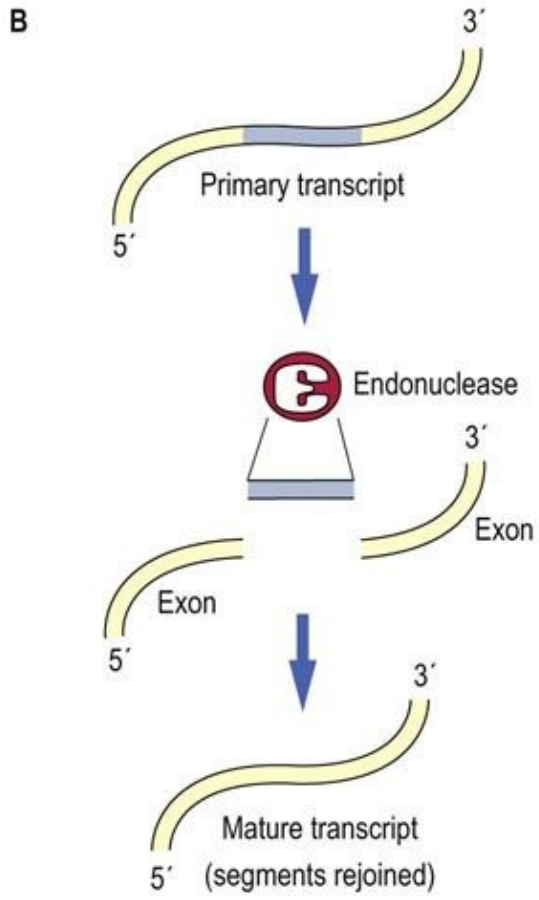

B
3′
Primary transcript
5′
Endonuclease
3′
Exon
Exon
5′
3′
Mature transcript
5′ (segments rejoined)

FIG. 33.7 RNA processing.
There are two general types of RNA processing events. Processing of an RNA transcript can involve (A) the removal of excess sequences by the action of endonucleases and exonucleases as in the processing of rRNA and tRNA genes, or (B) the removal of excess sequences and the rejoining of segments of the newly transcribed RNA as in splicing of mRNAs (see also Fig. 33.8).

In prokaryotes, processing of the pre-rRNA requires several RNases. Ribonuclease III (RNase III) cleaves the pre-rRNA in double-stranded regions. Such regions occur at each end of the 16 S and 23 S rRNAs and their cleavage liberates these rRNAs from the pre-rRNA transcript. The 16 S and 23 S rRNAs are further processed at the 5′ and 3′ ends; however, this trimming requires the presence of specific ribosomal proteins and occurs during ribosome assembly.

In all eukaryotes from yeast to mammals, pre-rRNA transcripts are processed in a manner similar to the prokaryotic pre-rRNA processing. Every 45 S rRNA transcript (~13.7 kb) includes a single copy of the 18 S, 5.8 S and 28 S rRNAs (in eukaryotes, the 5S rRNA is separately encoded). However, processing of the human rRNA is more complex. The pre-rRNA transcript must be cleaved at 11 different sites to generate the mature 18 S, 28 S, and 5.8 S rRNAs. Processing occurs on a huge ribonucleoprotein complex termed the **processome**. In addition to modifications by cleavage, the mature human rRNA contains 115 specific methyl group modifications (in most of these, the methyl group is added to the backbone ribosyl 2′-hydroxyl group, 2′-O-methyl) and 95 specific uridine to pseudouridine (ψ) conversions. These modifications are introduced into the pre-rRNA through the interaction with individual small nucleolar RNA protein complexes (**snoRNPs**, pronounced snorps). Each of these snoRNPs contains a unique guide RNA (~60–300 nt in length) and from one to four protein molecules. Each snoRNP is specific for a single, or at most a few, individual modification sites. The **snoRNA** (small ribonucleolar RNA) component of snRNPs contains structural motifs that belong to two groups, either the C/D box or the H/ACA box. These sequences direct binding of snoRNPs to sites in the pre-rRNA molecules by virtue of a complementary nucleotide stretch (~10–20 nucleotides). This correctly positions a methyl transferase (box C/D) or a pseudouridine synthase (box H/ACA) along the pre-rRNA for modification. The processome contains ≥100 snoRNAs as well as more than 100 individual proteins.

In addition to the rRNA genes, tRNAs are also synthesized in precursor form. As many as seven individual tRNAs can be synthesized from a single pre-tRNA

gene. Processing of the tRNAs from the pre-tRNAs requires RNase P, which cleaves each tRNA from the pre-tRNA by a single cleavage at its 5′ end. RNAse P is an RNA–protein complex containing a 377 nucleotide RNA and a 20 kDa protein, although the protein portion is not required for enzymatic activity, *i.e.* the RNA is catalytic by itself. The discovery of self-splicing RNA – that is, RNA with an enzymatic activity, a **ribozyme** – has led to new ideas about early cellular evolution, which was originally believed to start with amino acids and proteins. It is now believed that ribonucleotides and RNA may have been the most primitive catalytic biopolymers to form on earth, providing for genetic diversity, and that DNA and proteins may have developed later. A third enzyme, RNase D, trims away the extra 3′ nucleotides from the pre-tRNAs, leaving the invariant CCA that is found at the 3′ end of every tRNA. A few tRNAs also contain introns within the anticodon loop which must be removed during processing.

Advanced concept box Ribozymes

RNAs that act like enzymes

In some instances, RNAs have a catalytic activity similar to the type of activities previously ascribed only to proteins (i.e. ribonuclease activity). These unusual catalytic molecules are known as ribozymes. The substrate specificity of a ribozyme is determined by nucleotide base pairing between complementary sequences contained within the enzyme and the RNA substrate that it cleaves. Just like proteinaceous enzymes, the ribozyme will cleave its substrate RNA at a specific site and then release it, without itself being consumed in the reaction. Some RNA viruses, especially plant viruses and virus-like particles such as hepatitis virus delta agent (HVD) that utilize a rolling-circle replicative cycle, rely on the action of ribozymes to cleave viral RNAs from the pre-RNA product.

Because sequences required for ribozyme activity have been identified, ribozymes can be designed that will cleave allele-specific RNAs. Recombinant ribozymes are being considered as possible therapeutic agents for diseases such as muscular dystrophy (DM1), Alzheimer's, Huntington's, and Parkinson's that are caused by the inappropriate expression of an RNA or the

expression of a mutated RNA which participates in disease pathomechanisms. Recent experimental results in cell culture suggest that ribozyme strategies may be highly successful in therapeutic approaches to such diseases.

Pre-mRNA processing

Prokaryotes rapidly synthesize their mRNAs and typically do not process or modify them; both the 5′ and 3′ ends of prokaryotic mRNAs are naked and unprotected. Consequently, even newly synthesized mRNAs are rapidly degraded by normal cellular RNases. This is not a problem for these rapidly growing organisms; they are able to rapidly alter mRNA synthesis for immediate protein production. After their immediate needs have been met, they subsequently degrade their mRNAs and reuse the ribonucleotides for synthesis of other mRNAs. A typical half-life of prokaryotic mRNA is about 3 minutes. In contrast, eukaryotes take special precautions to stably maintain their mRNAs for continued use. mRNA half-lives in eukaryotes range from a few minutes, for some highly regulated transcription factors, to as long as 30 hours for some long-lived transcripts.

Eukaryotic mRNAs have longer half-lives than prokaryotic mRNAs because of protective modifications at their 3′ and 5′ ends

Eukaryotes have evolved methods to protect each end of the mRNA. At the 5′ end, a unique structure termed a ‘5′ cap’ is added. The cap consists of a **7-methylguanidine** residue that is attached in reverse orientation to the first nucleotide of the mRNA, *i.e.* by a 5′-to-5′ triphosphate linkage (m^7Gppp). Most cellular exo-RNases do not have the ability to hydrolyze this cap from the mRNA, so the 5′ end is immune in their presence. At the 3′ end of nearly all eukaryotic mRNAs (with the exception of histone mRNAs), a polyadenosine track is added, termed the **polyA tail**. The adenosine residues are not encoded by the DNA but instead are added by the action of poly(A) polymerase using ATP as a substrate. This polyA tail is frequently >250 nucleotides in length. Although it is still susceptible to the action of exo-RNases, the presence of the polyA tail

significantly reduces the turnover of mRNA, thereby increasing its lifetime. The presence of the polyA tail has historically been used to isolate mRNA from eukaryotic cells.

The spliceosome joins exons from pre-mRNA to form a mature mRNA

In the more complicated post-transcriptional processing of eukaryotic mRNAs, sequences called **introns** (intravening sequences) are removed from the primary transcript (pre-RNA, a form of hnRNA) and the remaining segments, termed **exons** (expressed sequences), are ligated to form a functional RNA. This process involves a large complex of proteins and auxiliary RNAs called small nuclear RNAs (snRNAs), which interact to form a spliceosome. The function of the five snRNAs (U1, U2, U4, U5, U6) in the spliceosome is to help position reacting groups within the substrate mRNA molecule, so that the introns can be removed and the appropriate exons can be spliced together precisely (Table 33.3). The snRNAs accomplish this task by binding, through base-pairing interactions, with the sites on the mRNA that represent intron/exon boundaries.

Table 33.3

The function of small nuclear RNAs (snRNAs) in the splicing of mRNAs

snRNA	Size	Function
U1	165 nt	Binds the 5′ exon/intron boundary
U2	185 nt	Binds the branch site on the intron
U4	145 nt	Helps assemble the spliceosome
U5	116 nt	Binds the 3′ intron/exon boundary
U6	106 nt	Displaces U1 after first rearrangement

The removal of an intron and rejoining of two exons can be considered to occur in three steps (Fig. 33.8). The first step involves the binding of the U1 snRNP to the exon/intron boundary at the 5′ end of the intron, along with the binding of the U2 snRNP to a target adenosine nucleotide, usually found about 30 nt from the 3′ end of the intron, Following binding of the U4/U5/U6 snRNP complex just upstream of the 5′ splice site, the intron loops back on itself positioning the ends of the introns in the correct orientation. In this process, the U6 snRNP displaces the U1 snRNP from the mRNA. Subsequently, a transesterification reaction between the 2′-hydroxyl of the target adenosine with the phosphodiester bond of the intron's 5′-guanosine residue breaks the upstream

splice site and forms a branched chain structure in which the target adenine has 2′-, 3′-, and 5′-phosphate groups. The looped structure of the intron is similar in appearance to a cowboy's lariat. Following a subsequent physical rearrangement which releases the U4 snRNP, a second transesterification reaction ligates the 3′ end of exon 1 with the 5′ end of exon 2. The splicosome then disassembles releasing the **lariat structure**, which is degraded and the mature mRNA is ready for further processing, including addition of the M7G cap and polyA tail, and nuclear export.

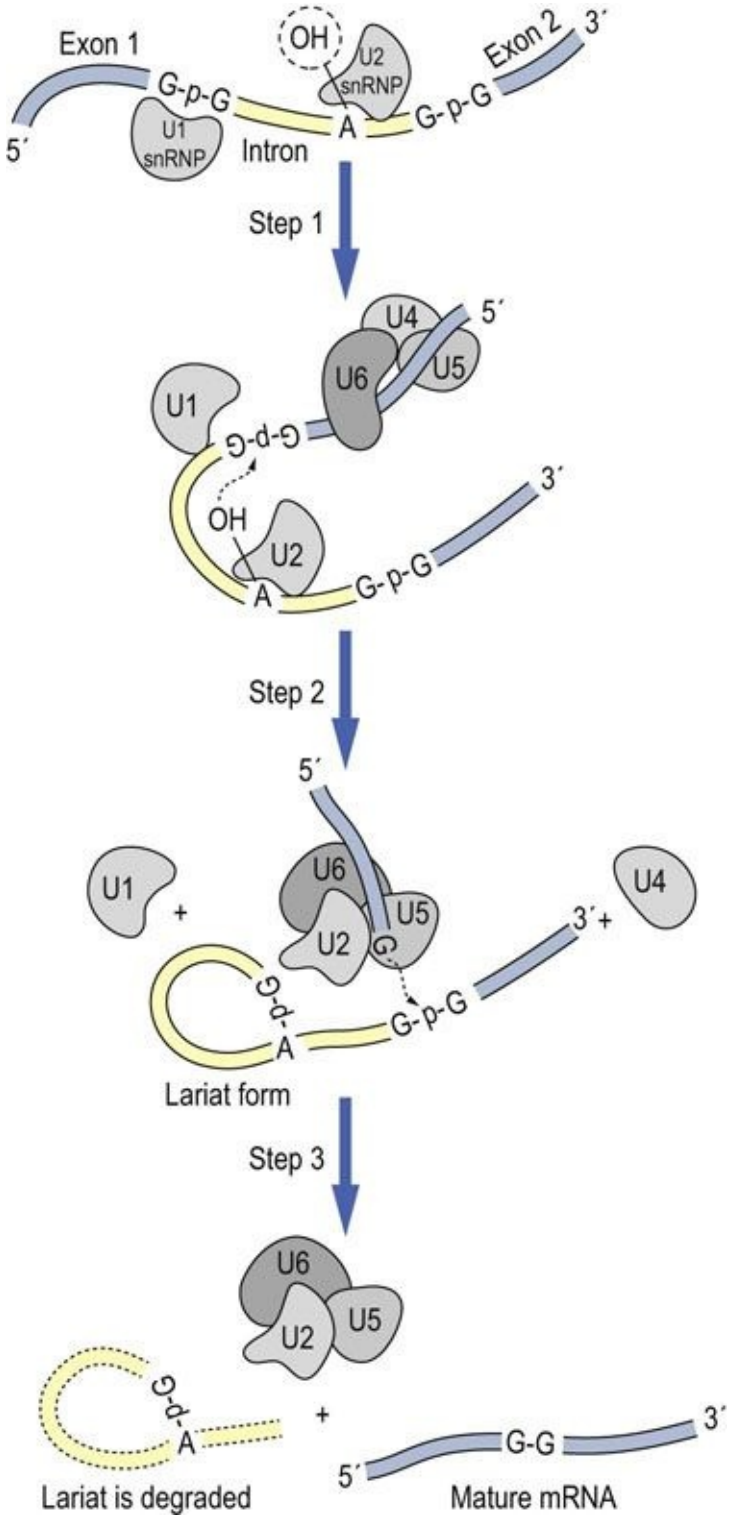


FIG. 33.8 RNA splicing.
RNA splicing is a multistep process catalyzed by ribonucleoprotein complexes, “simplified” in the above diagram. In a transesterification reaction, the phosphate bond of a guanosine residue at the 5′ exon/intron boundary is broken and joined to the 2′-OH of an adenine residue located in the middle of the intron. In a later step, the phosphate

bond at the 3′ intron/exon boundary is first cleaved and then the two exons are spliced together by reformation of a phosphodiester bond between the nucleotides at either end of the exons. The intron is eliminated in the form of a lariat structure, cyclized through a 2′,3′5′-phosporylated adenosine residue. N = any nucleotide.

Alternative splicing produces multiple mature mRNAs from a single pre-mRNA transcript

Most eukaryotic mRNAs consist of multiple introns and multiple exons. If splicing were consistent, only a single mature mRNA would result from the pre-mRNAs. However, many eukaryotic genes undergo a process called alternative splicing, in which different mRNA regions are removed from the pre-mRNA, resulting in multiple mature mRNAs with different sequences (see Fig. 33.3). When these different mRNAs are translated, multiple protein isoforms are made. In humans, almost 60% of pre-mRNAs give rise to multiple mature mRNAs following alternative splicing. About 80% of these alternatively spliced mRNAs result in alternations in the encoded proteins (splicing events may also occur in the 5′ and 3′ untranslated regions). Alternative splicing can result in insertion or deletion of amino acids in the protein sequence, shifts in reading frames, or even introduction of novel stop codons. Such alternative splicing can also add or remove mRNA sequences that can alter regulatory elements affecting translation, mRNA stability or subcellular localization. In vertebrates, the introns of many ribosomal protein genes host the sequences of the small nucleolar RNAs (snoRNAs) that function in modification of specific residues in rRNAs and tRNAs; however, many introns do not host these sequences and their function remains obscure.

Editosomes modify the nucleotide sequence of mature mRNAs

Finally, in some instances, mRNAs are post-transcriptionally edited via one of several different mechanisms. These include specific C-to-U modifications catalyzed by cytosine deaminases or A-to-I modifications catalyzed by adenosine deaminases or deletion of single or multiple U residues to change the nucleotide sequence of the mRNA. These changes lead to codon differences that result in altered protein sequences that differ from the sequences encoded at the gene level (see Chapter 35). A large multiprotein complex termed the editosome is involved in this process. The snoRNAs that function in rRNA and tRNA

modification also function in these RNA editing processes. These snoRNAs bind to the editing sites within the mRNAs and position the editosome with either cytosine deaminase or adenosine deaminase activity at the correct position to complete the modification.

Advanced concept box Genomic imprinting

snoRNAs also function in genomic imprinting. Genomic imprinting is an epigenetic process that results in differential expression of maternal and paternal genes in a developing embryo. This process occurs in both mammals and plants. Specific genes are methylated during meiosis and thereby inactivated. This methylation occurs differentially in oocyte and spermatocyte development, thereby uniquely inactivating either the maternal or the paternal alleles during embryonic development. Genomic imprinting can affect a large number of conditions in humans, including susceptibility to asthma, cancer, diabetes, obesity and a number of developmental disorders, including Angelman and Prader–Willi syndromes. These disorders are caused by deletion of the same region (band 11) of chromosome 15. However, paternal inheritance of this deletion results in Prader–Willi syndrome, whereas maternal inheritance results in Angelman disorder. They are dissimilar disorders because the deleted region contains two paternally expressed genes (small nuclear ribonucleoprotein-associated protein N and necdin) and one maternally expressed gene (ubiquitin-protein ligase E3A). At least 83 human genes are known to be imprinted and more than 1300 in mice.

Advanced concept box RNAi as a therapeutic option

Age-related macular degeneration (AMD) is the leading cause of blindness for the elderly in the developed world. AMD results

from an atrophy of the macula in the retina. The result of atrophy is a loss of central vision, which can lead to the inability to read or even to recognize faces. The most severe type of AMD (the wet form) causes vision loss due to the growth of blood vessels (neovascularization) in the retinal choriocapillaries, which, if untreated, leads to blood and protein leakage beneath the macula. This eventually causes scarring and irreversible damage to the photoreceptors.

One mechanism that causes neovascularization is the aberrant expression of the proangiogenic Vascular Endothelial Growth Factor (VEGF) in the retina, which results in blood vessel outgrowth. One treatment has been to use anti-VEGF antibodies (ranibizumab (Lucentis) or bevacizumab (Avastin)) injected once per month directly into the vitreous humor of the eye. These antibodies bind to and inactivate VEGF, thereby reducing the level of angiogenesis and prolonging eyesight.

The ability of a cell to downregulate specific mRNA levels coupled with the localized treatment area offered by the vitreous humor make the VEGF gene an ideal candidate for RNAi-mediated downregulation. Small RNA molecules complementary to VEGF mRNA are injected into the vitreous humor. When these molecules are taken into the cell, they function as siRNAs, causing degradation of VEGF mRNA and decreasing VEGF biosynthesis. Phase I clinical trials have shown that >75% of patients showed improved or stable vision 2 weeks after a single injection of the PF-655 siRNA and that intravitreal injection of the naked siRNAs were safe in doses up to 3 mg. A similar approach might also work for treatment of diabetic retinopathy. Other RNAi trials are currently under way for respiratory syncytial virus (RSV), hepatitis C, Huntington's disease, HIV and cancer.

Selective degradation or inactivation of ribonucleic acid

Interferon activates pathways that inhibit proliferation of RNA viruses

RNA viruses pose a major challenge for eukaryotic cells; however, natural defense mechanisms have evolved to limit viral infection. Because RNA viruses generally form a double-strand (ds) replicative intermediate during their life cycle, this dsRNA is a unique structure that is not generally found in eukaryotic cells. It can be recognized by dsRNA-binding proteins that will subsequently trigger responses to limit viral infection. One of these mechanisms involves the **dsRNA-activated protein kinase R (PKR)**. When activated by binding dsRNA, this enzyme can phosphorylate and inactivate the protein translation factor, eIF2α, thereby downregulating translation of viral RNA. Likewise, when activated by dsRNA, the enzyme 2′-5′-oligoadenylate synthase polymerizes ATP into a series of short nucleotides (2′-5′-oligoadenylate, 2–5 A) that differs from the 5′-3′ structure found in normal RNA. The most active form is a trimer, pppA-2′-p-5′-A-2′-p-5′-A. The accumulation of 2–5 A inhibits viral (and host) protein translation by activating an endoribonuclease (RNase L) that indiscriminately degrades both mRNAs and rRNAs within the cell. Both genes, PKR and 2–5 A synthase, are induced by interferon, which is itself upregulated by viral **infection**. This results in an efficient amplification mechanism, leading to programmed cell death (**apoptosis**) to limit the growth and spread of the virus.

RNAis exert wide-ranging effects on gene expression

Another important process in innate cellular immunity involves specific targeting of RNA sequences for rapid degradation. This process is termed **RNA interference** (RNAi) (see Chapter 35) or post-transcriptional gene silencing (PTGS). While this process is thought to have evolved as a defense against double-strand forms of RNA viruses, it also functions as an endogenous mechanism of gene regulation during development of many eukaryotes. PTGS begins when duplex RNA is recognized within cells by the enzyme Dicer. **Dicer** is an endonuclease active against dsRNAs. It produces short duplex RNA fragments from 21 to 25 nucleotides long, termed small interfering RNAs

(siRNAs) or microRNAs (miRNAs). While siRNAs generally arise from virally produced dsRNAs, miRNAs are transcribed from introns of normal genes or from other nonprotein coding genes, all of which show precise regulation and expression. These miRNA sequences form hairpin structures that are recognized by Dicer to liberate the multiple small miRNAs.

miRNAs interact with a multiprotein complex, termed the RNA-induced silencing complex, **RISC**. The RISC is composed of a number of components, including the enzyme **Argonaut**. Argonaut binds the small dsRNA fragments, unwinds them, and incorporates one of the strands (the guide strand) to act as an RNA targeting cofactor; the other strand is degraded. Using the incorporated si/miRNA, the RISC complex can identify sequences complementary to its RNA, whereupon it will either base pair and cleave the complementary RNAs triggering the rapid degradation of complementary RNAs (siRNAs) or will recruit additional proteins that repress the translation of the complementary RNAs via multiple mechanisms (miRNAs).

As many as 1000 miRNAs may function in humans. Some miRNAs target multiple genes; some genes have are targets of multiple miRNAs; and miRNAs are probably involved in some way in regulation of most metabolic pathways. miRNAs regulate central metabolic processes such as adipocyte differentiation and insulin secretion, which are involved in development of obesity and diabetes. Recently, the miR-21 miRNA was identified as an antiapoptotic signal. This miRNA maintains normal apoptotic responses in cells. When miR-21 is aberrantly expressed, as it is in many forms of cancer, it inhibits apoptosis and promotes cellular proliferation.

Summary

■ The major products of transcription are the rRNAs, tRNAs, and mRNAs. These RNAs perform specific functions within a cell: mRNAs carry the genetic information from nuclear DNA to ribosomes for protein synthesis; rRNAs interact with proteins to form ribosomes, the basic cellular machinery on which protein synthesis occurs; and tRNAs function as amino acid carriers that translate the information stored in the mRNA nucleotide sequence to the amino acid sequence of proteins.

■ In eukaryotic cells, each of these RNA classes is produced by a different, specific RNA polymerase (RNAPol I, II or III, respectively), while in bacterial cells a single RNA polymerase synthesizes all three classes of RNA.

■ The basic structures of rRNAs and tRNAs in eukaryotic and bacterial cells are similar. However, mRNAs from eukaryotic cells have a 5′ (m^7Gppp) cap and a 3′ ($[A]_n$) tail. Prokaryotic mRNA transcripts do not have these modifications on their 5′ and 3′ ends and can be polycistronic.

■ Most, but not all, eukaryotic mRNAs undergo a process called splicing to be functional, whereas prokaryotic mRNAs are functional as soon as they are synthesized. Splicing involves the removal of sequences called introns and the rejoining of exon sequences to each other to form a mature functional mRNA.

■ The process of transcription consists of three parts: initiation, elongation, and termination. Initiation involves the recognition and binding of promoter sequences by RNA polymerase and associated transcriptional cofactors. Elongation involves the selection of the appropriate nucleotide and formation of the phosphodiester bridges between each nucleotide in an RNA molecule. Finally, termination involves the dissociation of the RNA polymerase from the DNA template. This is mediated by either RNA secondary structure or specific protein factors.

■ Unique cellular mechanisms recognize double-stranded RNA that limit viral infection by multiple mechanisms, some that induce overall mRNA degradation and others that target specific mRNAs for degradation.

Active learning

1. Which commonly used antibiotics are directed at inhibition of bacterial RNA polymerase but do not affect the mammalian complex? Why are these drugs less effective against fungal

infections?
2. Review the pathogenesis of systemic lupus erythematosus, an autoimmune disease in which antibodies to ribonucleoprotein particles are implicated in the development of chronic inflammation.
3. Review the pathogenesis of the thalassemias, with emphasis on those variants in which gene mutations affect the synthesis, processing and splicing of the RNA for hemoglobin, leading to anemia.

Further reading

Atasoy, D, Betley, JN, Su, HH, *et al.* Deconstruction of a neural circuit for hunger. *Nature*. 2012; 488:172–177.

Bora, RS, Gupta, D, Mukkur, TK, *et al.* RNA interference therapeutics for cancer: challenges and opportunities (review). *Mol Med Report*. 2012; 6:9–15.

Bumcrot, D, Manoharan, M, Koteliansky, V, *et al.* RNAi-therapeutics: a potential new class of pharmaceutical drugs. *Nat Chem Biol*. 2006; 2:711–719.

Butler, MG. Prader–Willi syndrome: obesity due to genomic imprinting. *Curr Genomics*. 2011; 12:204–215.

David, CJ, Chen, M, Assanah, M, *et al.* HnRNP proteins controlled by c-Myc deregulate pyruvate kinase mRNA splicing in cancer. *Nature*. 2010; 463:364–368.

Davis, ME, Zuckerman, JE, Choi, CH, *et al.* Evidence of RNAi in humans from systemically administered siRNA via targeted nanoparticles. *Nature*. 2010; 464:1067–1070.

Ishida, M, Moore, GE. The role of imprinted genes in humans. *Mol Aspects Med*. 2013; 34:826–840.

Kaiser, PK, Symons, RC, Shah, SM, *et al.* RNAi-based treatment for neovascular age-related macular degeneration by Sirna-027. *Am J Ophthalmol*. 2010; 150:33–39.

Keaney, J, Campbell, M, Humphries, P. From RNA interference technology to effective therapy: how far have we come and how far to go? *Ther Deliv*. 2011; 2:1395–1406.

Liu, H, He, L, Tang, L. Alternative splicing regulation and cell lineage differentiation. *Curr Stem Cell Res Ther*. 2012; 7:400–406.

Mastroyiannopoulos, NP, Uney, JB, Leonidas, A, *et al.* The application of ribozymes and DNAzymes in muscle and brain. *Molecules*. 2010; 15:5460–5472.

Matlin, AJ, Clark, F, Smith, CWJ. Understanding alternative splicing: towards a cellular code. *Nat Rev Mol Cell Biol*. 2005; 6:386–398.

Zhou, J, Rossi, JJ. Progress in RNAi-based antiviral therapeutics. *Methods Mol Biol*. 2011; 721:67–75.

Websites

RNA polymerase:

www.rcsb.org/pdb/101/motm.do?momID=40

en.wikipedia.org/wiki/RNA_polymeraseα-Amanitin

toxnet.nlm.nih.gov/cgi-bin/sis/search/a?dbs+hsdb:@term+@DOCNO+3458

emedicine.medscape.com/article/1008902-overview

RNAi:

nobelprize.org/nobel_prizes/medicine/laureates/2006/index.html

www.youtube.com/watch?v=cK-OGB1_ELE

www.pbs.org/wgbh/nova/body/rnai-cure.html

Macular degeneration:

www.ahaf.org/macular/

www.lucentis.com/lucentis/

www.clinicaltrials.gov/ct2/show/NCT00470977?term=NCT00470977&rank=1

Spliceosome:

neuromuscular.wustl.edu/pathol/spliceosome.htm

www.rcsb.org/pdb/101/motm.do?momID=65

Alternative splicing:

en.wikibooks.org/wiki/Proteomics/Alternative_splicing_and_its_impact_on_protein_identification

www.eurasnet.info/alternative-splicing/what-is-alternative-splicing

CHAPTER 34

Protein Synthesis and Turnover

Jeffrey R. Patton and Gary A. Bannon

Learning objectives

After reading this chapter you should be able to:

■ Describe how various RNAs involved in protein synthesis interact to produce a polypeptide.

■ Outline the structure and redundancy of the genetic code.

■ Explain how proteins are targeted to specific subcellular organelles.

■ Describe the major steps in synthesis and degradation of a cytosolic protein.

Introduction

Translation is the process by which the information endoded in an mRNA is translated into the primary structure of a protein

Protein synthesis or translation represents the culmination of the transfer of genetic information, stored as nucleotide bases in deoxyribonucleic acid (DNA), to protein molecules that are the major structural and functional components of living cells. It is during translation that this information, expressed as a specific nucleotide sequence in a ribonucleic acid (mRNA) molecule, is used to direct the synthesis of a protein. The protein then folds into a three-dimensional structure that is defined, in large part, by its amino acid sequence. In order to translate an mRNA into protein, three main RNA components are necessary:

- ribosomes, containing ribosomal RNA (rRNA)
- messenger RNA (mRNA)
- transfer RNA (tRNA).

The ribosome, composed of rRNAs and a number of proteins, is the macromolecular machine on which all protein synthesis occurs. The information required to direct the synthesis of the primary sequence of the protein is contained in mRNA. The amino acids that are to be incorporated into the protein are attached to tRNAs. The ribosome interacts with the tRNA molecules and mRNA so that the correct amino acid is incorporated into the protein. The translation of mRNA begins near the 5′ end of the template and moves towards the 3′ end, and proteins are synthesized starting with their amino-terminal ends. Therefore, the 5′ end of the RNA encodes the amino-terminal end of the protein and the 3′ end of the RNA encodes the carboxyl-terminal end of the protein.

This chapter begins with an introduction to the genetic code and the components needed for protein synthesis. This is followed by presentation of the structure and function of the ribosome, detailing the process of translation by outlining the initiation, elongation, and termination of protein synthesis, and the mechanism by which proteins are targeted to specific locations in the cell. Following a discussion of post-translational modifications of proteins, the chapter ends with description of the role of a second macromolecular complex, the proteasome, in protein turnover.

The genetic code

The genetic code is degenerate and not quite universal

The mRNA to be used for translation has only four nucleotides – adenosine, A; cytidine, C; guanosine, G; and uridine, U – but it will encode a protein containing as many as 20 different amino acids. So there is not a one-to-one correspondence between nucleotide and amino acid sequence; instead, a series of three nucleotides in the mRNA, known as a codon, are required to specify each amino acid. When all combinations of four nucleotides are taken into account, three at a time, 64 possible codons result (Table 34.1). Three of these codons (UAA, UAG, UGA) are **stop codons**, used to signal the termination of protein synthesis; they do not specify an amino acid. The rest specify the 20 amino acids, which illustrates a feature of the genetic code known as **degeneracy**: more than one codon can specify a specific amino acid. For example, codons GUU, GUC, GUA, and GUG all code for the amino acid valine. Indeed, all the amino acids, with the exception of methionine (AUG) and tryptophan (UGG), have more than one codon. The codon AUG, which specifies only methionine, encodes methionine anywhere it appears in the RNA and it also marks the starting point for protein synthesis (see below for a few exceptions).

Table 34.1

The genetic code

1st position	2nd position				3rd position
	G	A	C	U	
G	Gly	Glu	Ala	Val	G
	Gly	Glu	Ala	Val	A
	Gly	Asp	Ala	Val	C
	Gly	Asp	Ala	Val	U
A	Arg	Lys	Thr	Met	G
	Arg	Lys	Thr	Ile	A
	Ser	Asn	Thr	Ile	C
	Ser	Asn	Thr	Ile	U
C	Arg	Gln	Pro	Leu	G
	Arg	Gln	Pro	Leu	A
	Arg	His	Pro	Leu	C
	Arg	His	Pro	Leu	U
U	Trp	Stop	Ser	Leu	G
	Stop	Stop	Ser	Leu	A
	Cys	Tyr	Ser	Phe	C
	Cys	Tyr	Ser	Phe	U

The genetic code is degenerate, meaning more than one codon can code for an amino acid, and in many cases changing the nucleotide at the third position does not change the amino acid encoded. In order to find the sequence(s) of the codons that encode a particular amino acid, one simply finds the amino acid in the table and combines the nucleotide sequence for each position. For example, methionine (Met) is encoded by the sequence AUG. To find the amino acid that matches a codon sequence, reverse this process.

The genetic code as specified by the triplet nucleotides is, for the most part, the same for bacteria and humans, and is referred to as ‘universal’. However, there are exceptions and some examples are found in bacteria and mitochondria. In bacteria, if the codons GUG and UUG occur at the beginning of protein synthesis, they can be read as a methionine codon. There are also minor differences in the genetic code in mitochondria; for instance, in vertebrate mitochondria there are additional codons that can encode methionine, UGA which is normally a stop codon, encodes tryptophan, and there are additional stop codons.

Another aspect of the genetic code is that once synthesis has started at an AUG codon for methionine, each successive triplet from that start point will be read in register without interruption until a termination codon is encountered. Thus the **reading frame** of the mRNA will be dictated by the start codon. Mutations that cause the addition or deletion of even single nucleotides will cause a reading frame shift, resulting in a protein with a different amino acid sequence after the mutation or a protein that is prematurely terminated if a stop codon is now in frame (nonsense mutation; Table 34.2).

Table 34.2

Effect of mutations on protein synthesis

Description of change in gene sequence	mRNA sequence	Protein sequence	Result of change
Normal gene	AUG GGG AAU CUA UCA CCU GAU C …	Met-Gly-Asn-Leu-Ser-Pro-Asp-…	Normal protein
Insertion of a **C**	AUG GG**C** GAA UCU AUC ACC UGA UC …	Met-Gly-Glu-Ser-Ile-Thr-Stop	Premature stop
Deletion of an **A**	AUG GGG AAU CUA UCC CUG AUC …	Met-Gly-Asn-Leu-Ser-Leu-Ile-…	Different sequence
Substitution **CG** for UC	AUG GGG AAU CUA **CG**A CCU GAU C …	Met-Gly-Asn-Leu-Arg-Pro-Asp-…	Substitution
Substitution **G** for A	AUG GGG AAU CU**G** UCA CCU GAU C …	Met-Gly-Asn-Leu-Ser-Pro-Asp-…	No change (silent)
Substitution **G** for C	AUG GGG AAU CUA U**G**A CCU GAU C …	Met-Gly-Asn-Leu-Stop	Premature stop

Mutations in a gene are transcribed into the mRNA and the resulting changes in the protein sequence are shown. Note that, depending on the position of the mutation, single nucleotide substitutions can result in **silent changes**, a change in a single amino acid (**missense**) or even premature termination (**nonsense**).

Clinical box Sickle cell anemia: mutation of the genetic code

Sickle cell anemia is an example of a disease in which a single nucleotide change within the coding region of the gene for the β-chain of hemoglobin A, the major form of adult hemoglobin, yields

an altered protein that has impaired function (see Chapter 5). The mutation that causes this disease is a single nucleotide change in a codon that normally specifies glutamate (GAG) and which now produces a codon that specifies valine (GUG). Under conditions of low oxygen tension, this single amino acid change allows the protein to polymerize into rod-shaped structures, resulting in deformation of red blood cells and in altered flow properties of the cells in vessels and capillaries. When oxygen is not bound to hemoglobin, the conformation of the protein, deoxyhemoglobin, exposes a hydrophobic patch, allowing the mutant proteins with valine to interact more readily and form rods. This substitution of an amino acid with an acidic side chain for an amino acid with a nonpolar, hydrophobic side chain is termed a nonconservative mutation. Conservative replacement of one amino acid by another with similar physical and chemical properties may have less severe consequences, *e.g.* an Arg → Lys or Asp → Glu mutation.

The machinery of protein synthesis

The ribosome is multi-step assembly line for protein synthesis

Ribosomes, the molecular machines that conduct protein synthesis, consist of a small and a large subunit that, when associated with each other, possess three specific sites at which tRNAs bind. These sites are known as the aminoacyl-tRNA, or A site, the peptidyl-tRNA, or P site, and the exit site, or E site. The A site is where a donor tRNA molecule, carrying the appropriate amino acid on its acceptor stem, is positioned before that amino acid is incorporated into the protein. The P site is the location in the ribosome that contains a tRNA molecule with the amino-terminal polypeptide of the newly synthesized protein still attached through its carboxyl terminus to its acceptor stem. It is within these sites that the process of peptide bond formation takes place. This process is catalyzed by a peptidyl transferase activity, which forms the peptide bond between the amino group of the amino acid in the A site and the carboxyl terminus of the nascent peptide attached to the tRNA in the P site. The E site is where the deacylated tRNA moves once the peptide bond is formed and it will soon be exiting the ribosome. The E site, which provides a third site of interaction between tRNA and mRNA on the ribosomes, appears to be essential for maintaining the reading frame and assuring the fidelity of translation.

- A site – donor tRNA-amino acid
- P site – tRNA-growing peptide chain
- E site – site occupied by deacylated tRNA.

Each amino acid has a specific synthetase that attaches it to all the tRNAs that encode it

There is a distinct tRNA molecule for most of the codons represented in Table 34.1. The amino acid is attached to the acceptor stem of the tRNA by an enzyme called aminoacyl-tRNA synthetase. This enzyme catalyzes the formation of an ester bond linking the 3′ hydroxyl group of the adenosine nucleotide of the tRNA to the carboxyl group of the amino acid (Fig. 34.1). The attachment of an amino acid to a tRNA is a two-step reaction. The carboxyl group of the amino acid is first activated by reaction with adenosine triphosphate (ATP) to form an

amino-acyladenylate intermediate, which is bound to the synthetase complex. The enzymology of activation of the carboxyl group of amino acids is similar to that for activation of fatty acids by thiokinase (Chapter 15), but, rather than transfer of the acyl group to the thiol group of coenzyme A, the aminoacyl group is transferred to the 3′-hydroxyl of the tRNA. The product is described as a **charged tRNA** molecule. At this point it is ready to bind to the A site of the ribosome, where it will contribute its amino acid to a growing peptide chain. There is a different synthetase specific for each of the 20 amino acids in protein. This synthetase attaches the appropriate amino acid to all the tRNAs that bind that amino acid.

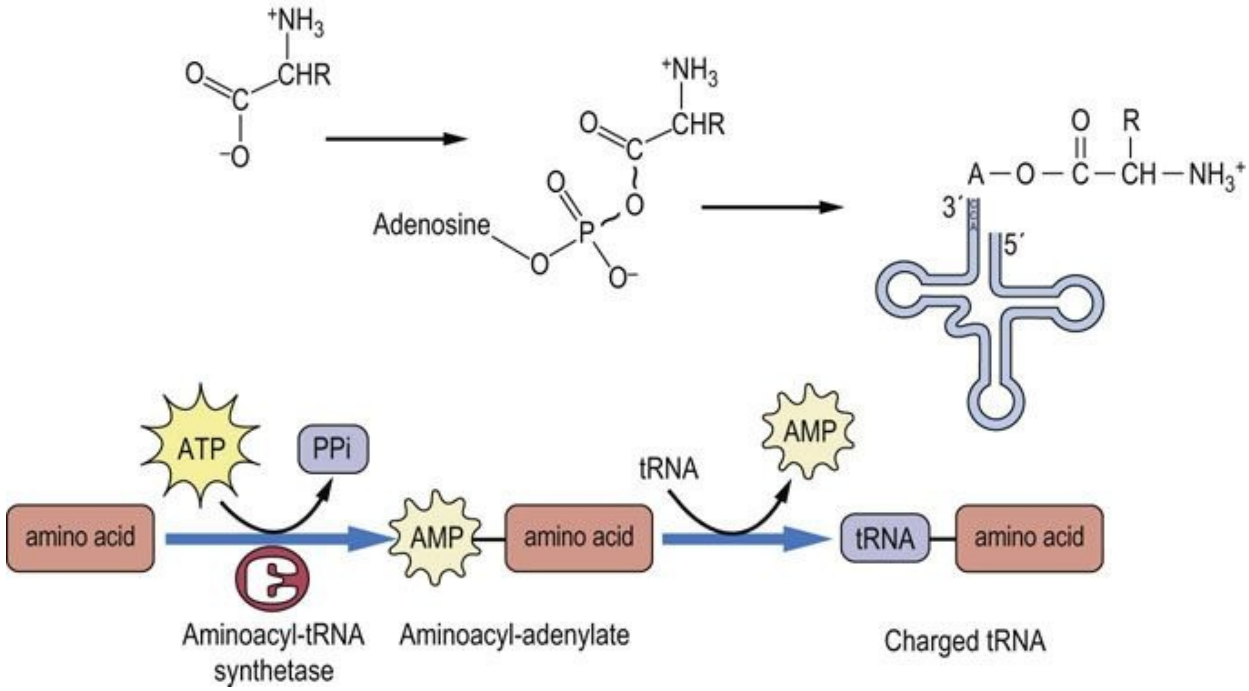


FIG. 34.1 Activation of an amino acid and attachment to its cognate tRNA. The amino acid must be activated by an aminoacyl-tRNA synthetase to form an aminoacyl-adenylate intermediate, before its attachment to the 3′ end of the tRNA. AMP, adenosine monophosphate; PPi, inorganic pyrophosphate.

Advanced concept box Fidelity of translation

Aminoacyl-tRNA synthetases have proofreading

ability.

To guarantee the accuracy of protein synthesis, mechanisms have evolved to ensure selection of the correct amino acid for acylation and for proofreading of already charged tRNAs. One such mechanism is found in the enzymes responsible for attaching an amino acid to the correct tRNA. The aminoacyl-tRNA synthetases have the ability not only to discriminate between amino acids before they are attached to the appropriate tRNA but also to remove amino acids that are attached to the wrong tRNA. In addition, these enzymes must discriminate between the multitude of tRNAs and be able to pair the correct tRNA with the appropriate amino acid. These abilities exhibited by the synthetases are accomplished by a series of hydrogen bonding interactions between the enzyme and the amino acid and between the enzyme and the tRNA. These mechanisms combine to ensure the accurate transfer of information from mRNA to protein.

Some flexibility in base pairing occurs at the 3′ base of the mRNA codon

Interaction of the charged tRNA with its cognate codon is accomplished by association of the anticodon loop in tRNA with the codon in mRNA through hydrogen bonding of complementary base pairs (Fig. 34.2). The base-pairing rules are the same as those for DNA (Chapter 32), except at the third position, or 3′ base, of the codon. At this position, nonclassical base pairs can form between this nucleotide and the first, or 5′ base, of the anticodon. The so-called **wobble hypothesis** of codon–anticodon pairing allows a tRNA that has an anticodon that is not perfectly complementary to the mRNA codon to recognize the sequence and allow for the incorporation of the amino acid into the growing peptide chain. Thus, if a guanine residue is at the 5′ position of the anticodon, it can form a base pair with either a cytidine or a uridine residue in the 3′ position of the codon. If the modified adenosine residue, inosine, occurs at the 5′ position of the anticodon, it can form a base pair with uridine, adenosine, or even cytidine at the 3′ position of the codon (Table 34.3). This would allow a tRNA with the

anticodon GAG to decode the codons CUU and CUC, both of which code for leucine. The wobble provides a mechanism for dealing efficiently with the degeneracy of the genetic code since the degeneracy always occurs in the third residue of the codon.

Table 34.3

Base-pairing possibilities between the third position or the 3′ nucleotide of the mRNA codon and the first position or 5′ nucleotide of the tRNA anticodon

Codon, third or 3′ position (mRNA)	Anticodon, first or 5′ position (tRNA)
G	C
U	A
A or G	U
C or U	G
A or C or U	I

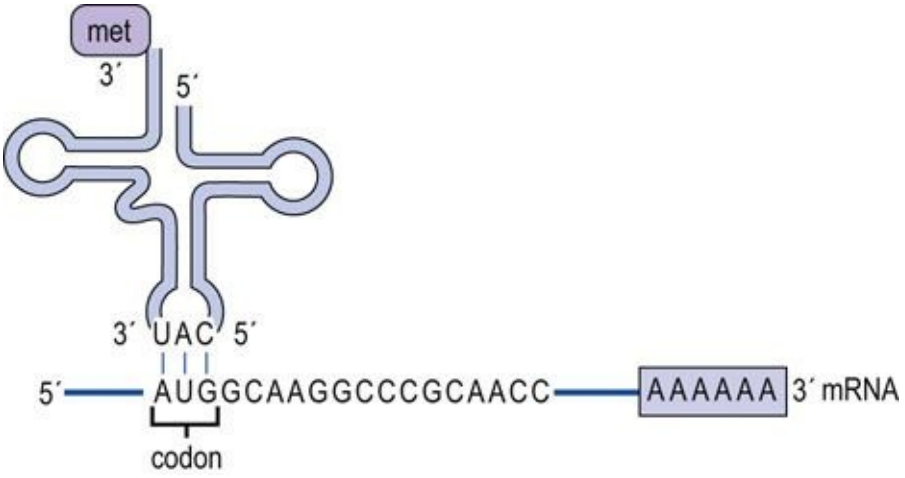


FIG. 34.2 Interaction of charged tRNA with mRNA.
The interaction of a charged tRNA with an mRNA occurs by base pairing of complementary bases in the anticodon loop and the codon of the mRNA.

How does the ribosome know where to begin protein synthesis?

The mRNA molecule carries the information that will be used to direct the synthesis of the protein. However, not all the information carried on the mRNA is used for this purpose. Most eukaryotic mRNAs contain regions both before

and after the protein-coding region, called 5′ and 3′ **flanking sequences** or 5′ and 3′ UTRs (untranslated regions). These sequences are involved in regulating the site and rate of protein synthesis and the stability of the mRNA. Thus, the protein coding region does not start immediately at the beginning of the mRNA, which raises the question of how the ribosome knows where to start synthesis. In the case of eukaryotic cells, the ribosome first binds to the 7-methylguanine 'cap' structure (Chapter 33) at the 5′ end of the mRNA, and then moves down the molecule until it encounters the first AUG codon (Fig. 34.3). This signals the ribosome to begin synthesizing the protein, beginning with a methionine residue, and to continue until it encounters one of the termination codons. On some viral and eukaryotic mRNAs, the first AUG is not used and, instead, an alternative start codon is defined by an internal ribosome entry site (IRES; see Fig. 34.3).

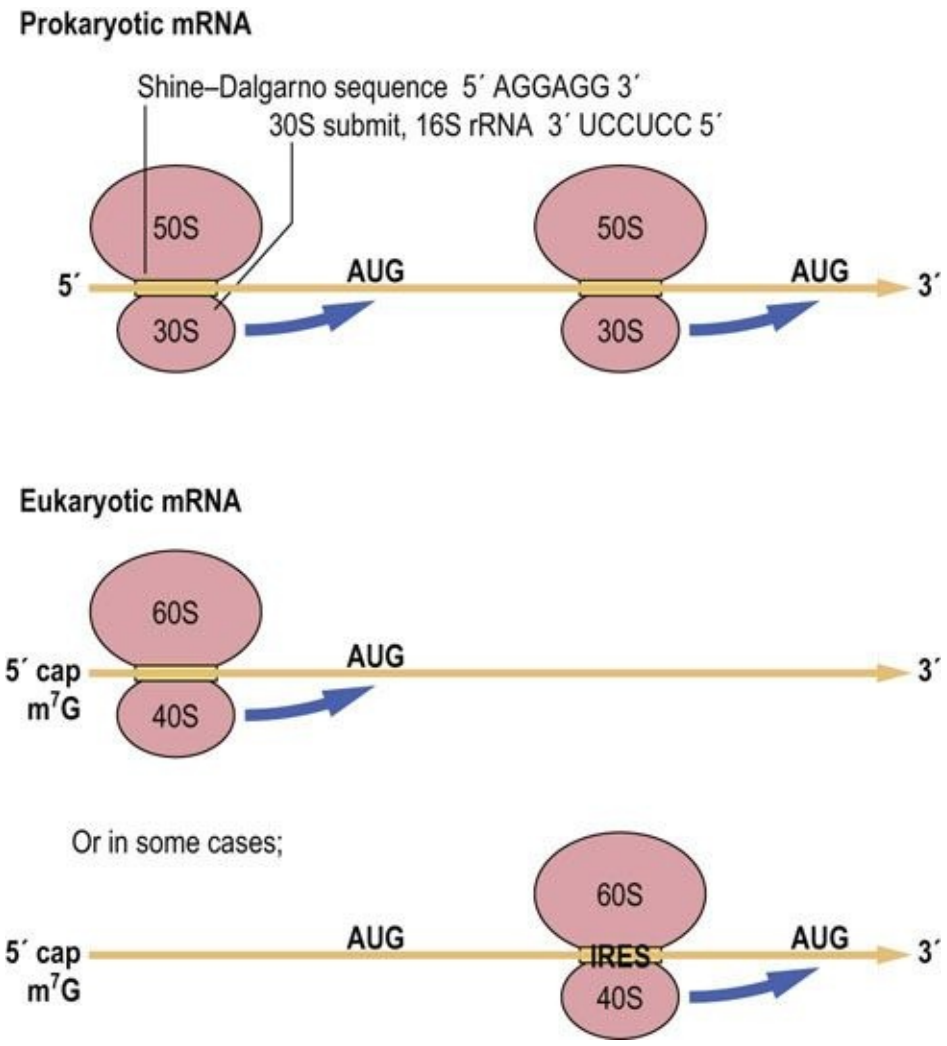


FIG. 34.3 Finding the protein-coding region.
The ribosome binds to the mRNA before locating the protein-coding region. Bacterial ribosomes (a portion of 16 S rRNA) bind to complementary sequences in the mRNA that are termed Shine–Dalgarno sequences, which are a short distance from the start of the protein-coding region. Eukaryotic ribosomes bind to the 5′ cap of mRNAs and then move down the mRNA until they encounter the first AUG codon or in a few cases they bind internally at an internal ribosome entry site (IRES) and then move to the AUG.

In the case of bacterial cells, there is no m^7G cap. Knowing what portion of the mRNA is to be used to synthesize a protein is complicated by the fact that there can be several proteins encoded by a single mRNA, each out of register with one another, so that proteins of different sequence may be obtained from the same ribonucleotide sequence. This problem has been solved by the discovery of a sequence in the mRNA that helps to precisely position the ribosome at the beginning of each protein-coding region. This sequence, known as the **Shine–Dalgarno sequence**, is found in most bacterial mRNAs and is complementary to a portion of the 16 S rRNA in the small bacterial ribosomal subunit (Fig. 34.3). Through the formation of hydrogen bonds, the ribosome is then positioned at the start of each protein-coding region.

The process of protein synthesis

Translation is a dynamic process that involves the interaction of enzymes, tRNAs, ribosomes and rRNAs, translation factors, and mRNA

Translation is normally divided into three steps:

- initiation
- elongation
- termination.

Initiation

Synthesis of a protein is initiated at the first AUG (methionine) codon in the mRNA

Initiation of protein synthesis in eukaryotes takes place when a dissociated small subunit of the ribosome forms a complex with eukaryotic initiation factors (eIF), eIF-1, eIF-1 A, and eIF-3, which binds to an eIF-2/Met-tRNA complex. This **preinitiation complex** is directed to the 5′ end of the mRNA by the binding of eIF-4F, and other factors, to the 5′ cap. The complex scans the mRNA until it locates the first AUG codon, using ATP to power this process. The large ribososmal subunit then binds to the small subunit/Met-tRNA/mRNA complex, and the Met-tRNA eventually is directed to the P site (Fig. 34.4), requiring eIF-5, hydrolyzing GTP and releasing initiation factors in the process. In eukaryotic cells, there are at least 12 different initiation factors. In prokaryotic cells, the process involves three initiation factors and the initiation complex first forms just 5′ to the coding region, as a result of the interaction of 16 S rRNA in the small subunit with the Shine–Dalgarno sequence on the mRNA. *N*-Formyl methionine (fmet), encoded by AUG, is the first amino acid in all bacterial proteins, instead of methionine.

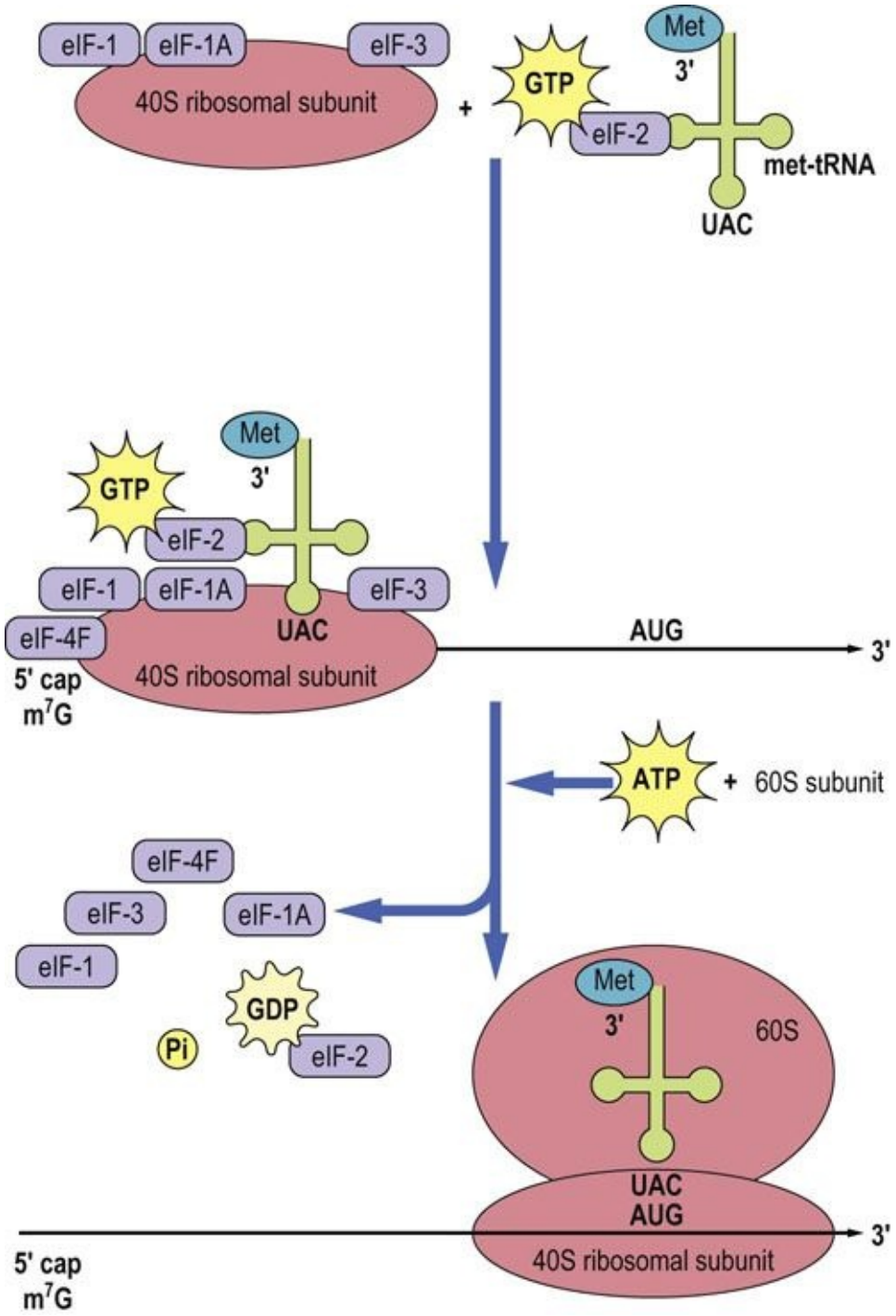


FIG. 34.4 Initiation of protein synthesis in eukaryotic cells.
The 40 S ribosomal subunit with bound initiation factors eIF-1, eIF-1A, and eIF-3, mRNA with eIF-4F bound to the 5′ cap, and met-tRNA bound to eIF-2 are brought together. Once these components are assembled, the complex translocates to the AUG, scanning the mRNA sequence and hydrolyzing ATP in the process. The 60 S ribosomal subunit completes the initiation complex and in the process the initiation factors are released. Note that, at initiation, the P site is occupied by the initiator met-tRNA, as shown in Fig. 34.5. GDP, guanosine diphosphate; eIF, eukaryotic initiation factor.

Elongation

Factors involved in the elongation stage of protein synthesis are targets of some antibiotics

After initiation is complete, the process of translating the information in mRNA into a functional protein starts. Elongation begins with the binding of a charged tRNA to the A site of the ribosome. In eukaryotic cells, the charged tRNA molecule is brought to the ribosome by the action of an elongation factor called eEF-1A (Fig. 34.5). For eEF-1A to be active, it must have a GTP molecule associated with it. If the charged tRNA is correct, one in which the anticodon of the tRNA forms base pairs with the codon on the mRNA, then GTP is hydrolyzed and eEF-1A is released. For the eEF-1A factor to bring another charged tRNA molecule to the ribosome, it must be regenerated by an elongation factor called eEF-1B which will promote the association of eEF-1A with GTP so that it may bind to another charged tRNA molecule (see Fig. 34.5). Once the correct charged tRNA molecule has been delivered to the A site of the ribosome, the **peptidyl transferase** activity of the ribosome catalyzes the formation of a peptide bond between the amino acid in the A site and the amino acid at the end of the growing peptide chain in the P site. The tRNA-peptide chain is now transiently bound to the A site. The ribosome is then moved one codon down the mRNA (towards the 3′ end), with the help of a factor known as eEF-2, and the tRNA in the A site, with the nascent peptide chain attached, moves to the P site. The uncharged tRNA originally in the P site moves to the E site so that there is a total of nine nucleotide pairs involved in stabilization of the ribosome–mRNA–tRNA complex. The whole process recycles for addition of the next amino acid (Fig. 34.6). The mechanics of this complex process are identical in prokaryotic cells, but the ribosomes and factors are different and this helps to explain the utility of antibiotics that preferentially inhibit protein synthesis in bacteria (Table 34.4).

Table 34.4

Selected antibiotics that affect protein synthesis

Antibiotic	Target
Tetracycline	Bacterial ribosome A site

Streptomycin	Bacterial 30 S ribosome subunit
Erythromycin	Bacterial 50 S ribosome subunit
Chloramphenicol	Bacterial ribosomepeptidyl transferase
Puromycin	Causes premature termination
Cycloheximide	Eukaryotic 80 S ribosome

Note that cycloheximide is toxic to humans.

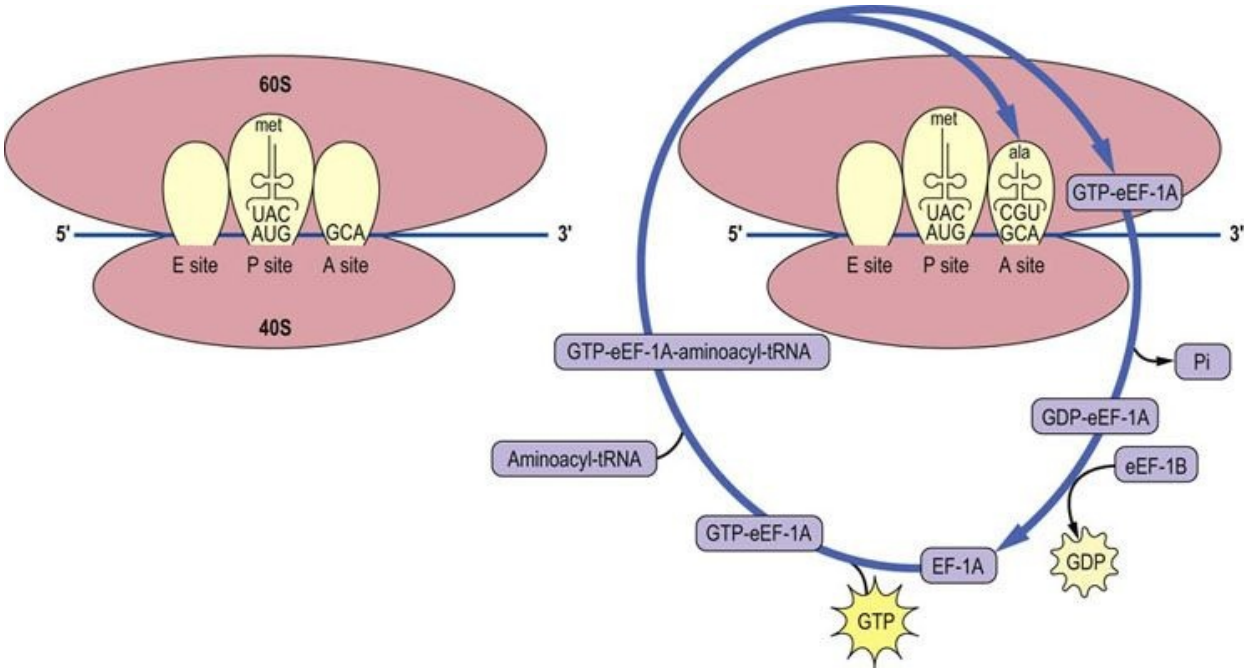


FIG. 34.5 Recycling of elongation factor eEF-1A.
A charged tRNA molecule is brought to the A site of the initiation complex, with the aid of eEF-1A with bound GTP, to begin the process of elongation. The factor is released once GTP is hydrolyzed and the process of recycling eEF-1A is aided by the exchange factor eEF-1B. Each successive amino acid addition requires that the correctly charged tRNA molecule be brought to the A site of the ribosome. ala, alanine; Pi, inorganic phosphate.

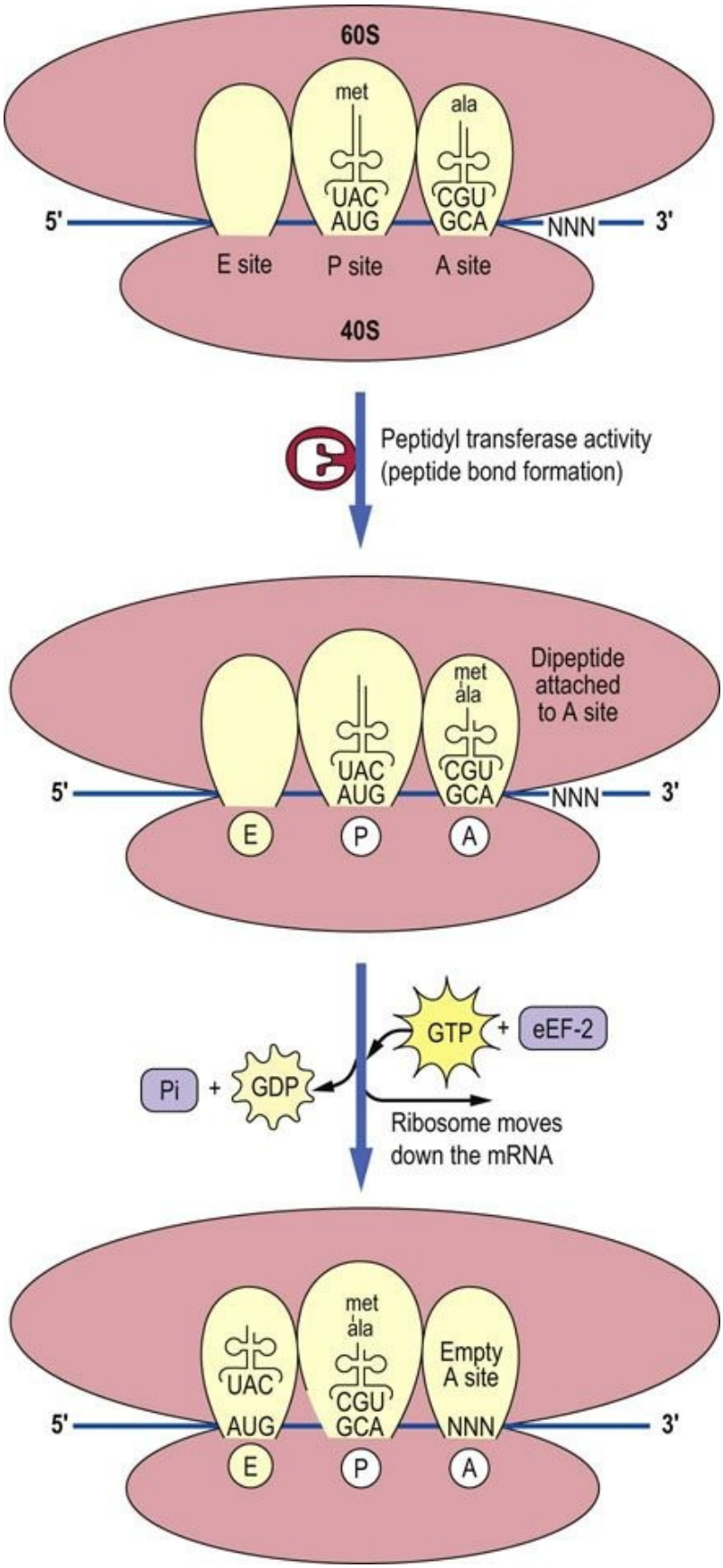

60S
met
ala
UAC
AUG
CGU
GCA
5'
NNN
3'
E site
P site
A site
40S
Peptidyl transferase activity
(peptide bond formation)
Dipeptide
attached
to A site
met
ala
UAC
AUG
CGU
GCA
5'
NNN
3'
E
P
A
GTP
+
eEF-2
Pi
+
GDP
Ribosome moves
down the mRNA
met
ala
UAC
AUG
CGU
GCA
Empty
A site
5'
NNN
3'
E
P
A

FIG. 34.6 Peptide bond formation and translocation.
The formation of the peptide bond between each successive amino acid is catalyzed by peptidyl transferase. Once the peptide bond is formed, an elongation factor (eEF-2 in this case) will move the ribosome down one codon on the mRNA, so that the A site is vacant and ready to receive the next charged tRNA. The E site is now occupied by the uncharged tRNA (met). NNN is the codon for next amino acid.

Termination

Termination of protein synthesis in both eukaryotic and bacterial cells is accomplished when the A site of the ribosome reaches one of the stop codons of the mRNA. Proteins called releasing factors recognize these codons, and cause the protein that is attached to the last tRNA molecule in the P site to be released (Fig. 34.7). This process is an energy-dependent reaction catalyzed by the hydrolysis of GTP, which transfers a water molecule to the end of the protein, thus releasing it from the tRNA. After release of the newly synthesized protein, the ribosomal subunits, tRNA, and mRNA dissociate from each other, setting the stage for the translation of another mRNA.

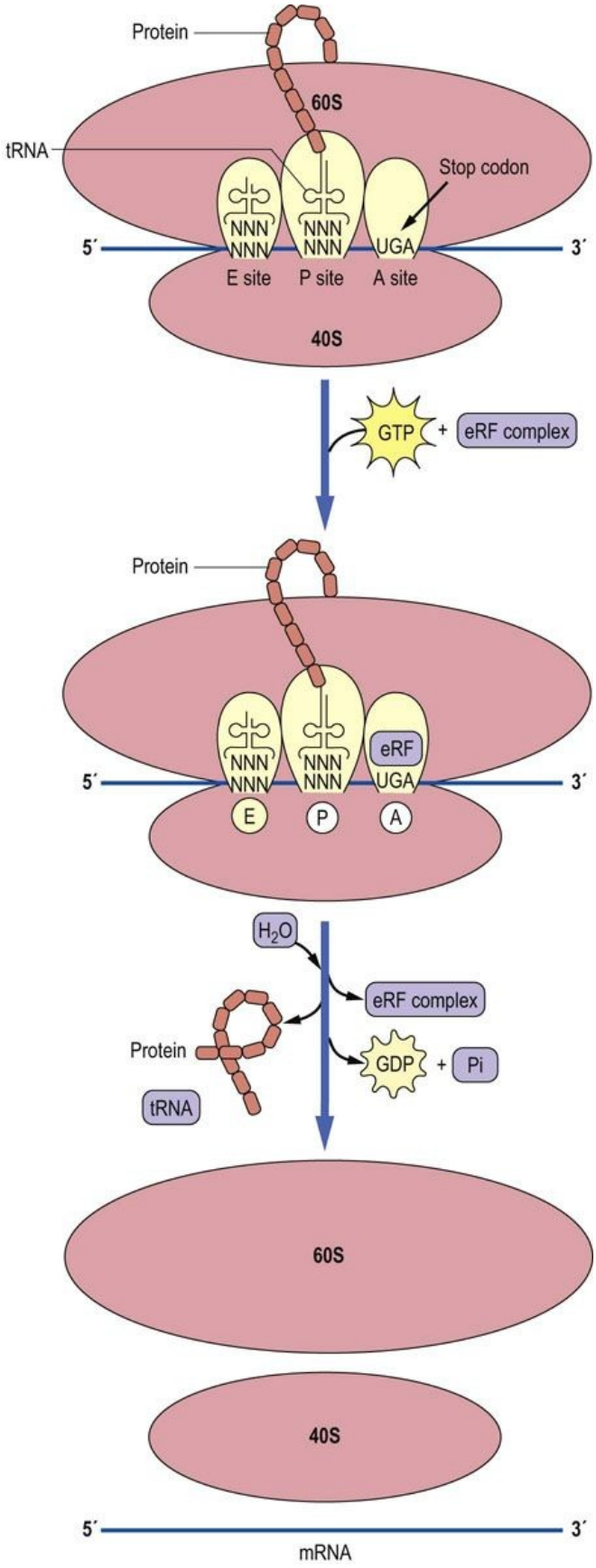

Protein
60S
tRNA
Stop codon
NNN
NNN
NNN
NNN
UGA
5′
3′
E site
P site
A site
40S
GTP
+
eRF complex
Protein
NNN
NNN
NNN
NNN
eRF
UGA
5′
3′
E
P
A
H_2O
eRF complex
Protein
GDP
+
Pi
tRNA
60S
40S
5′
3′
mRNA

FIG. 34.7 Termination of protein synthesis.
Termination of protein synthesis occurs when the A site is located over a termination codon. A releasing factor complex (eRF), with eRF-1 in the A site, will cause the completed protein to be released and the ribosome, mRNA, and tRNA will dissociate from each other to begin another cycle of translation.

Clinical box A noncompliant patient who was prescribed an antibiotic

A young man you were treating for a sinus infection returns to your clinic after 1 week, still complaining of sinus headaches and stuffiness. He explains that he began to feel better about 3 days after starting to take the antibiotic tetracycline, which you had prescribed. You inquire whether he continued to take the full dose of the drug, even after he began to feel better. He reluctantly admits that, as soon as he felt better, he stopped taking the drug. How do you explain to your patient that it is important that he takes the drug for as long as you prescribed it, even if he feels better after only a few days?

Comment.

As a physician, you know that tetracycline is a broad-spectrum antibiotic that inhibits the protein synthetic machinery of the bacterial cell by binding to the A site of the ribosome (Table 34.4). You also know that, if the drug is removed, protein synthesis can resume. If the drug is not taken for the entire period recommended, bacteria may begin to grow again, leading to the resurgence of the infection. Further, those bacteria that begin to grow after early termination of treatment are likely to be the most resistant to the drug. Because of the selection for more resistant mutant strains, the secondary infection is likely to be more difficult to control.

Advanced concept box Protein synthesis: peptidyl transferase

Peptidyl transferase is not your typical enzyme. It is a ribozyme.

Peptidyl transferase is the activity responsible for peptide bond formation during protein synthesis. This enzyme activity catalyzes the reaction between the amino group of the aminoacyl-tRNA in the A site and the carboxyl carbon of the peptidyl-tRNA in the P site, forming a peptide bond from an ester bond. The activity is located in the ribosome, but none of the ribosomal proteins has the capacity to catalyze this reaction. Crystal structures of the ribosome have shown that the peptidyl-transferase center is composed entirely of rRNA, and although proteins or specific amino acids are important for positioning the tRNAs or in stabilizing rRNA structure, the catalytic activity resides in the rRNA.

Protein folding and endoplasmic reticulum (ER) stress

ER stress, the result of errors in protein folding, develops in many chronic conditions, including obesity, diabetes and cancer

For a newly synthesized protein to become functionally active, it must be folded into a unique three-dimensional structure. Since there are too many random conformations that the newly synthesized proteins could possibly adopt, the proteins achieve their native structures with the help of a class of proteins called **chaperones**. Chaperones promote the correct folding, assembly and organization of proteins and macromolecular structures, including the nucleosome and electron transport complexes. In the ER they bind to exposed hydrophobic regions of unfolded proteins, preventing the misfolding of proteins and formation of nonspecific aggregates by shielding the interactive surfaces. The chaperones promote the correct folding of newly synthesized proteins by cycles of protein (substrate) binding and release regulated by an ATPase activity and by cofactor proteins. **Heat shock proteins** are a group of chaperone proteins that are expressed by cells in response to high temperature; they assist in the refolding of denatured proteins, not only as a result of heat but also in response to physical and chemical stresses. A protein, **GRP78/BiP** (glucose regulated protein 78/binding immunoglobulin protein) binds and traps misfolded proteins in the ER, preventing their further transport and secretion. The proteins are then routed to the **ER-associated degradation (ERAD) pathway**, which facilitates export to the cytosol and proteasomal (see below) degradation of the misfolded protein.

A condition known as ER stress develops if chaperone and ERAD activity is overwhelmed and protein aggregates accumulate in the lumen of the ER (e.g. as a result of a protein mutation, a glycosyl transferase deficiency or inhibitor, such as tunicamycin; Fig. 27.11). ER stress activates the **unfolded protein response (UPR)**, which induces higher expression of several chaperone and ERAD proteins. In addition, the protein PERK (protein kinase RNA-like ER kinase) oligomerizes and autophosphorylates, then phosphorylates eIF-2, which inhibits initiation (Fig. 34.4) and slows down the rate of synthesis of many proteins. If

homeostasis is not restored by the UPR, apoptosis is eventually initiated to eliminate the cell. A number of human diseases are characterized by ER stress and the UPR, including some forms of cystic fibrosis and retinitis pigmentosa. ER stress and the UPR also inhibit insulin signaling, causing insulin resistance, and play a role in development of pathology in obesity and type 2 diabetes mellitus (Chapter 21).

Protein targeting and post-translational modifications

Protein targeting

An mRNA can have several bound ribosomes at one time and this is known as a **polyribosome or polysome** (Fig. 34.8). There are two general classes of polysomes found in cells: those that are free in the cytoplasm and those that are attached to the endoplasmic reticulum. Those mRNAs encoding proteins destined for the cytoplasm or nucleus are translated primarily on polysomes free in the cytoplasm, while mRNAs encoding membrane and secreted proteins are translated on polysomes attached to the ER. Regions of the ER studded with bound ribosomes are described as the **rough endoplasmic reticulum (RER)**.

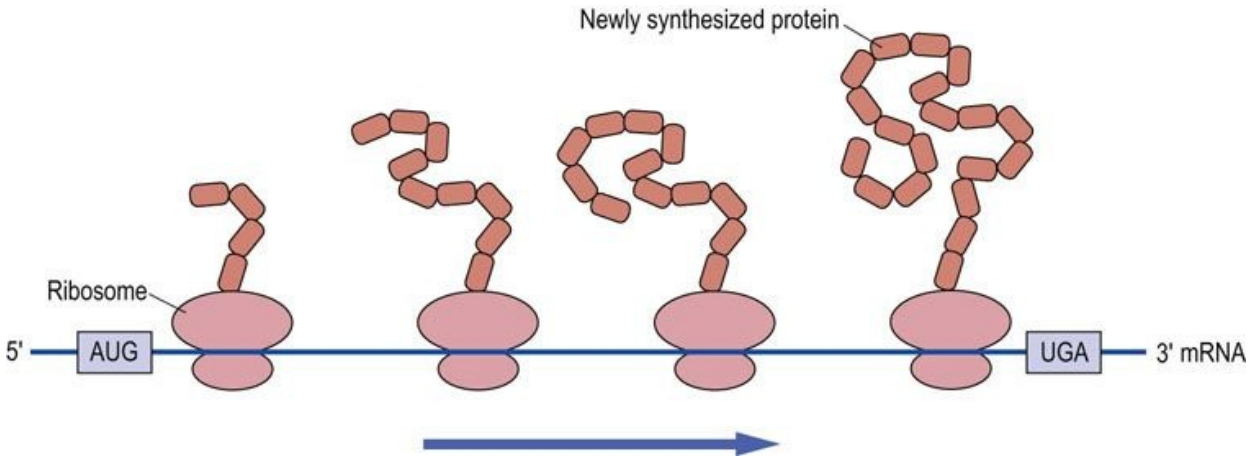


FIG. 34.8 Protein synthesis on polysomes.
Protein can be synthesized by several ribosomes bound to the same mRNA, forming a structure known as a polysome.

Cellular fate of proteins is determined by their signal peptide sequences

Proteins that are destined for export, for insertion into membranes, or for specific cellular organelles, must in some way be distinguished from proteins that reside in the cytoplasm. The distinguishing characteristic of proteins targeted for these

locations is that they contain a signal sequence usually comprising the first 20–30 amino acids on the amino-terminal end of the protein. In the case of secretory or membrane proteins, shortly after the signal sequence is synthesized, it is recognized by a ribonucleoprotein complex known as the **signal recognition particle (SRP)**, composed of a small RNA and six proteins. The SRP binds to the signal sequence and halts translation of the remainder of the protein. This complex then binds to the SRP receptor located on the membrane of the endoplasmic reticulum. After the SRP has delivered the ribosome-bound mRNA with its nascent protein to the endoplasmic reticulum, the signal sequence is inserted through the membrane, the SRP dissociates, and translation continues with the polypeptide chain being moved, as it is synthesized, across the membrane into the interstitial space of the endoplasmic reticulum (Fig. 34.9). The protein is then transferred to the Golgi apparatus and then to its final destination.

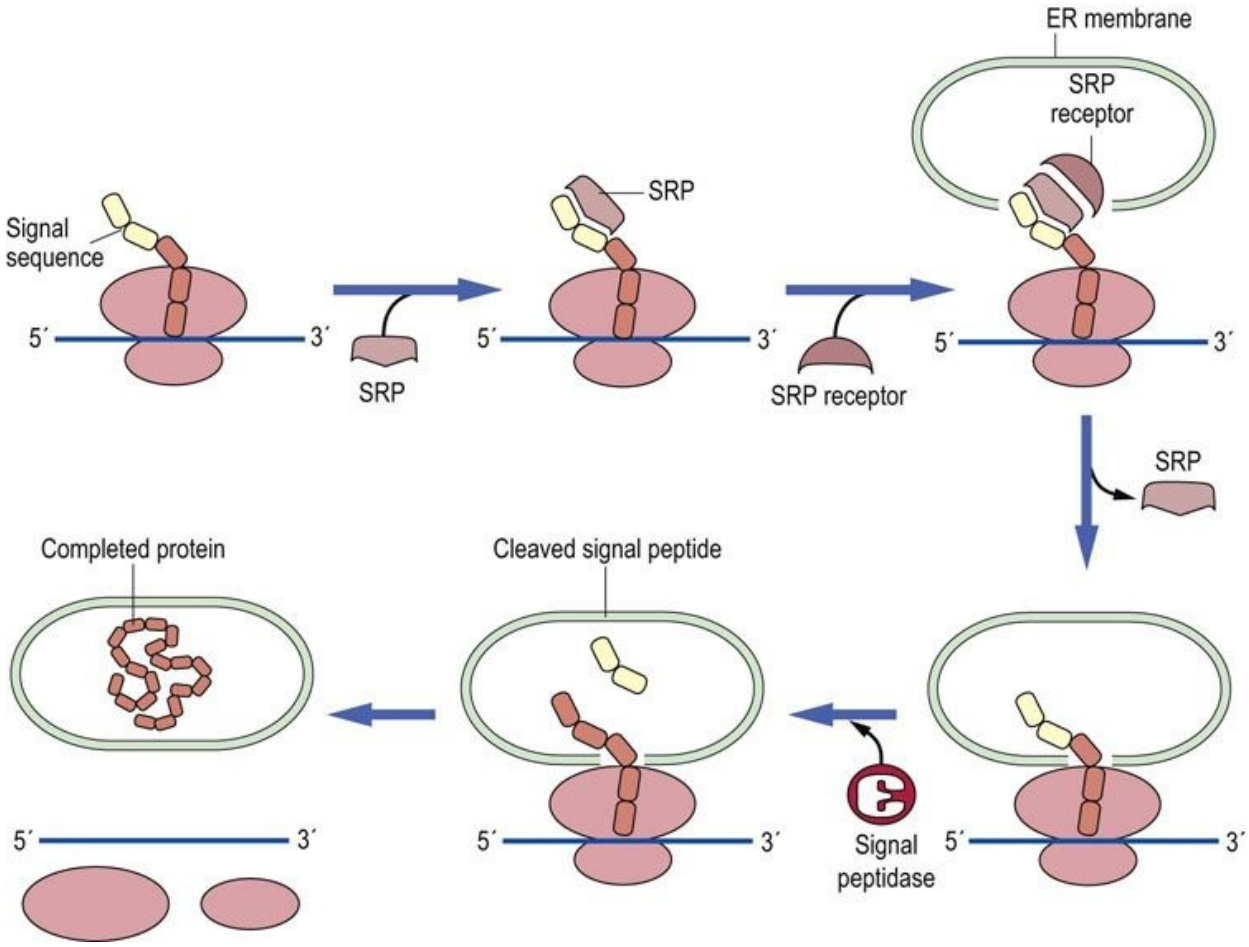


FIG. 34.9 Protein synthesis on the endoplasmic reticulum.
The signal sequence of the protein being translated is bound to an SRP and this complex is recognized by the SRP docking protein (SRP receptor) on the endoplasmic reticulum (ER), where the signal sequence is inserted through the membrane. A signal peptidase typically removes the signal sequence but not on every protein. Once synthesis of the protein is complete, the protein is inserted into the membranes or secreted.

While Man-6-P in high-mannose oligosaccharides is the signal that directs some glycoproteins to the lysosomal compartment (Fig. 27.14), targeting of proteins to other subcellular compartments is directed by a signal peptide sequence in the peptide chain. In contrast to secretory and membrane proteins, mitochondrial and nucleoplasmic proteins are transported after their translation is complete. In the case of proteins destined for the mitochondrion, they may have two signal sequences on their *N*-terminal end, depending on whether they are destined for the matrix or the intermembrane space. Mitochondrial proteins must be unfolded before they can be transported through transporters in the inner and outer membranes (TIM and TOM; Fig. 9.3). In contrast, nuclear proteins may have nuclear localization signals anywhere in the protein sequence, but exposed on the protein surface, and do not have to be unfolded before transport. Nuclear pores are very large complexes that can accommodate the recognition and transport of a protein in its native state.

Post-translational modification

Most proteins require post-translational modification before they become biologically active

The endoplasmic reticulum and Golgi apparatus are major sites of post-translational modification of proteins. In the endoplasmic reticulum, an enzyme called **signal peptidase** removes the signal sequence from the amino-terminus of the protein, resulting in a mature protein that is 20–30 amino acids shorter than that encoded by the mRNA. In the endoplasmic reticulum and Golgi apparatus, carbohydrate side chains are added and modified at specific sites on the protein (Chapter 27). One of the common amino-terminal modifications of eukaryotic cells is the removal of the amino-terminal methionine residue that initiates protein synthesis. Finally, many proteins, *e.g.* the hormones insulin and glucagon, are synthesized as preproproteins and proproteins that must be proteolytically cleaved for them to be active. The cleavage of a precursor to its

biologically active form is usually accomplished by a specific protease, and is a regulated cellular event.

Proteasomes: cellular machinery for protein turnover

Unlike DNA, proteins are not repaired, but degraded

Protein degradation is a complex process that is critical to cellular regulation. There are a number of reasons why a protein would need to be degraded. The protein is just worn out – it has aged passively by gradual denaturation during normal environmental stress. Proteins might also be modified by reaction with reactive intracellular compounds, such as glycolytic intermediates, or by reaction with products of lipid peroxidation during oxidative stress (Chapter 37). Other proteins might be part of cellular responses to hormones or perhaps are components of the cell cycle or are transcription factors; they need to be removed, sometimes rapidly, to attenuate the response or signal. Some of these latter proteins have characteristic amino acid sequences or *N*-terminal residues that promote their rapid turnover. The **PEST (ProGluSerThr)** sequence marks some proteins for rapid turnover, while proteins with *N*-terminal arginine generally have short half-lives, compared to proteins with *N*-terminal methionine.

The proteasome is a multicatalytic, complex for designed for degradation of cytosolic proteins

Since protein degradation is a destructive process, it must be sequestered within specific cellular organelles. Lysosomes, for example, ingest and degrade damaged mitochondria and other membranous organelles. However, most soluble, cytoplasmic proteins are degraded in structures called proteasomes. The 26 S proteasome (Fig. 34.10) consists of two types of subunits: a 20 S multimeric, multicatalytic protease (MCP) and a 19 S ATPase. The proteasome is a barrel-shaped structure, formed by a stack of four rings of seven homologous monomers, α-type subunits in the outer ends and β-type subunits on the inner rings of the barrel. The proteolytic activity – three different types of threonine proteases – resides on β-subunits with active sites facing the inside of the barrel, thereby protecting cytoplasmic proteins from inappropriate degradation. The

ATPase subunits are attached at either end of the barrel and act as gatekeepers, allowing only proteins destined for destruction to enter the barrel. The proteins are unfolded in a process that requires ATP and degraded by the protease activities to small peptides, 6–9 amino acids in length, that are released into the cytoplasm for further degradation.

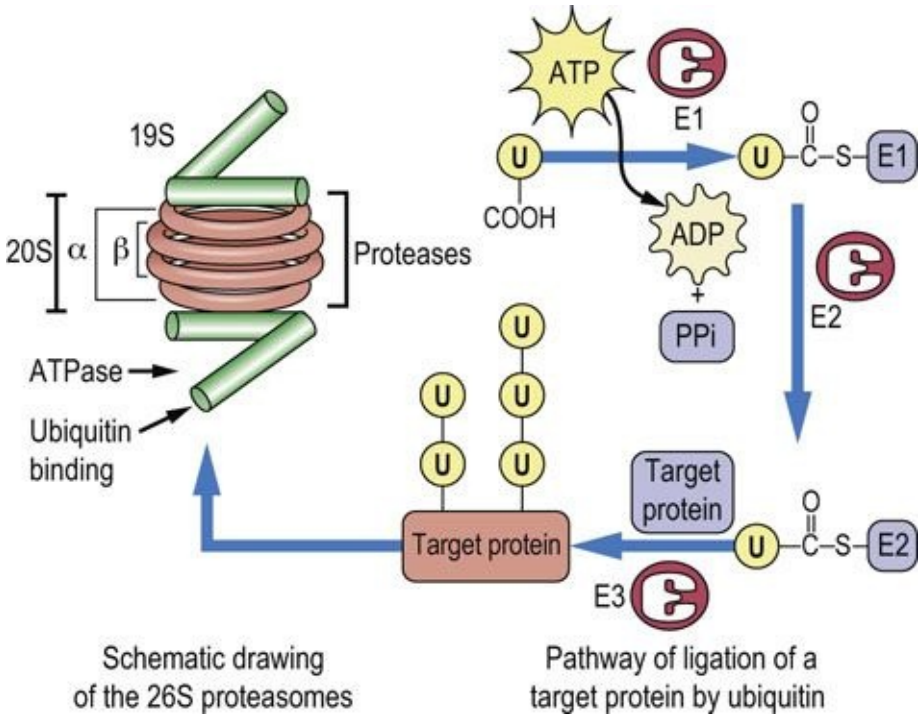


FIG. 34.10 Structure of the proteasome and the role of ubiquitin (U) in protein turnover. The proteasome is shown at the left, a barrel-shaped structure. The 20 S multicatalytic protease in the middle rings of the barrel has protease activity on the inner face. The 19 S caps at the end of the barrel have ubiquitin binding and release and ATPase activity, and control access of proteins to the inside of the barrel for degradation. The ubiquitin cycle is involved in marking proteins for degradation by the proteasome. Ubiquitin is first activated as a thioester derivative of ubiquitin-activating enzyme E1; it is then transferred to ubiquitin-conjugating enzyme E2, then to a lysine residue on the target protein, catalyzed by ubiquitin-ligase E3. Ubiquitin is frequently copolymerized on target proteins. The more highly ubiquinated the protein, the more susceptible it is to proteasomal degradation. Note that the drawing is not to scale; the proteasome is a 26 S macromolecular complex (≥2,000,000 Da); target proteins are smaller, while ubiquitin is less than 10,000 Da.

Ubiquitin targets proteins to the proteasome for

degradation

Proteins destined for destruction are directed to the proteasome primarily by covalent modification with a very highly conserved, 76-amino acid residue protein called ubiquitin**,** which is found in all cells. Ubiquitin must be activated to fulfill its role (see Fig. 34.10); this is accomplished by a ubiquitin-activating enzyme called E1. Activation occurs when E1 is attached via a thioester bond to the *C*-terminus of ubiquitin by ATP-driven formation of an ubiquitin–adenylate intermediate. The activated ubiquitin is then attached to a ubiquitin carrier protein, known as E2, by a thioester linkage. A ubiquitin protein ligase known as E3 transfers the ubiquitin from E2 to a target protein, forming an **isopeptide bond** between the carboxyl-terminus of ubiquitin and the ε-amino group of a lysine residue on a target protein. Denatured and oxidized proteins are commonly ubiquitinated by this mechanism. The 19 S subunit of the proteasome has a ubiquitin-binding site that allows proteins with a covalently attached ubiquitin protein to enter the barrel; the ubiquitin is then released by a ubiquitinase activity and recycled to the cytosol for reuse. Polymerization of ubiquitin on target proteins (**polyubiquitination**; Fig. 34.10) significantly enhances the degradation of the protein.

The ubiquitin pathway leading to proteasomal degradation of proteins is complex. Although the number of E1 enzymes is typically small, there are several E2 and E3 proteins with different target specificities, and there are six different ATPase activities associated with the 19 S proteasome subunit. The ATPases are thought to be involved in denaturation of ubiquinated proteins, opening the proteolytic core, and transporting the peptide chain into the core of the proteasome. All these variations in components of the pathway, as well as changes during the cell cycle and in response to hormonal stimulation, provide a flexible and regulated pathway for protein turnover.

Advanced concept box Inhibiting the proteasome to treat cancer

Multiple myeloma is a cancer of plasma cells (B lymphocytes), which are normally found in bone and synthesize antibodies as part of the immune system (see Chapter 2). The unchecked growth of these plasma cells results in anemia, tumors in the bones, a compromised immune response, and unfortunately usually a poor prognosis. Patients with recurring multiple myeloma have few

treatment options, but a new drug that is an inhibitor of protease activity in the proteasome, bortezomib, has been added to the arsenal. It has been shown to improve the chances for survival of these patients and increase the length of time before remission, especially when combined with other therapies such as radiation or additional chemotherapy. Bortezomib inhibits the degradation of proteins involved in programmed cell death (apoptosis) of cancer cells, thereby enhancing the self-destruct signal in myeloma cells. Second-generation proteosome inhibitors targeting different parts of the complex are currently in clinical trials.

Summary

■ Protein synthesis is the culmination of the transfer of genetic information from DNA to proteins. In this transfer, information must be translated from the four-nucleotide language of DNA and RNA to the 20-amino acid language of proteins.
■ The genetic code, in which three nucleotides in mRNA (codon) specify an amino acid, represents the translation dictionary of the two languages.
■ The tRNA molecule is the bridge between the sequence of the nucleotides in mRNA and the amino acids protein. It tRNA accomplishes this task by virtue of its anticodon loop, which interacts with specific codons on the mRNA and also with amino acids via its amino acid attachment site located on the 3′ end of the molecule.
■ The process of translation consists of three parts: initiation, elongation, and termination.
■ Initiation involves the assembly of the ribosome and charged tRNA at the initiation codon (AUG) of the mRNA. This assembly process is mediated by initiation factors and requires the expenditure of energy in the form of GTP.
■ Elongation is a stepwise addition of the individual amino acids to a growing peptide chain by the action of the ribozyme, peptidyl transferase. The charged tRNA molecules are brought to the ribosome by elongation factors at the expense of GTP hydrolysis.
■ Termination of protein synthesis occurs when the ribosome reaches a stop codon and releasing factors catalyze the release of a protein.
■ After release, the newly synthesized protein must be correctly folded with the help of ancillary proteins called chaperones, and targeted to specific subcellular compartments by signal sequences.
■ Many newly synthesized proteins must also be modified, by a variety of chemical and structural changes, before they are biologically active.

Active learning

1. Review the mechanism of action of various drugs that inhibit protein synthesis on the bacterial ribosome.
2. List the mechanisms that assure the fidelity of DNA and protein synthesis.
3. Describe the signal sequences that target proteins to the

lysosome, the mitochondria or the nucleus.
4. Discuss the role of the *N*-terminal amino acid as a factor regulating the rate of turnover of a cytoplasmic protein.
5. Explain how viruses take control of the cellular protein translation machinery during viral infections to favor the synthesis of viral proteins.

Further reading

Back, SH., Kaufman, RJ. Endoplasmic stress and type 2 diabetes. *Annu Rev Biochem*. 2012; 81:767–793.

Garcia-Robles, I, Sanchez-Navarro, J, de la Pena, M. Intronic hammerhead ribozymes in mRNA biogenesis. *Biol Chem*. 2012; 393:1317–1326.

Grudnik, P, Bange, G, Sinning, I. Protein targeting by the signal recognition particle. *Biol. Chem*. 2009; 390:775–782.

Korostelev, A, Noller, HF. The ribosome in focus: new structures bring new insights. *Trends Biochem Sci*. 2007; 32:434–441.

Rodnina, MV, Beringer, M, Wintermeyer, W. How ribosomes make peptide bonds. *Trends Biochem Sci*. 2007; 32:20–26.

Schmeing, TM, Ramakrishnan, V. What recent ribosome structures have revealed about the mechanism of translation. *Nature*. 2009; 461:1234–1242.

Walter, P, Ron, D. The unfolded protein response: from stress pathway to homeostatic regulation. *Science*. 2011; 334:1081–1086.

Wang, S, Kaufman, RJ. The impact of the unfolded protein response on human disease. *J Cell Biol*. 2012; 197:857–867.

Websites

Genetic code. www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi?mode = c.

IRES. www.iresite.org.

Proteasome:

http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/P/Proteasome.html

Ubiquitin. www.rcsb.org/pdb/101/motm.do?momID=60.

Ribosome:

www.mrc-lmb.cam.ac.uk/ribo/homepage/mov_and_overview.html

www.weizmann.ac.il/sb/faculty_pages/Yonath/home.html

CHAPTER 35

Regulation of Gene Expression

Basic Mechanisms

Jeffrey R. Patton, D. Margaret Hunt and Andrew Jamieson

Learning objectives

After reading this chapter you should be able to:

■ Describe the general mechanisms of regulation of gene expression, with an emphasis on initiation of transcription.

■ Describe the many levels at which gene expression may be controlled, using steroid-induced gene expression as a model.

■ Explain how alternative mRNA splicing, alternate promoters for the start of mRNA synthesis, post-transcriptional editing of the mRNA, and the inhibition of protein synthesis by small RNAs, can modulate expression of a gene.

■ Explain how the structure and packaging of chromatin can affect gene expression.

■ Explain how genomic imprinting affects gene expression, depending on whether alleles are maternally or paternally inherited.

Introduction

Despite identical DNA is all cells, gene expressions varies significantly with sex, and in time and place in the body

The study of genes and the mechanism whereby the information they hold is converted into proteins and enzymes, hormones, and intracellular signaling molecules is the realm of molecular biology. Except for the erythrocyte, all cells in the body have the same DNA complement. One of the most fascinating aspects of this work is the study of the mechanisms that control differential gene expression, both in time and in place, and the consequences if these control mechanisms are disrupted.

The goal of this chapter is to introduce the basic concepts involved in regulation of protein-coding genes and how these processes are involved in the causation of human disease. The basic mechanism of gene regulation will be described first, followed by a discussion of a specific gene regulation system to highlight various aspects of the basic mechanism. The chapter will end with a discussion of various ways in which the gene regulatory apparatus can be adapted to suit the needs of different tissues and situations.

Basic mechanisms of gene expression

Gene expression is regulated at several different levels

The control of gene expression in humans occurs principally at the level of transcription, the synthesis of mRNA. However, transcription is just the first step in the conversion of the genetic information encoded by a gene into the final processed gene product, and it has become increasingly clear that post-transcriptional events allow for exquisite control of gene expression. The sequence of events involved in the ultimate expression of a particular gene may be summarized as:

initiation of transcription
→ processing the transcript
→ transport to cytoplasm
→ translation of transcript into protein
→ post-translational processing of the protein

At each of these steps conditions allow for the cell to either proceed to the next step or attenuate or halt the process. For instance, if the processing of the RNA is not correct or complete, the resulting mRNA would be useless or possibly destroyed. In addition, if the mRNA is not transported out of the nucleus, it will not be translated. Clearly, during the growth of a human embryo from a single fertilized ovum to a newborn infant, there must be numerous changes in the regulation of genes, to allow the differentiation of a single cell into many types of cells that develop tissue-specific characteristics. Similarly, at puberty there are changes in the secretion of pituitary hormones that result in the cyclic secretion of ovarian and adrenal hormones in females and the production of secondary sexual characteristics. Such programmed events are common in all cellular organisms and the production of these phenotypic changes in cells – and thus the whole organism – arises as a result of changes in the expression of key genes. The expression of genes essential for these processes varies depending on the cell type and the stage of development, but the mechanisms underlying the changes are available to basically all cells. In humans and most other eukaryotes, mechanisms that regulate gene expression are numerous; some of the requirements and options available at each stage are outlined in Table 35.1.

Table 35.1

Requirements and options in the control of gene expression

Process	Requirements	Options
Transcription of mRNA	Chromatin is relaxed (condensed chromatin is a poor template)	Allele-specific transcription
	DNA in hypomethylated state (methylation of promoter inhibits transcription)	Selection of alternative promoters giving different start sites
	Correct *trans*-acting factors are present (such as transcription factors and cofactors)	
Processing of mRNA	mRNA is 5′ capped	
	PolyA is added to 3′ end of most messages	Many transcripts are alternatively spliced, increasing coding potential
	For most mRNAs the transcript is spliced	mRNAs can be edited to change the coding sequence, changing an amino acid or creating a stop codon
		Signals in the 3′ untranslated regions (UTR) of mRNAs can stabilize or mark the RNA for destruction
Translation of mRNA	mRNA must be transported to cytoplasm	mRNA can be localized in specific regions of the cytoplasm, such as the ends of axons, for local translation
	All the factors needed for protein synthesis	Alternative start codons due to internal ribosome entry site (IRES)
		Translation on free ribosomes or rough endoplasmic reticulum
		miRNAs can inhibit translation
Turnover of proteins	Unique protein half-life	Structural proteins tend to turn over slowly collagens
		Cell cycle proteins are quickly turned over to limit mitosis
		Some proteins contains sequences that target them for rapid degradation

Gene transcription depends on key elements in the region of the gene

The key step in the transcription of a protein-coding gene is the conversion of the information held within the DNA of the gene into messenger RNA, which can then be used as a template for synthesis of the protein product of the gene. For expression of a gene to take place, the enzyme that catalyzes the formation of mRNA, RNA polymerase II (RNAPol II), must be able to recognize the so-called start points for transcription of the gene. RNAPol II uses one strand of the DNA template to create a new, complementary RNA start point (often called the **primary transcript or pre-mRNA**), which is then modified in various ways, (commonly including the addition of a **7-methylguanosine cap (m^7Gppp)** cap at the 5′ end and the polyA tail at the 3′ end, and the removal of introns to form a mature mRNA (Chapter 33). However, RNAPol II cannot initiate transcription alone; it requires other factors to assist in the recognition of critical gene sequences and other proteins to be bound in the vicinity of the start point for transcription.

Promoters

Promoters are usually upstream of the transcription start point of a gene

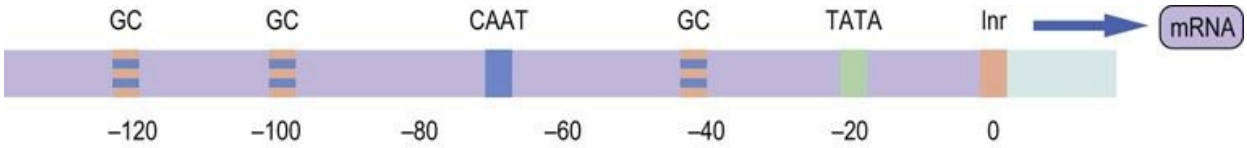
Sequences that are relatively close to the start of transcription of a gene and control its expression are collectively known as the promoter. Since this is usually within a few hundred or a few thousand nucleotides of the start point, it is usually referred to as the proximal promoter. The promoter sequence acts as a basic recognition unit, signaling that there is a gene that can be transcribed and providing the information needed for the RNAPol II to recognize the gene and to correctly initiate RNA synthesis, both at the right place and using the correct strand of DNA as template. The promoter also plays an important role in determining that the RNA is synthesized at the right time in the right cell. Most control regions in the promoter are upstream (5′) of the transcription start point, and therefore are not transcribed into RNA. Occasionally, some elements of the promoter may be downstream of the start point for RNA synthesis, and may actually be transcribed into RNA. The structure of promoters varies from gene to gene but there are a number of key sequence elements that can be identified within the promoter. These elements may be present in different combinations, some elements being present in one gene and absent in another. Sequences further away from the transcription start site, some of which are known as **enhancers**, may also have a major impact on the transcription of a gene. These elements are often said to be part of the distal promoter for the gene.

The efficiency and specificity of gene expression are conferred by *cis*-acting elements

A promoter exerts its effect because it is on the same piece of DNA as the gene being transcribed and is referred to as a ***cis*-acting sequence** or element to emphasize that it affects only the neighboring gene on the same chromosome. Since the promoter is critical to gene expression, it is often regarded as being part of the gene it controls, since without it the mRNA would not be made.

The nucleotide sequence immediately surrounding the start of transcription of a gene varies from gene to gene. However, the first nucleotide in the mRNA transcript tends to be adenosine, usually followed by a pyrimidine-rich sequence, termed the **initiator (Inr)**. In general, it has the nucleotide sequence Py_2CAPy_5 (Py-pyrimidine base) and is found between positions –3 to +5 in relation to the starting point. In addition to Inr, most promoters possess a sequence known as the **TATA box** approximately 25 bp upstream from the start of transcription.

The TATA box has an 8 bp consensus sequence that usually consists entirely of adenine-thymine (A-T) base pairs, although very rarely a guanine-cytosine (G-C) pair may be present. This sequence appears to be very important in the process of transcription, as nucleotide substitutions that disrupt the TATA box result in a marked reduction in the efficiency of transcription. The positions of Inr and the TATA box relative to the start are relatively fixed (Fig. 35.1). Having said this, it must be pointed out that there are many eukaryotic genes that do not have an identifiable TATA box and other sequences are essential for delineating the start of transcription.



	Consensus site	DNA binding	Factor bound
TATA	TATAAAA	~10 bp	TBP
CAAT	GGCCAATCT	~22 bp	NFI/CTF
GC box	GGGCGG	~20 bp	SP-1

FIG. 35.1 Idealized version of a promoter comprising various elements. Each promoter element has a specific consensus sequence that binds ubiquitous transcription-activating factors. Binding of transcription factors encompasses the consensus site and a variable number of anonymous adjacent nucleotides, depending on the promoter element. CTF, a member of a protein family whose members act as transcription factors; TBP, TATA-binding protein; NFI, nuclear factor I; SP-1, ubiquitous transcription factor.

In addition to the TATA box, other commonly found *cis*-acting promoter elements have been described. For example, the **CAAT box** is often found upstream of the TATA box, typically about 80 bp from the start of transcription. As in the case of the TATA box, it may be more important for its ability to increase the strength of the promoter signal rather than in controlling tissue-or time-specific expression of the gene. Another commonly noted promoter element is the GC **box**, a GC-rich sequence; multiple copies may be found in a single promoter region.

Figure 35.1 lists some of the common *cis*-acting elements seen within promoters. These promoter elements bind protein factors (transcription factors)

that recognize the DNA sequence of each particular element. Some transcription factors stimulate transcription, others suppress it; some are expressed ubiquitously, others are expressed in a tissue-or time-specific fashion. Thus the array of factors bound to a promoter region can vary from cell to cell, tissue to tissue, and be affected by the state of the organism. These factors, bound to promoter sequences, determine how actively the RNApol II copies the DNA into RNA.

Alternative promoters

Alternative promoters permit tissue or developmental stage-specific gene expression

Although it is clear that a promoter is essential for gene expression to occur, a single promoter may not possess the tissue specificity or developmental stage specificity to allow it to direct expression of a gene at every correct time and place. Some genes have evolved a series of promoters that confer tissue-specific expression. In addition to the use of different promoters that are physically separated, each of the alternative promoters is often associated with its own first exon and, as a result, each mRNA and subsequent protein has a tissue-specific 5′ end and amino acid sequence. A good example of the use of alternative promoters in humans is the gene for dystrophin, the muscle protein that is deficient in Duchenne muscular dystrophy (see Chapter 20). This gene uses alternative promoters that give rise to brain-, muscle-, and retinal-specific proteins, all with differing *N*-terminal amino acid sequences.

Advanced concept box Identifying the function and specificity of nucleotide sequences

Consensus sequences are nucleotide sequences that contain unique core elements that identify the function and specificity of the sequence, for example the TATA box. The sequence of the element may differ by a few nucleotides in different genes but a core, or consensus, sequence is always present. In general, the differences do not influence the effectiveness of the sequence. These consensus sequences are arrived at by comparing the promoters of the same genes from different species of eukaryotes,

by comparing the promoter sequences from genes that bind the same transcription factor, or by determining the actual sequence of DNA that serves as the binding element for the factor (see Fig. 35.1).

Enhancers

Enhancers modulate the strength of gene expression in a cell

Although the promoter is essential for initiation of transcription, it is not necessarily alone in influencing the strength of transcription of a particular gene. Another group of elements, known as enhancers, can regulate the level of transcription of a gene but, unlike promoters, their position may vary widely with respect to the start point of transcription and their orientation has no effect on their efficiency. Enhancers may lie upstream or downstream of a promoter and may be important in conferring tissue-specific transcription. For instance, a nonspecific promoter may initiate transcription only in the presence of a tissue-specific enhancer. Alternatively, a tissue-specific promoter may initiate transcription but with a greatly increased efficiency in the presence of a nearby enhancer that is not tissue specific. In some genes, for example immunoglobulin genes, enhancers may actually be present downstream of the start point of transcription, within an intron of the gene being actively transcribed.

Response elements

Response elements are binding sites for transcription factors and coordinately regulate expression of multiple genes, *e.g.* in response to hormonal or environmental stimuli

Response elements are nucleotide sequences that allow specific stimuli, such as steroid hormones (steroid response element; SRE), cyclic AMP (cyclic AMP response element; CRE), or insulin-like growth factor-1 (IGF-1, insulin response

element; IRE), to stimulate or repress gene expression. Response elements are often part of promoters or enhancers where they function as binding sites for particular transcription factors. Response elements in promoters are *cis*-acting sequences that are typically 6–12 bases in length. A single gene may possess a number of different response elements, possibly having transcription stimulated by one stimulus and inhibited by another. Multiple genes may possess the same response element, and this facilitates coinduction or corepression of groups of genes, such as in response to a hormonal stimulus.

Transcription factors

Transcription factors are DNA binding proteins that regulate gene expression

Promoters, enhancers and response elements are part of the gene; transcription factors are the proteins that recognize these structures. These sequence-specific DNA-binding proteins bind to specific nucleotide sequences and bring about differential expression of the gene during development and also within tissues of the mature organism (Fig. 35.2). Many transcription factors act positively and promote transcription, while others act negatively and promote gene silencing. The unique pattern of transcription factors present in the cell will determine in large part which portion of the genome is transcribed into RNA in any given cell. Transcription factors are sometimes referred to as **trans-acting factors** to emphasize that, as soluble proteins, they can diffuse within the nucleus and act on multiple different genes on different chromosomes.

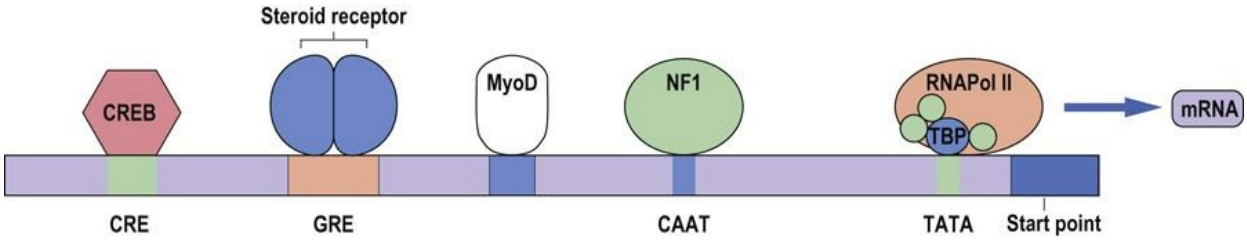


FIG. 35.2 Regulation of gene expression by specific regulatory elements. Binding of transcription factors to a steroid response element modulates the rate of transcription of the message. Different elements have varying effects on the level of transcription, some exerting greater effects than others, and may also activate tissue-specific expression. CRE, cyclic AMP response element; CREB, CRE-binding protein;

GRE, glucocorticoid response element; MyoD, muscle cell-specific transcription factor; NF1, nuclear factor 1. The proteins are shown in a linear array for convenience, but they interact physically with one another, both because of their size and the folding of DNA.

There are other kinds of proteins involved in regulating transcription besides the sequence-specific transcription factors. The so-called general transcription factors, such as TFIIA, B, D, E, etc., form a complex with RNAPol II and this complex is necessary for the initiation of transcription. These **general transcription factors** are needed for the successful use of every promoter; they vary somewhat with the class of gene, being generally different for RNA polymerase I, II, and III. In eukaryotic cells, and mammalian cells in particular, the RNA polymerases cannot recognize promoter sequences themselves. It is the task of the gene-specific factors to create a local environment that can successfully attract the general factors, which in turn attract the polymerase itself. However, there is emerging evidence that the RNA polymerase complex itself may also be important in the regulation of gene expression.

In addition, other proteins can bind to the sequence-specific transcription factors and modulate their function by repressing or activating gene expression; these factors are often called **coactivators** or **corepressors**. Thus, the overall rate of RNA transcription from a gene is the result of the complex interplay of a multitude of transcription factors, coactivators and corepressors. Since there are thousands of these factors in a cell, there is an almost unimaginably large number of combinations that can occur, and thus the control of gene expression can be very specific and very subtle.

In prokaryotes, the *cis*-elements that control the start site and, in general, the initiation of transcription, are placed closer to the starting point. These *cis*-elements are fewer in number and there is much less variety (Chapter 33), when compared with those from eukaryotes. In addition, there are fewer *trans*-acting factors and the control of gene expression is much less subtle in prokaryotes. However, understanding the limitations of the control of gene expression in prokaryotes allows one to appreciate the flexibility of the strategies for control found in eukaryotes.

Advanced concept box What is a ‘gene’? Transcription unit versus gene

Exactly what a ‘gene’ is has become increasingly difficult to define in recent years. The initial notion that a gene was a piece of DNA

that gave rise to a single gene product – one gene, one protein – has been challenged. It is now clear that many functional products – different mRNA species or different protein products – may arise from a single region of transcribed DNA, as a result of differences either at the level of transcription or at the post-transcriptional level. Thus there is now a tendency to refer to such 'genes' as transcription units. The transcription unit encapsulates not only those parts of the gene such as the promoters, exons, and introns, classically regarded as the gene unit, but also the molecular elements that modify the transcription process from the initiation of transcription to the final post-transcriptional modifications. This is a shift away from the notion of a gene being one strand of DNA with exons and introns, to one of a gene being a complex structure that directs a dynamic process, giving rise to the final gene product or products at various stages of development of an organism.

Initiation of transcription requires binding of transcription factors to DNA

For transcription to occur, transcription factors must bind to DNA. The protein known as **TATA-binding protein (TBP)** binds to the region of the TATA box. TBP is a general transcription factor, associated with the complex of RNAPol II and a variable number of other proteins. Binding of TBP to the TATA box directs the positioning of the transcription apparatus at a fixed distance from the start point of transcription and thus allows RNAPol II to be positioned exactly at the site of initiation of transcription. Once RNAPol II and a number of other transcription factors have bound to the region of the start point, transcription can occur. When transcription begins, many of the transcription factors required for binding and alignment of RNAPol II are released, and the polymerase travels along the DNA, forming the pre-mRNA transcript.

Transcription factors have highly conserved DNA binding sites

The binding of transcription factors to DNA involves a relatively small area of

the transcription factor protein, which comes into close contact with the major and/or minor groove of the DNA double helix to be transcribed. The regions of these proteins that contact the DNA are called DNA-binding domains or motifs, and are highly conserved between species. There are a variety of DNA-binding domains, some of which occur in multiple transcription factors or multiple times in the same factor. **Four common classes of DNA-binding domain are the helix-turn-helix and helix-loop-helix motifs, zinc fingers (below), and leucine zippers.** Most known sequence-specific transcription factors contain at least one of these DNA-binding motifs and proteins with unknown function that contain any of these motifs are likely to be transcription factors. The average transcription factor has 20 or more sites of contact with DNA, which amplifies the strength and specificity of the interaction.

In addition to a DNA-binding domain, sequence-specific transcription factors also have a **transcription-regulatory domain** that is required for their ability to modulate transcription. This domain may function in a variety of ways. It may interact directly with the RNA polymerase–general transcription factor complex, it may have indirect effects via coactivators or corepressor proteins, or it may be involved in remodeling the chromatin (below) and so alter the ability of the promoter to recruit other transcription factors.

One way to characterize the interaction of a transcription factor with a particular DNA sequence is to use a technique termed the **electrophoretic mobility shift assay (EMSA)**; (see Fig. 35.3). This method has been used to aid in the purification of transcription factors, to identify them in complex mixtures (such as a cellular extract), to delineate the size of the binding site, and to estimate the strength of the interaction between the factor and the DNA sequence it recognizes.

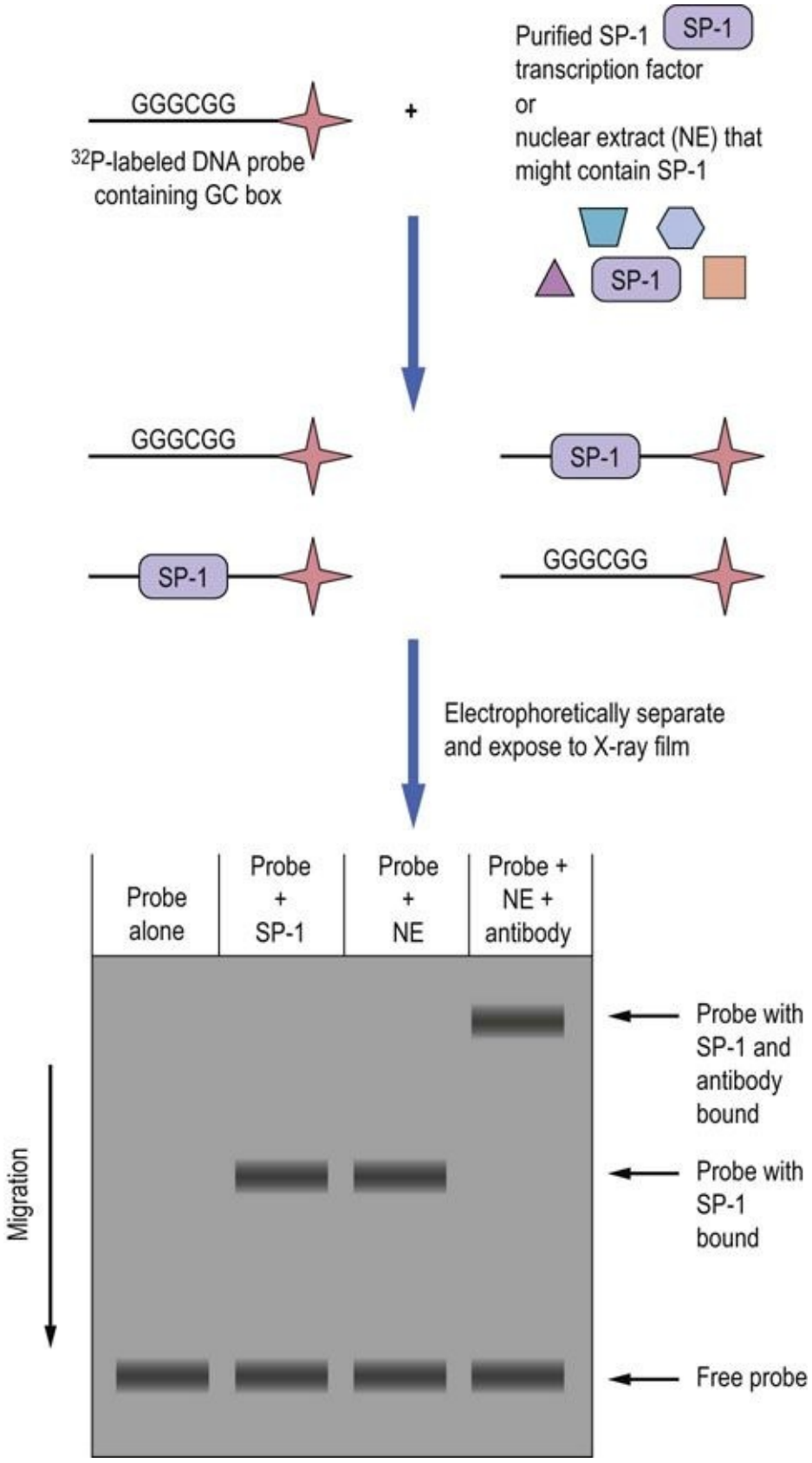

GGGCGG
32P-labeled DNA probe containing GC box
+
Purified SP-1 transcription factor or nuclear extract (NE) that might contain SP-1
SP-1
SP-1
GGGCGG
SP-1
SP-1
GGGCGG
Electrophoretically separate and expose to X-ray film
Probe alone
Probe + SP-1
Probe + NE
Probe + NE + antibody
Migration
Probe with SP-1 and antibody bound
Probe with SP-1 bound
Free probe

FIG. 35.3 Electrophoretic mobility shift assay (EMSA).
The basic components of the EMSA method are included in the diagram. In this example, a DNA probe that is labeled with ^{32}P is first incubated with either the purified transcription factor SP-1 or with nuclear extract (NE) that contains SP-1, and then subjected to native (nondenaturing) gel electrophoresis. If SP-1 binds to the probe it has a slower mobility in the gel than probe without protein bound (free probe). If an antibody to SP-1 is included in the reaction with NE, then the probe/anti-SP-1 antibody complex migrates even more slowly, confirming that the protein bound to the probe is indeed SP-1. EMSA can be used to help characterize any nucleic acid–protein interaction, including the interaction of RNA and proteins.

Steroid receptors

Steroid receptors possess many characteristics of transcription factors and provide a model for the role of zinc finger proteins in DNA binding

Steroid hormones have a broad range of functions in humans and are essential to normal life. They are derived from a common precursor, cholesterol, and thus share a similar structural backbone (Chapter 17). However, differences in hydroxylation of certain carbon atoms and aromatization of the steroid A-ring give rise to marked differences in biological effect. Steroids bring about their biological effects by binding to steroid-specific hormone receptors; these receptors are found in the cell cytoplasm and nucleus. For the type I (cytoplasmic) receptors, the steroid ligand induces structural changes that lead to dimerization of the receptor and exposure of a **nuclear localization signal (NLS)**; this signal, as well as dimerization, is commonly blocked by a heat shock protein that is released on steroid binding. The ligand–receptor complex now enters the nucleus, where it binds to DNA at a specific response element, the SRE, alternatively called the **hormone response element (HRE)**. SREs may be found many kilobases upstream or downstream of the start of transcription. The steroid–receptor complex functions as a sequence-specific transcription factor and binding of the complex to the SRE results in activation of the promoter and initiation of transcription (Fig. 35.4) or in some cases in the repression of transcription. As might be expected, because of the large number of steroids found in humans, there are correspondingly large numbers of distinct steroid receptor proteins, and each of these recognizes a consensus sequence, an SRE, in the region of a promoter.

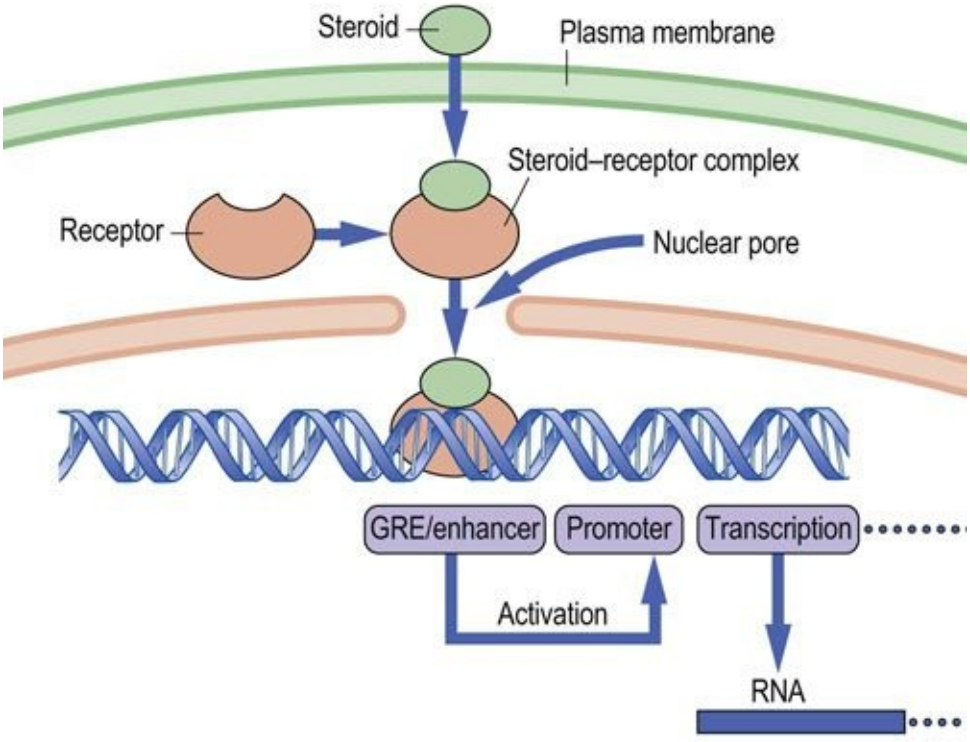


FIG. 35.4 Regulation of gene transcription by glucocorticoids.
Steroids induce dimerization of receptor molecules that, in turn, bind to an enhancer, activating transcription of the gene. GRE, glucocorticoid response element.

The zinc finger motif

A zinc finger motif in steroid receptors binds to the steroid response element in DNA

Central to the recognition of the SRE in the DNA, and to the binding of the receptor to it, is the presence of the so-called zinc finger region in the DNA-binding domain of the receptor molecule. Zinc fingers consist of a peptide loop with a zinc atom at the core of the loop. In the typical zinc finger, the loop comprises two cysteine and two histidine residues in highly conserved positions relative to each other, separated by a fixed number of intervening amino acids; the Cys and His residues are coordinated to the zinc ion. The zinc finger mediates the interaction between the steroid receptor molecule and the SRE in the major groove of the DNA double helix, thus enhancing the efficiency of, and conferring specificity on, the promoter. Zinc finger motifs are generally organized as a series of tandem repeat fingers; the number of repeats varies in

different transcription factors. The precise structure of the steroid receptor zinc finger differs from the consensus sequence; the two are compared in Figure 35.5.

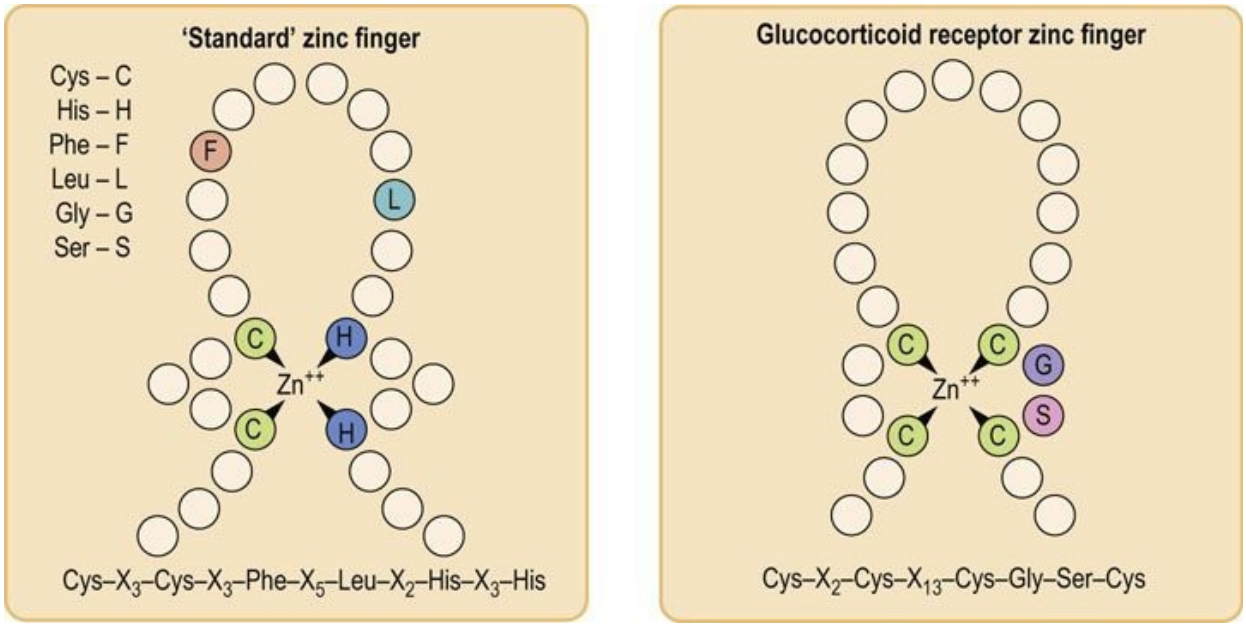


FIG. 35.5 A 'standard' zinc finger and a steroid receptor zinc finger.
Zinc fingers are commonly occurring sequences that allow protein binding to double-stranded DNA. C, cysteine; G, glycine; S, serine; X, any intervening amino acid.

Zinc finger proteins recognize and bind to short **palindromic sequences** of DNA. Palindromes are DNA sequences that read the same (5′ to 3′) on the antiparallel strands, *e.g.* 5′-GGATCC-3′, which reads the same 5′ to 3′ sequence on the complementary strand. The dimerization of the receptor and recognition of identical sequences on opposite strands strengthen the interaction between receptor and DNA and thus enhance the specificity of SRE recognition.

Organization of the steroid receptor

Steroid receptors are products of a highly conserved gene family

One central feature of all the steroid receptor proteins is the similarity in organization of their receptor molecules. Each receptor has a DNA-binding domain, a transcription-activating domain, a steroid hormone-binding domain,

and a dimerization domain. There are three striking features about the structure of the steroid hormone receptors:

■ The DNA-binding region always contains a highly conserved zinc finger region, which, if mutated, results in loss of function of the receptor.

■ The DNA-binding regions of all the steroid hormone receptors have a high degree of homology to one another.

■ The steroid-binding regions show a high degree of homology to one another.

These common features have identified the steroid receptor proteins as products of a gene family. It would appear that, during the course of evolution, diversification of organisms has resulted in the need for different steroids with varied biological actions and, consequently, a single ancestral gene has undergone duplication and evolutionary change over millions of years, resulting in a group of related but slightly different receptors (Fig. 35.6).

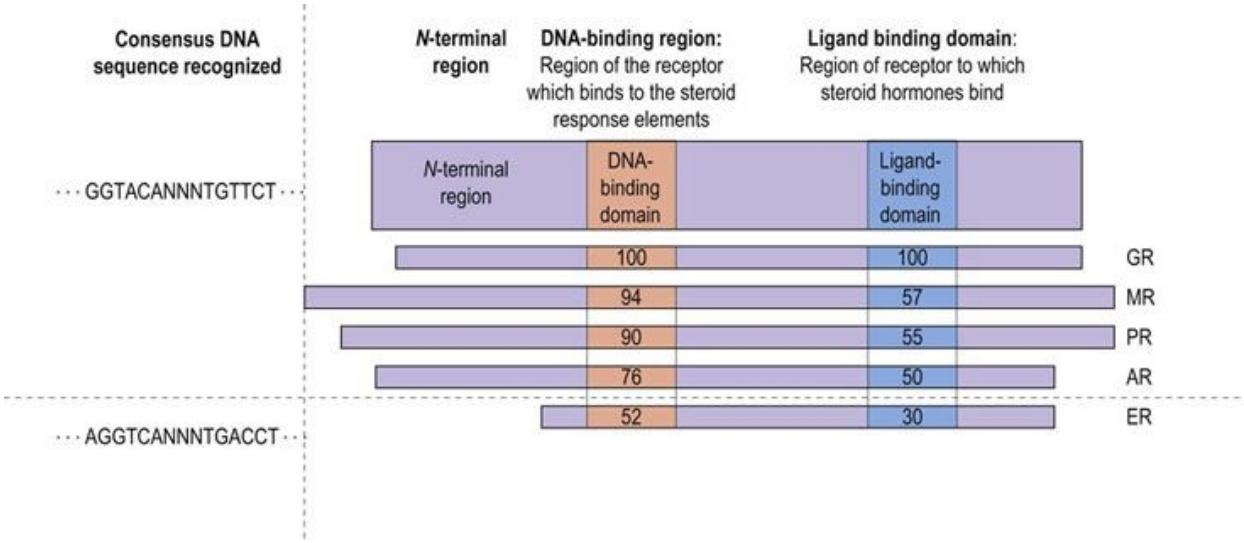


FIG. 35.6 Similarity between different steroid receptors.
The DNA-binding and hormone-binding regions of steroid receptors share a high degree of homology. The estrogen receptor is less similar to the glucocorticoid receptor than are the others. AR, androgen receptor; ER, estrogen receptor; GR, glucocorticoid receptor; MR, mineralocorticoid receptor; PR, progesterone receptor; NNN, any three nucleotides. Numbers denote percent amino acid homology to sequence in GR.

Advanced concept box Steroid receptor gene family: the thyroid

hormone receptors

The steroid receptor gene family, although large, is in fact only a subset of a much larger family of so-called nuclear receptors. All members of this family have the same basic structure as the steroid hormone receptors: a hypervariable *N*-terminal region, a highly conserved DNA-binding region, a variable hinge region, and a highly conserved ligand-binding domain (see Fig. 35.6). They are separated into two basic groups. **Type I (cytoplasmic) receptors** are a group of receptor proteins that form homodimers and bind specifically to steroid hormone response elements only in the presence of their ligand, such as the glucocorticoid receptor. **Type II (nuclear) receptors** form homodimers that can bind to response elements in the absence of their ligand, and may also form heterodimers with other type II receptor subunits, to form active units. The type II receptors include the thyroid hormone, vitamin D, and retinoic acid receptors.

Alternative approaches to gene regulation in humans

Promoter access

Chromatin structure affects access of transcription factors to genes, and thereby affects gene expression

DNA in the cell nucleus is packaged into nucleosomes and higher-order structures in association with histones and other proteins (Chapter 32). Thus, the promoters of some genes may not be readily accessible to transcription factors, even if the transcription factors themselves are present in the nucleus. It has become evident that the degree of packaging of a promoter and the presence, absence or precise location of nucleosomes on a promoter can have major effects on the degree of access for both sequence-specific transcription factors and the RNAPol II complex associated with general transcription factors. Condensed chromatin, termed heterochromatin, where the DNA is tightly associated with the nucleosomes, is usually not a good template for transcription and usually it is necessary for chromatin remodeling (see below) to occur before transcription proceeds. Euchromatin, comprised of DNA that is not as tightly associated with nucleosomes, typically contains regions of active gene transcription. Remodeling may also be necessary in portions of euchromatin, depending on the cell or tissue, but the initial state of the chromatin is less condensed. Portions of a chromosome that are packaged into heterochromatin in one type of cell may be in the form of euchromatin in a different cell, allowing for regulation of gene expression at the level of DNA accessibility. Certain portions of chromosomes, such as the centromeres and telomeres, are examples of regions of the genome that are packaged as heterochromatin in all cells.

Histone packaging, nucleosome stability and therefore the accessibility of DNA are controlled by reversible acetylation and deacetylation of lysine residues in the amino-terminal regions of the core histones, particularly histones H3 and H4. Histone acetyl transferases (HAT) transfer acetyl groups from acetyl-CoA to the amino group of lysines, neutralizing the charge on the lysine residue and weakening the strength of histone–DNA interactions, thereby permitting the relaxation of the nucleosome. Conversely, enzymes that remove

the acetyl groups and promote the local condensation of chromosomes are known as histone deacetylases (HDACs).

Reversible acetylation and deacetylation of histones is important in controlling the activation of promoters. Indeed, some transcription factors or transcription coactivators have HAT activity themselves and can remodel chromatin. In the case of some promoters, binding of a transcription factor may result in repositioning of the nucleosomes the next time the DNA replicates, making the gene more likely to be transcribed after cell division. The dynamic interplay of chromatin structure, transcription factor and cofactor binding is important in determining whether a gene is transcribed and how efficiently the RNA polymerase synthesizes it.

Methylation of DNA regulates gene expression

Methylation is one of several epigenetic modifications of DNA; patterns of DNA methylation at birth affect risk for a number of age-related diseases

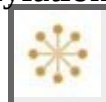
Certain nucleotides, principally cytidine at the 5 position on the pyrimidine ring, can undergo enzymatic methylation without affecting Watson–Crick pairing. The methylated cytidine residues are usually found associated with a guanosine, as the dinucleotide CpG, and in double-stranded DNA the complementary cytidine is also methylated, giving rise to a palindromic sequence:

5′ ^{m}CpG 3′

3′ GpC^{m} 5′

The presence of the methylated cytidine can be examined by susceptibility to restriction enzymes (Chapter 36) that cut DNA at sites containing CpG groups only if the CpG is unmethylated, compared to other restriction enzymes that cut whether or not the CpG is methylated. In addition, a bisulfite sequencing technique that relies on the differential reactivity of methyl cytidine can be used to more precisely map the sites of methylation. Many genes in humans (about 50%) have what are called **CpG islands (CPI)** in the region of their promoters. Generally, these CPI have been found to be unmethylated except in certain pathologic states and cancer. It has become clear that **methylation is generally associated with regions of DNA that are less actively transcribing RNA**. Demethylation of a promoter may be required for the initiation of transcription, and demethylation of a coding sequence of the gene may also be required for

efficient transcription. Regulation of the methylation state of promoters may be a more dynamic process than previously believed, for instance a decrease in the methylation of certain gene promoters in muscle after exercise.



Advanced concept box Epigenetic regulation of gene expression

Methylation is one aspect of the study of **epigenetics**, a broad field that, in general, addresses heritable modifications of DNA and protein that do not alter the sequence of DNA, but affect gene expression. Epigenetic control mechanisms encompass **DNA methylation, and histone acetylation, methylation or phosphorylation.** Nutrigenomics addresses the role of nutritional factors in the regulation of gene expression. There is some evidence that early nutritional intervention during a short time frame can cause nutritional imprinting by epigenetic mechanisms, which might prime the gene expression machinery for the development of diseases much later in life. Thus, the risk for age-related diseases, such as metabolic syndrome, obesity, atherosclerosis, diabetes, arthritis and cancer, may be affected by diet and lifestyle factors during youth.

The impact of epigenetic factors may change our approach to medical care, emphasizing the importance of preventive medicine and early intervention for control of age-related diseases, since the starting point to disease susceptibility may happen many years before the onset of the first symptoms. Hypermethylation of tumor suppressor genes is commonly observed in human cancers. Drugs that inhibit DNA methyl transferases are being tested as a means to uncover these repressed genes for treatment of leukemias. Genes that negatively regulate cell growth are often repressed by deacetylation of histones, creating a more compact (untranscribable) form of chromatin. **Histone deacetylase (HDAC) inhibitors** are also being tested as therapeutic agents for treatment of rapidly growing cancers, such as lymphomas.

Alternative splicing of mRNA

Alternative splicing yields many variants of a protein from a single pre-mRNA

In Chapter 33, the concept of the splicing of the initial transcript or pre-mRNA was introduced. Most pre-mRNAs can be spliced in alternative ways and the percentage of the alternative splicing of multi-exon transcripts is at least 50%, but with some estimates it may be as high as 80%. This process may provide sufficient diversity to explain individual uniqueness, despite similarities in the gene complement. Thus, by alternative splicing, a particular exon or exons may be spliced out on some but not all occasions. Since most genes **h**ave a number of exons (the average is about seven), some pre-mRNAs can eventually give rise to many different versions of the mRNA and, likewise, the final protein. The proteins may differ by only a few amino acids or may have major differences and often have different biological roles. For example, whether an exon is deleted or not may affect where in the cell the protein is localized, whether a protein remains in the cell or is secreted, and whether there are specific isoforms in skeletal versus cardiac muscle. Alternative splicing may also yield a truncated protein, known as a **dominant negative mutant protein**, that can inhibit the function of the full-length protein. Alternative splicing is regulated, so that particular splice forms may only be seen in certain cells or tissues, at defined stages of development or under differing conditions. In the human brain, there is a family of cell surface adhesion proteins, the **neurexins**, which mediate the complex network of interactions between approximately 10^{12} neurons. The neurexins are among the largest human genes, and hundreds, perhaps thousands, of neurexin isoforms are generated from only three genes by alternative promoters and splicing, providing for a diverse range of intercellular communications required for the development of sophisticated neural networks. The neurexins probably have an equally complex set of ligand isoforms, providing tremendous flexibility for reversible cellular interactions during the development of the central nervous system.

Clinical box Alternative splicing and tissue-specific expression of a gene: a girl with a swelling on the

neck

A 17-year-old girl noticed a swelling on the left side of her neck. She was otherwise well, but her mother and maternal uncle have both had adrenal tumors removed. Blood was withdrawn and sent to the laboratory for measurement of calcitonin, which was greatly increased. Pathology of the excised thyroid mass confirmed the diagnosis of medullary carcinoma of the thyroid. This family has a genetic mutation causing the condition known as multiple endocrine neoplasia type IIA (MEN IIA). MEN IIA is an autosomal dominant cancer syndrome of high penetrance caused by a germline mutation in the RET protooncogene. About 5–10% of cancers result from germline mutations, but additional somatic mutations are required for cancer to develop.

Comment.

Expression of the calcitonin gene provides an example of how different mechanisms may regulate gene expression and give rise to tissue-specific gene products that have very different activities. The calcitonin gene consists of five exons and uses two alternative polyadenylation signaling sites. In the thyroid gland, the medullary C cells produce calcitonin by using one polyadenylation signaling site associated with exon 4 to transcribe a pre-mRNA comprising exons 1–4. The associated introns are spliced out and the mRNA is translated to give calcitonin; elevated calcitonin is diagnostic for this condition. However, in neural tissue, a second polyadenylation signaling site next to exon 5 is used. This results in a pre-mRNA comprising all five exons and their intervening introns. This larger pre-mRNA is then spliced and, in addition to all the introns, exon 4 is also spliced out, leaving an mRNA comprising exons 1–3 and 5, which is then translated into a potent vasodilator, calcitonin gene-related peptide (CGRP).

Editing of RNA at the post-transcriptional level

The editosome modifies the internal nucleotide sequence of mature mRNAs

RNA editing involves enzyme-mediated alteration of mature mRNAs, before translation. This process, performed by editosomes (Chapter 33) may involve the insertion, deletion or conversion of nucleotides in the RNA molecule. Like alternative splicing, the substitution of one nucleotide for another can result in tissue-specific differences in transcripts. For example, *APOB*, the gene for human apolipoprotein B (apoB), a component of low-density lipoprotein, encodes a 14.1 kb mRNA transcript in the liver and a 4536-amino acid protein product, apoB100 (Chapter 18). However, in the small intestine, the mRNA is translated into a protein product, called apoB48, which is 2152 amino acids long (~48% of 4536), those amino acids being identical to the first 2152 amino acids of apoB100. The difference in protein size occurs because, in the small intestine, nucleotide 6666 is 'edited' by the deamination of a single cytidine residue, converting it to a uridine residue. The resulting change, from a glutamine to a stop codon, causes premature termination, yielding apoB48 in the intestine (Fig. 35.7).

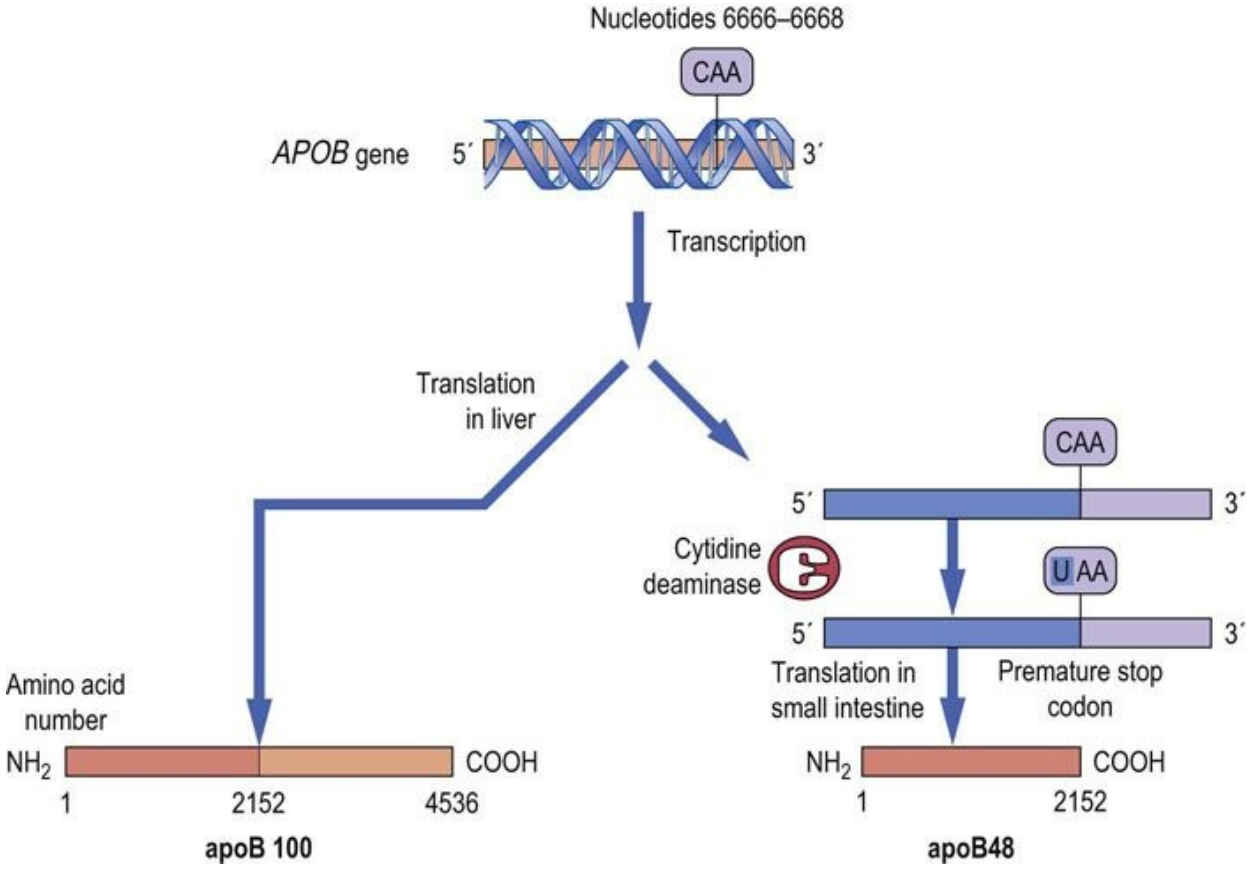


FIG. 35.7 RNA editing of the *APOB* gene in man gives rise to tissue-specific transcripts.
In the small intestine, nucleotide 6666 of apoB mRNA is converted from a cytosine to uracil by the action of the enzyme cytidine deaminase. This change converts a glutamine codon in apoB100 mRNA to a premature stop codon and, when the mRNA is translated, the truncated product apoB48 is produced. (See also Chapter 18.)

In addition to this cytidine deaminase, there are other enzymes that modify other mRNAs prior to translation, such as the **ADARs (*a*denosine *d*eaminases *a*cting on *R*NA)**. ADAR1 catalyzes the deamination of adenosine to inosine residues in dsRNAs; the RNA editing is essential for development of hematopoietic stem cells, and mutations in this enzyme in mice cause early embryonic death. ADAR2 modifies a neuronal glutamate receptor mRNA which results in the change of a single amino acid required for the function of the receptor; deficiency of this enzyme leads to seizures and neonatal death in mice.

Clinical box Iron status regulates translation of an iron carrier protein:

a man with breathlessness and fatigue

A 57-year-old Caucasian male presented to his family doctor with breathlessness and fatigue. He noticed that his skin had become darker. Clinical evaluation indicated cardiac failure with impaired left ventricular function as a result of dilated cardiomyopathy, a low serum concentration of testosterone and an elevated fasting concentration of glucose. Serum ferritin concentration was greatly increased, at >300 µg/L, and the diagnosis of hereditary hemochromatosis was suspected. The man was treated by regular phlebotomy until his serum ferritin was <20 µg/L (normal value 30-200 µg/L), at which point the phlebotomy interval was increased to maintain the serum ferritin concentration at <50 µg/L.

Comment.

In conditions of iron excess, for example in hemochromatosis, there is an increase in the synthesis of ferritin, an iron-binding and storage protein. Conversely, in conditions of iron deficiency, there is an increase in the synthesis of the transferrin receptor protein, which is involved in the uptake of iron. In both cases, the RNA molecules themselves are unchanged, and there is no change in the synthesis of the respective mRNAs. However, both the ferritin mRNA and the transferrin receptor mRNA contain a specific sequence known as the iron-response element (IRE), a specific IRE-binding protein can bind to mRNA. In iron deficiency, the IRE-binding protein binds the ferritin mRNA, prevents translation of ferritin, and binds the transferrin receptor mRNA and prevents its degradation. Thus, in iron deficiency, ferritin concentrations are low and transferrin receptor concentrations are high. In states of iron excess, the reverse process occurs and translation of ferritin mRNA increases, whereas transferrin receptor mRNA undergoes degradation, serum ferritin concentrations are high, and transferrin receptor concentrations are low (Fig. 35.8). About 10% of the US population carry the gene for hereditary hemochromatosis, but only homozygotes are affected with the disease. (See also discussion of iron metabolism and hemochromatosis in Chapter 22.)

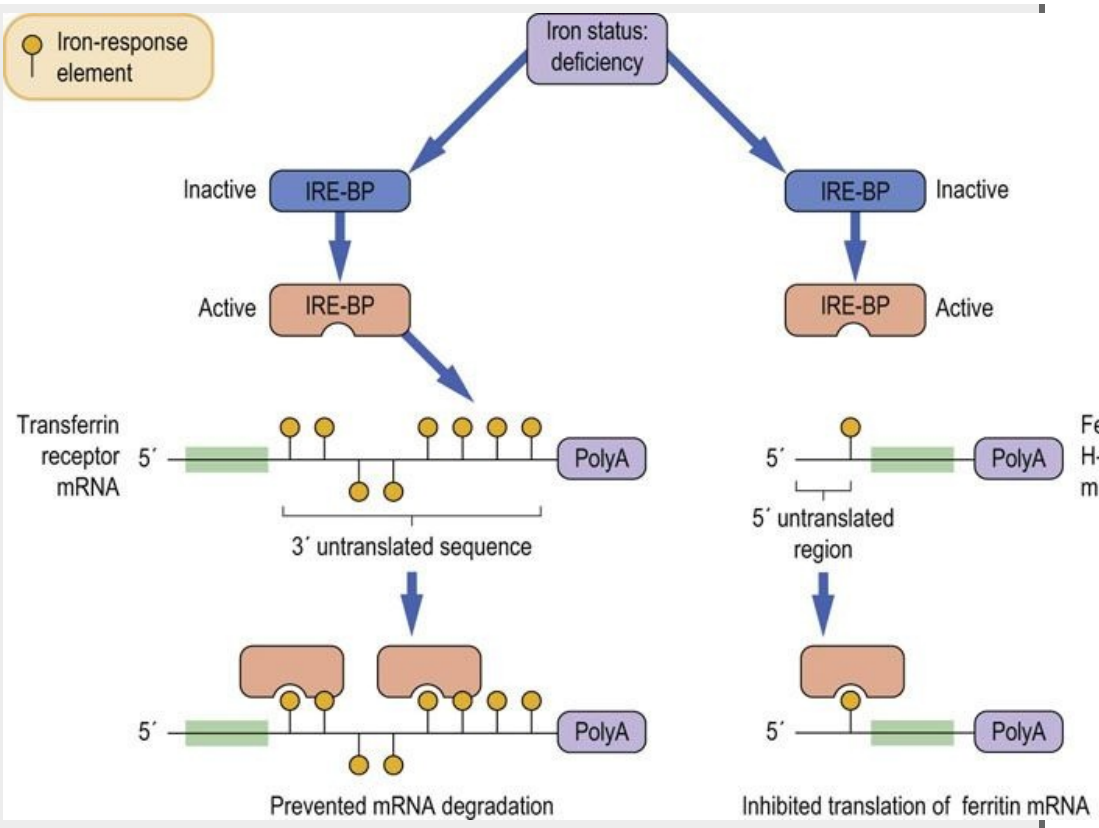


FIG. 35.8 Regulation of mRNA translation by iron status.
The binding of a specific binding protein to the iron response element (IRE) of the mRNA of iron-responsive genes can alter the translation of the mRNA into functioning proteins in different ways. When iron is deficient, the iron response element binding protein (IRE-BP) is activated and can bind to the 3′ end of the mRNA for the transferrin receptor. This prevents the degradation of the mRNA, and thus increases the amount of transferrin receptor that can be made (*left*) and thus increases the amount of iron that the receptor can deliver to the cell. However, the IRE-BP also binds to the 5′ end of the ferritin mRNA and prevents its translation (*right*). Ferritin is a protein that sequesters and stores iron in the cytoplasm, and less is needed in times of iron deficiency. (See also Fig. 22.8.)

RNA interference

RNA interference (RNAi), discussed in more detail in Chapter 33, is another way to control gene expression. At the heart of RNAi are very small noncoding RNAs, about 20–30 nucleotides long, known as micro RNAs (miRNAs). These

are involved in the attenuation or repression of translation by binding to the 3′ UTR of an mRNA and recruiting factors that inhibit protein synthesis, or by the destruction of the mRNA by an alternative pathway. During embryogenesis and in certain pathologic states, such as cancer, there are changes in the pattern of miRNA expression in cells, thereby changing gene expression in ways that might alter cell fate or favor cellular proliferation. RNAi holds promise in the treatment of human diseases where the inhibition of the expression of a gene product or the destruction of RNA would be therapeutic, such as in viral infections or cancer.

Preferential activation of one allele of a gene

Human genes are biallelic, but sometimes only one allele of the gene is expressed

The normal complement of human chromosomes comprises 22 pairs of autosomes and two sex chromosomes. In each of the pairs of autosomes the genes are present on both chromosomes: they are biallelic. Under normal circumstances, both genes are expressed without preference being given to either allele of the gene – that is, both the paternal and maternal copies of the gene can be expressed, unless there is a mutation in one allele that prevents this from occurring.

The situation with regard to sex chromosomes is slightly different. Sex chromosomes are of two types, X and Y, the X being substantially larger than the Y. Females have two X chromosomes, whereas males have one X and one Y chromosome. A region of the Y chromosome is identical to a region of the X chromosome but the X chromosome also contains genes that have no matching partner on the Y chromosome, and some genes on the Y chromosome are specific to the Y chromosome, for example SRY, a sex-determining gene. Such genes are said to be monoallelic; they offer no choice as to which allele of the gene will be expressed.

Apart from the specific cases of sex chromosomes, there would appear to be no reason why both alleles of a gene cannot be expressed. However, in humans, genes have been identified that are biallelic but only one allele – either maternal or paternal – is preferentially expressed, despite the fact that both alleles are perfectly normal or identical. As a result, only 50% of the gene product is produced but the product is functionally active. As outlined in Table 35.2, three

different mechanisms have been identified that can restrict the expression of biallelic genes in humans (Table 35.2).

Table 35.2

Examples of types of restriction of biallelic genes in humans

Genomic imprinting	For autosomal genes, imprinting may be tissue specific – monoallelic expression in some tissues, biallelic in others. Examples include insulin-like growth factor 2 (IGF-2) and Wilms' tumor susceptibility gene (WT1)
Allelic exclusion	Tissue-specific production of a single allele product; for example, synthesis of a single immunoglobulin light or heavy chain in a B cell from one allele only
X chromosome inactivation	Some genes on the X chromosome in females. Males exhibit only one allele of the X-linked gene but females have two, and one is inactivated by switching off nearly the entire X chromosome

For some genes, although two alleles exist in any particular cell, only one of these alleles is active. Hence the gene behaves as if it were monoallelic although it is, in fact, biallelic.

Advanced concept box X-chromosome inactivation

Males have one X-chromosome whereas females have two. Thus, genes on the X-chromosome are biallelic in females but monoallelic in males. In females, however, one of the X-chromosomes in each cell is inactivated at an early stage of embryogenesis. The inactivated X may be the paternally derived or the maternally derived X-chromosome; for any particular cell, which one is inactivated is random, but the descendants of that cell will have the same X inactivated. The inactivated X-chromosome can still express a few genes, however, including XIST (inactive X–Xi-specific transcript) that codes for an RNA which plays an important role in X-chromosome inactivation. There is methylation of CpG islands on most of the genes on the inactivated chromosome and this represses transcription. The inactivated X-chromosome is reactivated during oogenesis in the female.

Summary

■ The control of gene expression involves both transcriptional and post-transcriptional events that regulate the expression of a gene in both time and place and in response to numerous developmental, hormonal and stress signals.
■ DNA sequences and DNA-binding proteins control gene expression. The DNA sequences include *cis*-acting promoters, such as the TATA box, and enhancers and response elements.
■ The DNA-binding proteins are *trans*-acting transcription factors that bind with high specificity to these sequences, and facilitate the binding and positioning of RNAPol II for synthesis of pre-mRNA.
■ Other factors that affect the conversion of gene to protein include access of the transcriptional apparatus to the gene, enzymatic modification of histones and nucleotides in the DNA, factors that effect alternative intron splicing, post-transcriptional editing of pre-mRNA, RNA interference, and restricted expression of biallelic genes.

Active learning

1. How are steroid response elements identified in the genome? Discuss the consequences of a mutation in an SRE versus a mutation in the SRE binding protein. How does the zinc finger protein for the glucocorticoid receptor differ from that for the androgen or estrogen receptor?
2. What are the biochemical consequences of *APOB* gene editing in humans? Compare the effects of editing to introduce a substitution, compared to an insertion or deletion, in an mRNA molecule.
3. Some genes have promoters that have no TATA box (TATA-less genes). Without this box, what determines where the RNAPol II complex will start transcription?
4. Compare the total number of genes to the number of translated proteins that may be synthesized by the human genome. What fraction of the genes encode transcription factors? Compare the concentration of transcription factors to the concentration of glycolytic enzymes in the cell.

Further reading

Barres, R, Yan, J, Egan, B, *et al.* Acute exercise remodels promoter methylation in human skeletal muscle. *Cell Metabolism*. 2012; 15:405–411.

Berkhout, B, Jeang, KT. RISCy business: microRNAs, pathogenesis, and viruses. *J Biol Chem*. 2007; 282:26641–26645.

Brena, RM, Costello, JF. Genome-epigenome interactions in cancer. *Hum Mol Genet*. 2007; 16:R96–R105.

Glaser, KB. HDAC inhibitors: clinical update and mechanism-based potential. *Biochem Pharmacol*. 2007; 74:659–671.

Goodrich, JA, Tjian, R. Unexpected roles for core promoter factors in cell-type specific transcription and gene regulation. *Nat Rev Genet*. 2010; 11:549–558.

Kaelin, WG. Use and abuse of RNAi to study mammalian gene function. *Science*. 2012; 337:421–422.

Iijima, T, Wu, K, Witte, H, *et al.* SAM68 regulates neuronal activity-dependent alternative splicing of neurexin-1. *Cell*. 2011; 147:1601–1614.

Li, Q, Lee, J-A, Black, DL. Neuronal regulation of alternative pre-mRNA splicing. *Nature Rev Neurosci*. 2007; 8:819–831.

Nishikura, K. Functions and regulation of RNA editing by ADAR deaminases. *Annu Rev Biochem*. 2010; 79:321–349.

Oakley, RH, Cidlowski, JA. Cellular processing of the glucocorticoid receptor gene and protein: new mechanisms for generating tissue-specific actions of glucocorticoids. *J Biol Chem*. 2011; 286:3177–3184.

Wang, ET, Sandberg, R, Luo, S, *et al.* Alternative isoform regulation in human tissue transcriptomes. *Nature*. 2008; 456:470–476.

Witt, O, Deubzer, HE, Milde, T, *et al.* HDAC family: what are the cancer relevant targets? *Cancer Letts*. 2009; 277:8–21.

Websites

Alternate splicing. www.ensembl.org/info/docs/genebuild/ase_annotation.html.

Catalog of genetic diseases. www.ncbi.nlm.nih.gov/omim.

Epigenetics. http://epigenie.com/.

Genomic imprinting. www.geneimprint.com/site/home.

Noncoding RNA:

RNA editing. http://dna.kdna.ucla.edu/rna/index.aspx.

RNA interference. www.rnaiweb.com/.

www.eurasnet.info/tools/asdatabases

www.mirbase.org

www.noncode.org

CHAPTER 36

Regulation of Gene Expression

Genomics, Proteomics and Metabolomics

Andrew R. Pitt and Walter Kolch

Learning objectives

After reading this chapter you should be able to:

- Describe what the terms genomics, transcriptomics, proteomics, and metabolomics mean.
- Discuss the differences between the -omics methods and their particular challenges.
- Describe the structure of a gene and the regulation of gene transcription.
- Give several examples of methods used in genomics and transcriptomics.
- Give examples of methods used in proteomics.
- Describe methods used in metabolomics.
- Discuss biomarkers.

Introduction

The human genome is organized into 46 chromosomes consisting of 22 pairs of autosomal chromosomes, which are shared by both sexes, and the sex-determining chromosomes, X and Y. One set of autosomal chromosomes is derived from each parent. One X chromosome is contributed by the mother and another, either X or Y from the father. Most mammalian genes consist of multiple exons, which are the parts that eventually constitute the mature mRNA, and introns, which separate the exons and are removed from the primary transcript by splicing.

Many of the complex biological functions are generated by interaction between genes rather than by individual genes

Surprisingly, the 3 billion bases of the human genome only harbor 22,000–24,000 protein-coding genes. This is only about four times the number of genes of yeast and twice as many as the fruit fly *Drosophila melanogaster*, and less than many plants. Thus, many of the complex biological functions that characterize humans are generated by **combinatorial interaction between genes** rather than by individual genes being responsible for a specific function. This insight has replaced the dogma of one gene encoding one protein with one function. Mammalian cells use **alternative splicing** and alternative gene promoters to produce 4–6 different mRNA from a single gene, so that the number of protein-coding mRNAs, the transcriptome, may be as large as 100,000.

Post-translational modifications add further levels of complexity

This complexity is further augmented at the protein level by post-translational modifications and targeted proteolysis that could generate an estimated 500,000–1,000,000 functionally different protein entities, which comprise the proteome. An estimated 10–15% of these proteins function in metabolism, which collectively describes the processes used to provide energy and the basic low-molecular-weight building blocks of cells, such as amino acids, fatty acids and sugars. It also includes the processes that convert exogenous substances such as drugs or environmental chemicals.

The human metabolome database (www.hmdb.ca/) currently contains >40,000 entries. The real size of the metabolome is unknown, but is expected to increase with the number of environmental substances an organism is exposed to. The relationship between the different -omes is depicted in Figure 36.1.

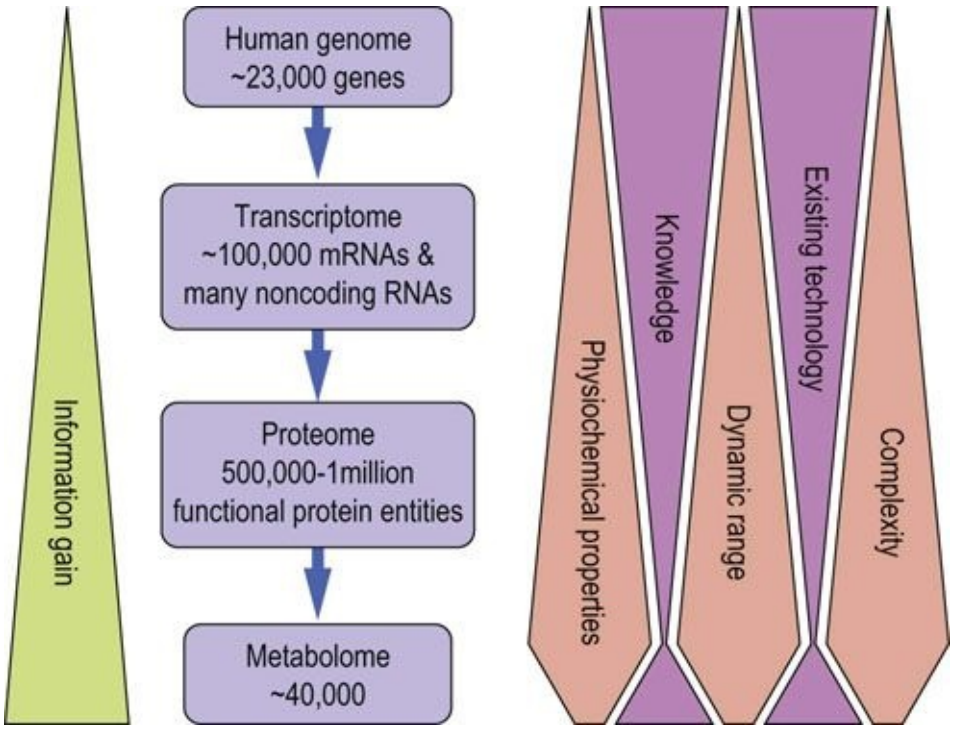


FIG. 36.1 Relationship between the -omics.
Complexity, diversity in physicochemical properties, and dynamic range increase as we move from genes to transcripts and proteins, but may decrease again at the level of metabolites. This presents a huge technologic challenge, but also represents a rich source of information gain, especially if the different -omics disciplines can be integrated into a common view.

Studies of genome, transcriptome, proteome and metabolome pose different challenges

The genome and transcriptome consist entirely of the nucleic acids DNA and RNA, respectively. Their uniform physicochemical properties have enabled efficient and ever cheaper methods for amplification, synthesis, sequencing and highly multiplexed analysis. The proteome and metabolome pose much bigger analytical challenges as they consist of molecules with widely different

physicochemical properties and highly variable abundance. For instance, the concentration of proteins in human serum spans 12 orders of magnitude, while under normal conditions genes are equally abundant. The genome is relatively static, whereas the transcriptome and proteome are more dynamic and can change quickly in response to internal and external cues. The most pronounced dynamic responses may manifest themselves in the metabolome as it directly reflects the interactions between organism and environment. Thus, the complexity increases as we move from the genome to the transcriptome, proteome and metabolome, while our knowledge decreases. The analysis of all -omics data requires large and sophisticated bioinformatics resources, and also has stimulated the development of **systems biology,** which uses mathematical and computational modeling to interpret the functional information about biological processes contained in these data.

Genomics

Genome analysis provides a way to predict the probability of a condition, but without providing information whether and when this probability will manifest itself

The 'whether and when' information can better be gained from the transcriptome, proteome and metabolome. They give a dynamic picture of the current state of an organism, and lend themselves to monitor changes in that state, *e.g.* during disease progression or treatment. Thus, the information provided by the -omics technologies is complementary, and their use for diagnostic purposes is mainly limited by the complexity of the equipment and analysis. Genomics and transcriptomics are making their way into the clinical laboratory, and are poised to become part of routine diagnostics in the next few years.

Many diseases have an inheritable genetic component

Many diseases are caused by genetic aberrations and many more manifest a genetic predisposition or component. The Online Mendelian Inheritance in Man (OMIM) database (www.omim.org/) currently lists more than 2,800 gene mutations that are associated with >4,700 phenotypes which cause or predispose to disease. These numbers suggest that many diseases are caused by mutations in single genes, and that many more have an inheritable genetic component. Thus, the genome holds a rich source of information about our physiology and pathophysiology. We now have a broad arsenal of techniques for genome analysis at our disposal, which allow the detection of gross abnormalities down to single nucleotide changes, and which are increasingly being used for clinical diagnostics.

Advanced concept box The human genome project

The Human Genome Project (HGP) officially began in 1990 and culminated with the deposition of the completed sequence into public databases in 2003. However, in-depth analysis and interpretation will go on for much longer. The HGP was unique in

several ways. It was the first global life science project, being coordinated by the Department of Energy and the National Institutes of Health (USA). The Wellcome Trust (UK) became a major partner in 1992, and further significant contributions were made by Japan, France, Germany, China, and other countries. More than 2800 scientists from 20 institutions around the world contributed to the paper describing the finished sequence in 2004. It also was conducted on an industrial scale with industrial-style logistics and organization. In fact, the HGP received competition from Celera Genomics, a private company founded in 1998, and the first draft sequences of the human genome were published in two parallel papers in 2001. The HGP used a **'clone-by-clone' approach** where the genome was cloned first and then these large clones were divided into smaller portions **and** sequenced. Celera followed a fundamentally different strategy, **shotgun sequencing**, where the whole genome is broken up into small pieces that can be sequenced directly with the full sequence being assembled afterwards. This approach is much faster but less reliable in producing continuous sequences, and much less able to mend gaps in the assembled sequence. The 2001 draft genomes estimated the existence of 30,000–35,000 genes. The refined HGP 2003 sequence confirmed 19,599 protein-coding genes and identified another 2188 DNA predicted genes, a surprisingly low number. They are contained in 2.85 billion nucleotides covering more than 99% of the euchromatin, *i.e.* gene-containing DNA. Many thousands of genomes have been sequenced since then, and the human reference genome is constantly updated. As by 2010 there are only ~250 gaps compared to the 150,000 in the draft sequence, and the current sequence is extremely accurate. Genome sequences are accessible publicly through all major nucleotide databases.

Karyotyping, comparative genome hybridization (CGH), chromosomal

microarray analysis (CMA) and fluorescence in situ hybridization (FISH)

Karyotyping assesses the general chromosomal architecture

Early successes of exploiting genome information for the diagnosis of human disease was the discovery of trisomy 21 as the cause of **Down syndrome** in 1959, and the discovery of the **Philadelphia chromosome** as associated with **chronic myelogenous leukemia** (CML) in 1960. Since then karyotyping has identified a large number of chromosomal aberrations, including amplifications, deletions and translocations, especially in tumors. The method is based on simple staining of chromosome spreads by Giemsa or other stains which reveal a banding pattern characteristic for each chromosome that is visible through the light microscope. Although it only reveals crude information such as number, shapes and gross alterations of general chromosomal architecture, it is still a mainstay of clinical genetic analysis.

Comparative genome hybridization compares two genomes of interest

A refinement of karyotyping is **comparative genome hybridization (CGH)**. The principle of CGH is to compare two genomes of interest, usually a diseased against a normal control genome. The genomes that are to be compared are labeled with two different fluorescent dyes. The fluorescently labeled DNAs are then hybridized to a spread of normal chromosomes and evaluated by quantitative image analysis (Fig. 36.2). As fluorescence has a large dynamic range (i.e. the relationship between fluorescence intensity and concentration of the probe is linear over a wide range), CGH can detect regional gains or losses in chromosomes with much higher accuracy and resolution than conventional karyotyping. Losses of 5–10 megabases (Mb) and amplifications of <1 Mb are detectable by CGH. However, balanced changes, such as inversions or balanced translocations, escape detection as they do not change the copy number and hybridization intensity.

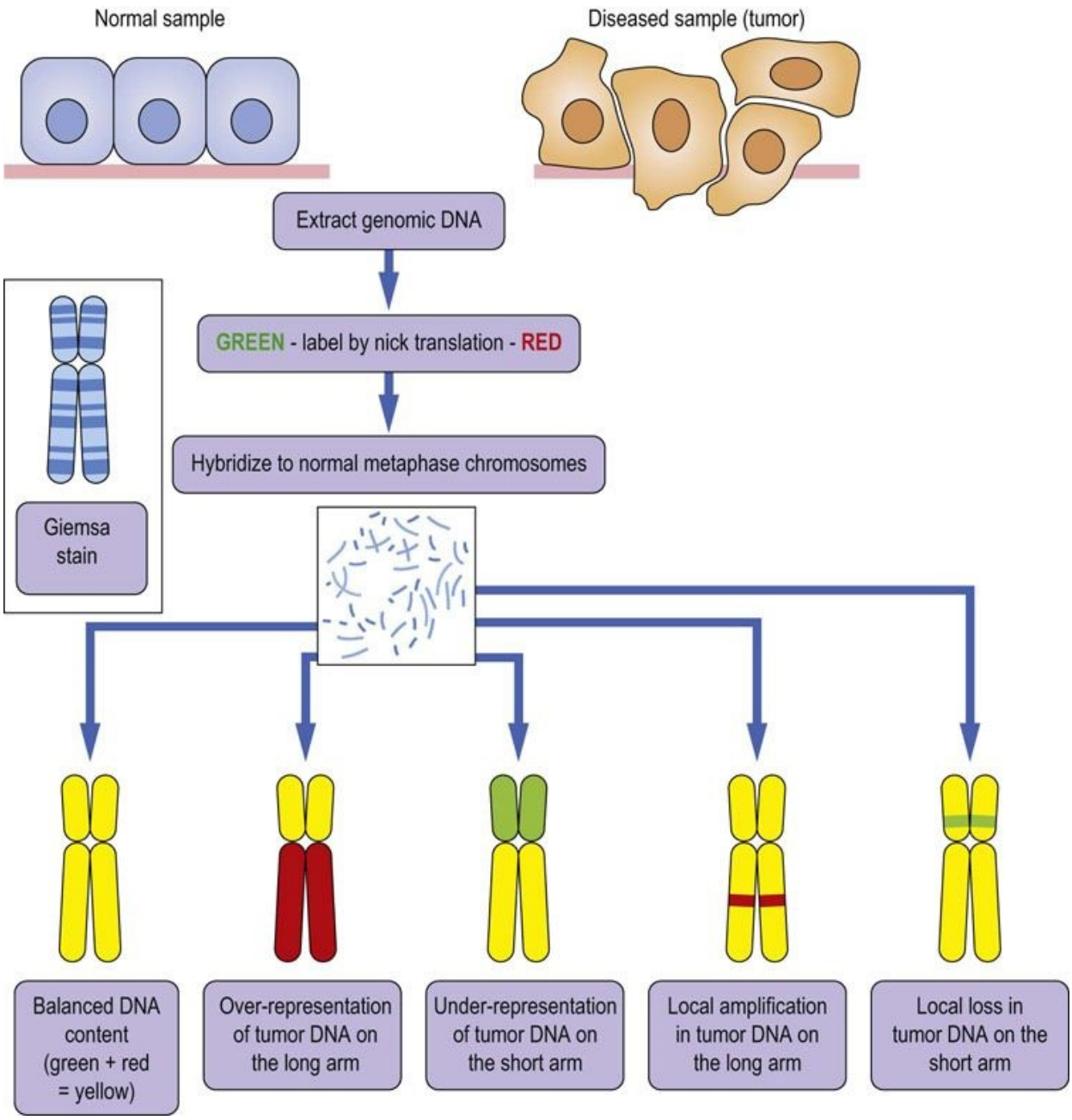


FIG. 36.2 Principles of comparative genome hybridization (CGH).
Genomic DNA is isolated from a normal and a diseased sample (here, from a tumor) to be compared. The DNA is labeled by nick translation with green or red fluorescent dyes, and hybridized to a normal chromosome spread. If the DNA content between samples is balanced, equal amounts of the control (green) and tumor (red) DNA will hybridize, resulting in a yellow color. Global or local amplifications or losses of genetic material will reveal themselves by a color imbalance.

In chromosomal microarray analysis the labeled DNA is

hybridized to an array of oligonucleotides

Further improvements in resolution are afforded by **chromosomal microarray analysis (CMA)**. In this method, the labeled DNA is hybridized to an array of oligonucleotides. Modern oligonucleotide synthesis and array manufacturing can produce arrays containing many million oligonucleotides on chips of the size of a microscope slide. By choosing the oligonucleotides so that they cover the region of interest to an equal extent, a very high resolution can be achieved, allowing the detection of copy number changes at the level of 5–10 kb in the human genome. CMA is used in **prenatal screening for the detection of chromosomal defects**. As the probe DNA can be amplified by the polymerase chain reaction (PCR; Fig. 36.3), only minute amounts of starting material are required. Notably, neither CGH nor CMA provides information about the ploidy. As long as the DNA content is balanced and carries no further aberrations, a tetraploid genome would be indistinguishable from a diploid genome.

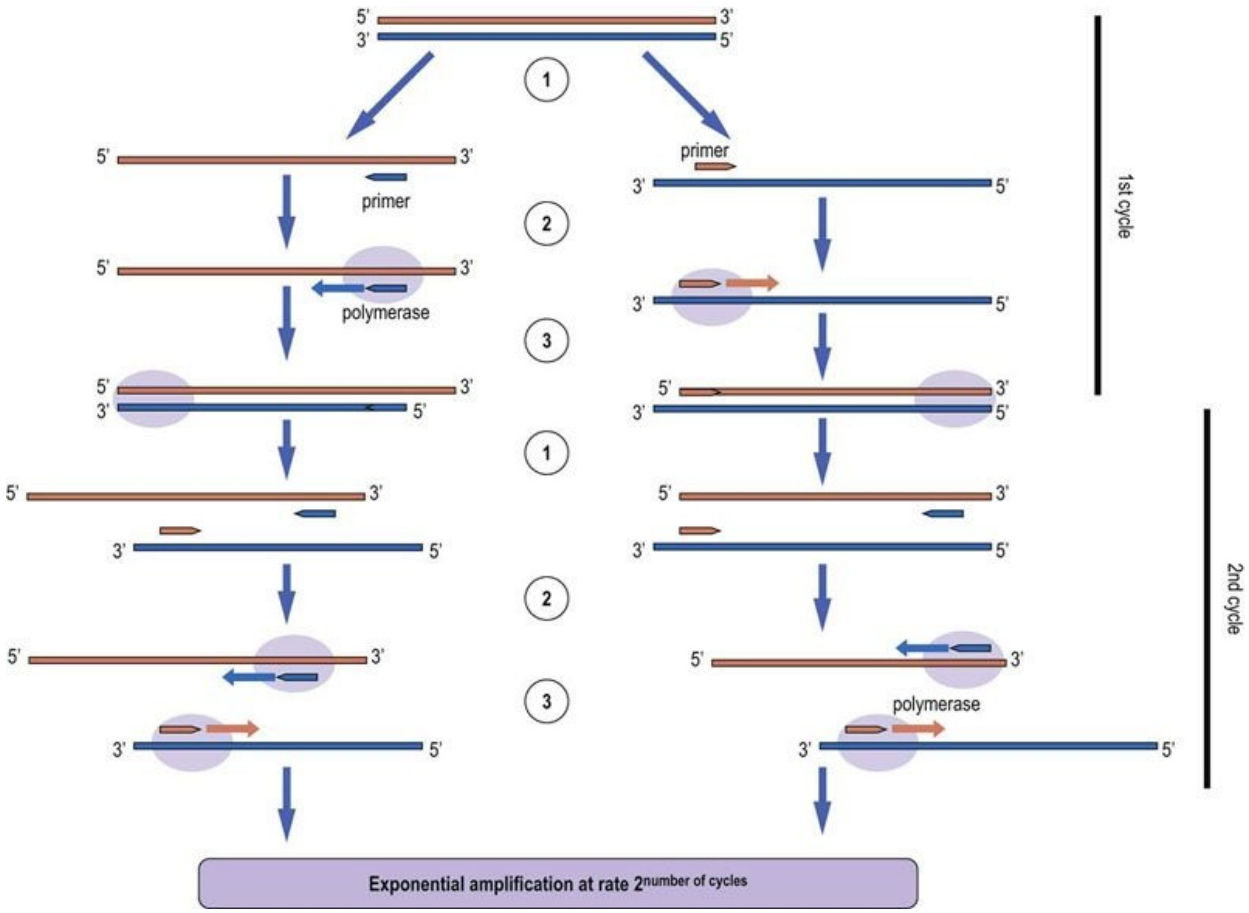


FIG. 36.3 Polymerase chain reaction (PCR).
This method is widely used for the amplification of DNA and RNA. The nucleic acid template is heat-denatured and specific primers are annealed by lowering the temperature **(step 1).** The primers are extended using reverse transcriptase if the template is RNA, or a DNA polymerase if the template is DNA **(step 2).** The result is a double-stranded product **(step 3),** which is heat denatured so that the cycle can start again. Typically, between 25 and 35 cycles are used. The amplification is exponential, and hence PCR enables us to analyze minute amounts of DNA or RNA down to the single cell level. The use of heat-stable and high-fidelity DNA polymerases permits amplification of fragments up to several thousand base pairs long. Many variations of PCR have been developed for a wide range of applications, such as molecular cloning, site-directed mutagenesis, generation of labeled probes for hybridization experiments, quantitation of RNA expression, DNA sequencing, genotyping, and many others.

Fluorescence in situ hybridization can be used when the gene in question is known

If the gene of interest is known, the respective recombinant DNA can be labeled and used as a probe on chromosome spreads. This method, called **fluorescence in situ hybridization (FISH)**, can detect gene amplifications, deletions and chromosomal translocations. Using different colored fluorescent labels, several genes can be stained simultaneously.

Gene mutations can be studied by sequencing

Efforts to find individual disease genes were hampered by our insufficient knowledge of the genome and by the lack of high-resolution mapping methods. This situation dramatically changed with the completion of the human genome sequence in 2003 and the rapid development of novel technologies, called **next-generation sequencing (NGS)**, which made sequencing very fast and affordable.

Four principles of DNA sequencing

There are four principles of DNA sequencing. (i) The **Maxam–Gilbert method** uses chemicals to cleave the DNA at specific bases and then separate the fragments on high-resolution gels, allowing the sequence to be read from the size of the fragments. (ii) **The Sanger method** uses a polymerase to synthesize DNA in the presence of small amounts of chain-terminating nucleotides (Fig. 36.4). This and the Maxam–Gilbert method were the first successful DNA sequencing techniques. While the latter has become obsolete, the Sanger method

is still widely used. (iii) The extension of the complementary DNA strand is measured when a matching nucleotide is added. (iv) The ligation of a synthetic oligonucleotide to the DNA target to be sequenced, which only occurs when a nucleotide pair in the oligonucleotide matches the sequence of the target DNA at the correct position, is monitored. Variations of methods ii–iv are incorporated in NGS workflows.

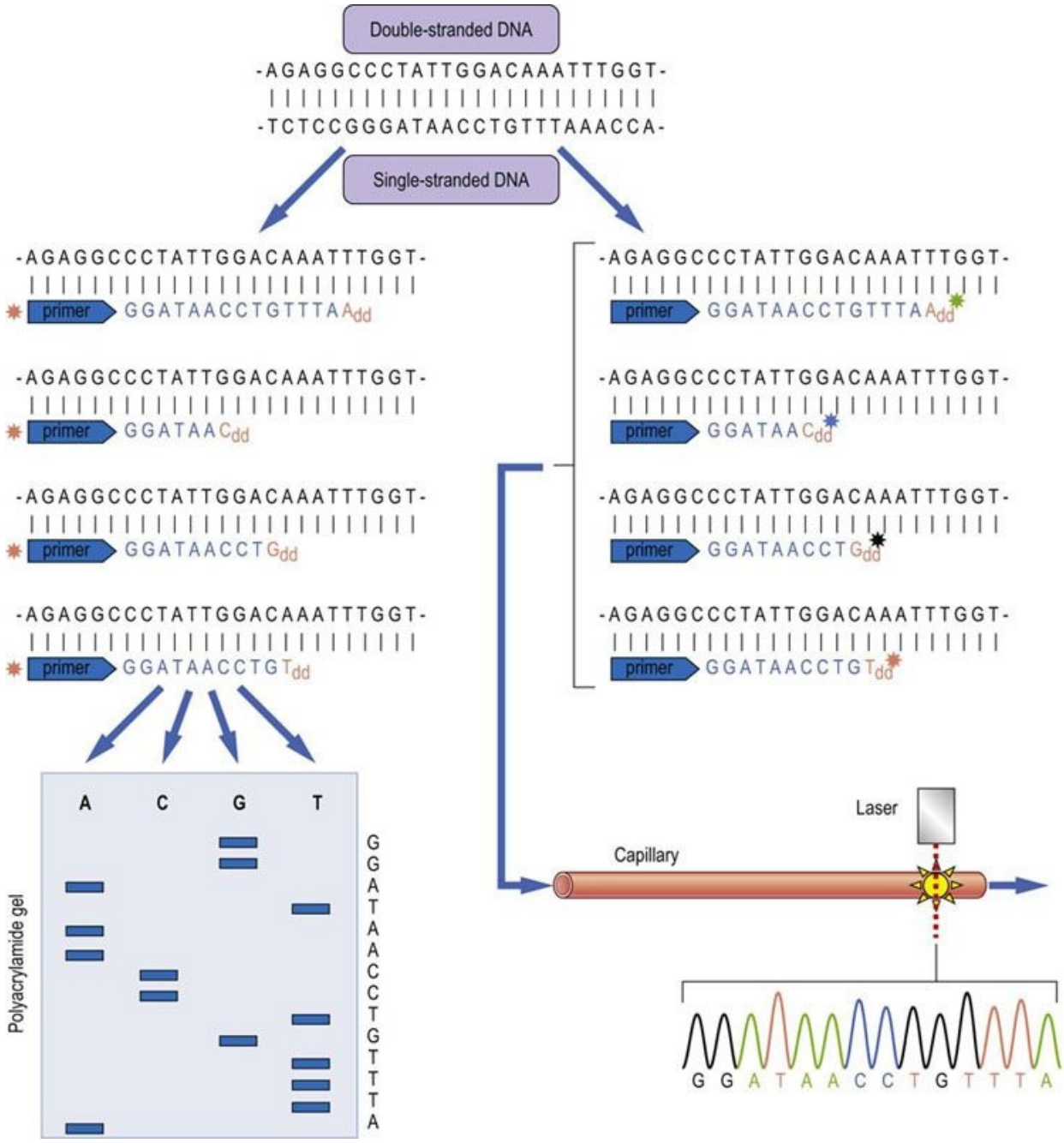


FIG. 36.4 DNA sequencing using Sanger's chain termination method.
Double-stranded DNA is heat-denatured to generate single-stranded DNA. Primers (usually hexamers of random sequence) are annealed to generate random initiation sites for DNA synthesis, which is carried out in the presence of DNA polymerase, deoxynucleotides (dNTPs) and small amounts of dideoxynucleotides (ddNTPs). The ddNTPs lack the 3'-hydroxyl group which is required for DNA strand elongation. They terminate the synthesis, leading to fragments of different sizes, each ending a specific nucleotide. These fragments can be separated by polyacrylamide gel electrophoresis and the sequence read from the 'ladder' of fragments on the gel. To visualize the fragments, either the DNA can be labeled by adding in radioactive or fluorescent dNTPs, or the primers can be labeled (as indicated in the figure) with a fluorescent dye. Using ddNTPs labeled with different dyes permits all four reactions being mixed together and separated by capillary electrophoresis. Online laser detection enables direct reading of the sequence. This 'capillary DNA sequencing' gives longer reads than gels, permits multiplexing, and high throughput. It was the method used for most of the sequencing in the Human Genome Project.

There are several NGS methods using different ways to read the DNA sequence

All NGS methods share the principle of conducting many millions of parallel sequencing reactions in microscopic compartments on arrays or nanobeads. These sequence pieces are assembled into complete genome sequences using sophisticated bioinformatics methods. While the first human genome sequence cost $3 billion and took more than10 years to complete, thanks to NGS we now can sequence a human genome in a single day for ~$1000. Thus, NGS has enabled the large-scale hunt for gene mutations by direct sequencing. Notable examples of such projects are **the Cancer Genome Projects** executed by the Wellcome Trust Sanger Centre in the UK and the US National Cancer Institute. The aim of these projects is to establish a systematic map of mutations in cancer and utilize this map for risk stratification, early diagnosis and choice of the best treatment in patients.

Single nucleotide polymorphisms (SNPs) are useful in identification and assessment of disease risk

Genomes in a population vary slightly by small changes, most often just concerning single nucleotides, called **single nucleotide polymorphisms (SNPs)**. The most common way to examine SNPs is by direct sequencing or array-based

methods. For the first method, DNA is usually amplified by PCR and then sequenced. For the second method, oligonucleotide arrays containing all possible permutations of SNPs are probed with genomic DNA, so that successful hybridization only occurs when the DNA sequences match exactly.

Systematic SNP mapping has proven useful in studying genetic identity and inheritance, and also in the identification and risk assessment of genetic diseases

The initial human genome sequences yielded ca. 2.5 million SNPs, while by 2012 more than 180 million SNPs were known. **The International HapMap Project** (www.hapmap.org) systematically catalogues genetic variations based on large-scale SNP analysis in 270 humans of African, Chinese, Japanese and Caucasian origin.

Genome-wide association studies (GWAS) try linking the frequency of SNPs to disease risks

Although GWAS studies have discovered new genes involved in diseases like Crohn's disease or age-related macular degeneration, the typically low risk associated with individual SNPs hampers such correlations, especially in multigenetic diseases. There is much debate how feasible it is to overcome this limitation by examining very large cohorts.

Epigenetic changes are heritable traits not reflected in the DNA sequence

Although the genome as defined by its DNA sequence is commonly viewed as the hereditary material, there are also other heritable traits that are not reflected by changes in the DNA sequence

These traits are called epigenetic changes (see also Chapter 35 and Box on p. 461). They comprise **histone modifications such as acetylation and methylation** that affect chromatin structure. Another modification is **methylation of the DNA** itself, which occurs at the N5 position of cytosines, typically in the context of the sequence CpG. Methylation of CpG clusters, so-

called CpG islands, in gene promoters can shut down the expression of a gene. These methylation patterns can be heritable by a poorly understood process called genomic imprinting (Chapter 35). **Aberrations in gene methylation patterns can cause diseases** and are common in human tumors, often serving to silence the expression of tumor suppressor genes.

The mapping of promoter methylation patterns is very important

The most common methods to analyze DNA methylation rely on the fact that bisulfite converts cytosine residues into uracil, but leaves 5-methylcytosine intact (Fig. 36.5). This change in the DNA sequence can be detected by several methods, including DNA sequencing of the treated versus untreated DNA, differential hybridization of oligonucleotides that specifically detect either the mutated or unchanged DNA, or array-based methods. The latter methods, similar to SNP analysis, also rely on differential hybridization to find bisulfite-induced changes in the DNA, but due to the ability to put millions of oligonucleotide probes on an array, are able to interrogate large numbers of methylation patterns simultaneously. The main limitations are the possibility that bisulfite modification may be incomplete, giving rise to false positives, and the severe general DNA degradation that occurs during the harsh conditions of bisulfite modification. Some new NGS methods can detect DNA methylation directly, and this will accelerate progress in epigenomics. **The epigenome is more variable between individuals than the genome**. Hence it will require a greater effort to map it systematically, but also holds more individual information that can be useful for designing personal medicine approaches.

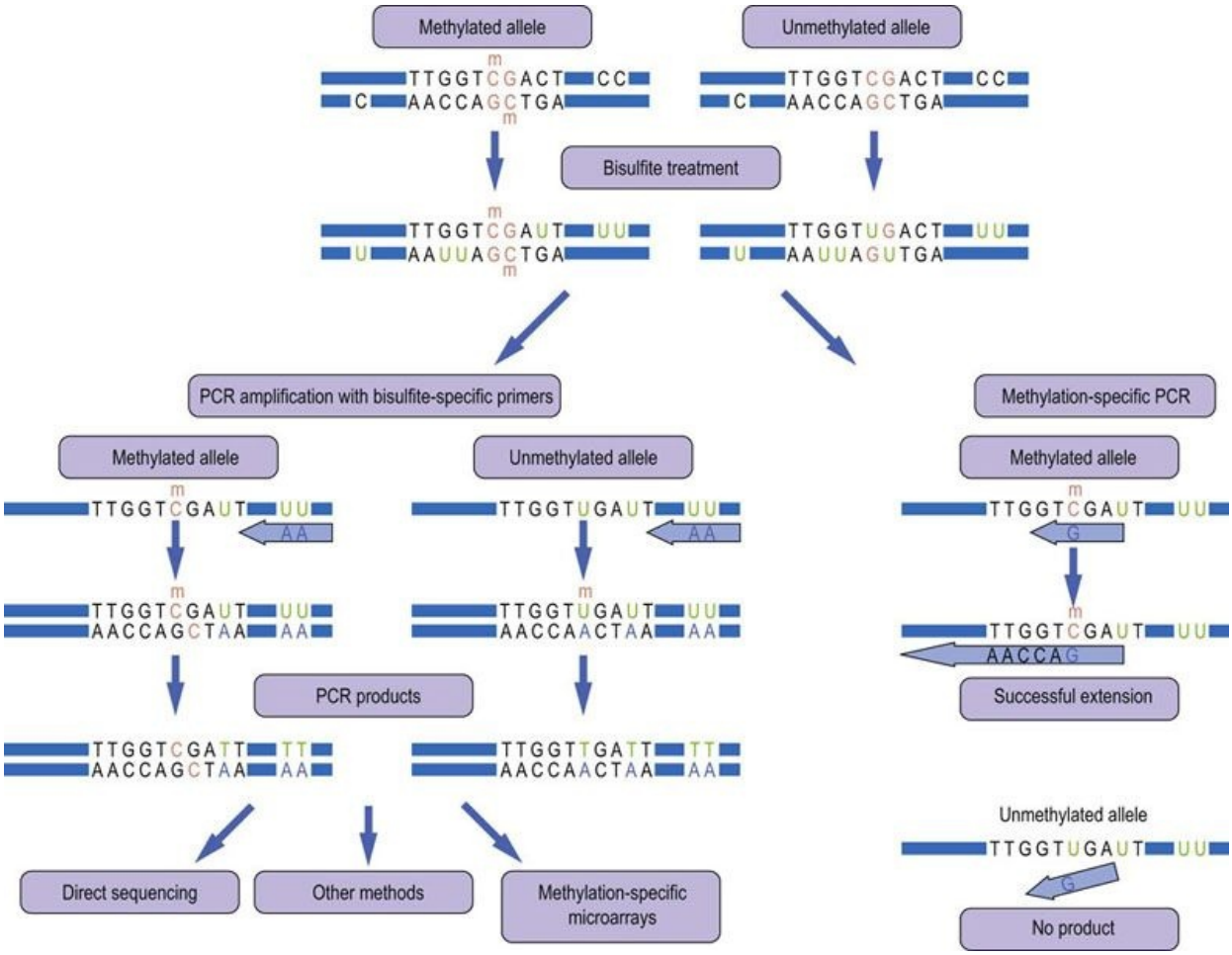


FIG. 36.5 Analysis of DNA methylation.
DNA methylation typically occurs on cytosine in the context of 'CpG islands' (colored orange), which are enriched in the promoter regions of genes. Bisulfite converts cytosine residues to uracil, but leaves 5-methylcytosine residues unaffected. This causes changes in the DNA sequence that can be detected in various ways. Many methods use a PCR amplification step with primers that will selectively hybridize to the modified DNA (left panel). The PCR products have characteristic sequence changes where the unmodified cytosine-guanosine base pairs are replaced by thymidine-adenosine, whereas the original sequence is maintained when the cytosine was methylated. There are many methods to analyze these PCR products. The most common are direct sequencing or hybridization to a microarray that contains oligonucleotides representing all permutations of the expected changes. Another common method is methylation-specific PCR (MSP) where the primer is designed so that it can only hybridize and extend if the cytosine was methylated and hence preserved during bisulfite treatment.

Gene expression and transcriptomics

The transcriptome represents the complement of RNAs that is transcribed from the genome

The transcription of a gene yields an average of 4–6 mRNA variants, which are translated into different proteins. The largest part of the transcriptome consists of noncoding RNAs which fulfill important structural and regulatory functions.

The transcriptome is naturally more dynamic than the genome and may differ widely between different cell types, tissues and different conditions. **Genes represent the DNA sequences that correspond to functionally distinguishable units of inheritance.** This definition goes back to experiments performed by Gregor Mendel, the father of genetics, in the 1860s, who showed that the color of pea plants is inherited as discrete genetic units. About 100 years later Marshall Nirenberg defined a simple relationship, *i.e.* **'Gene makes RNA makes protein'**, which anchored the concept that genes encode the information to make proteins and RNA is the messenger that transports that information (hence the name mRNA). It has turned out that each step is highly regulated and diversified.

Humans possess only approximately 23,000 protein-coding genes, which, however, give rise to an average of 4–6 mRNA transcripts generated by differential splicing, RNA editing and alternative promoter usage

The translation of these mRNAs into proteins (Chapter 34) is also a highly regulated process, so that no direct general correlations between mRNA expression and protein concentrations can be drawn. **Protein-coding genes only constitute 1–2% of the human genome sequence,** and the assumption that most transcripts originate from genes, has recently been superseded by the discovery that more than 80% of the genome can be transcribed. While some of these **noncoding RNAs** serve structural functions, *e.g.* as part of ribosomes, the large majority regulates gene transcription, mRNA processing, mRNA stability and protein translation (see Advanced Concept Box above). **Thus, the largest part of the transcriptome seems dedicated to regulatory functions,** and these regulatory RNAs also can be transcribed from portions of protein-encoding genes. Thus, the concept what constitutes a gene is likely due for revision in the coming years.

Advanced concept box

NONCODING RNAs (ncRNAs)

Noncoding RNAs (ncRNAs) is a summary name for RNAs which do not encode proteins. They comprise abundant species such as **transfer RNAs** and **ribosomal RNAs**, which are involved in protein translation. Several ncRNAs function as molecular guides that participate in processes which require sequence-specific recognition, such as RNA splicing or telomere maintenance. However, the vast majority of ncRNAs seems to have regulatory functions in gene expression. The call to fame came with the award of the Nobel Prize to Andrew Fire and Craig Mello in 2006 '*for their discovery of RNA interference – gene silencing by double-stranded RNA*'. These **small interfering (si) RNAs** are part of an enzyme complex that targets and cleaves mRNAs with high specificity conferred by the siRNA sequence. siRNAs have now become a powerful tool in the arsenal of the molecular biologist to downregulate the expression of selected mRNAs with high specificity and efficiency. **Micro RNAs (miRNAs)** are also small RNAs that are either transcribed under control of their own promoter or often also as part of introns in protein-coding genes. They originate from longer transcripts and are more extensively processed than siRNAs. Functionally, an important distinction is that **siRNAs are very specific**, requiring a perfect match to their targets, while **miRNAs have imperfect sequence recognition** and therefore act upon a larger number of targets, often regulating whole sets of genes. Another difference is that siRNAs induce mRNA degradation, while miRNAs can also prevent mRNA translation. The human genome encodes >1000 miRNAs, which may regulate as much as 60% of genes, thus playing a major role in the control of gene expression. Due to their pleiotropic targeting, **miRNAs** can affect whole programs of gene expression, and aberrant miRNA expression has been implicated in many human diseases, including cancer, obesity and cardiovascular disease.

Gene structure and gene expression

Genes consist of promoter regions, exons and introns. The promoter regulates transcription, while exons are the constituents of mRNAs

Protein-coding genes typically consist of a promoter region that contains the transcriptional start site and participates in the control of transcription, and exons and introns (Fig. 36.6). Introns are removed from the primary transcript by splicing and the mature mRNA only contains the exons. The role of introns is not clear, but they can contain regulatory sequences, including alternative promoters that can modulate gene expression. The usage of **alternative promoters** and **alternative splicing** increase the protein variants that a gene can encode.

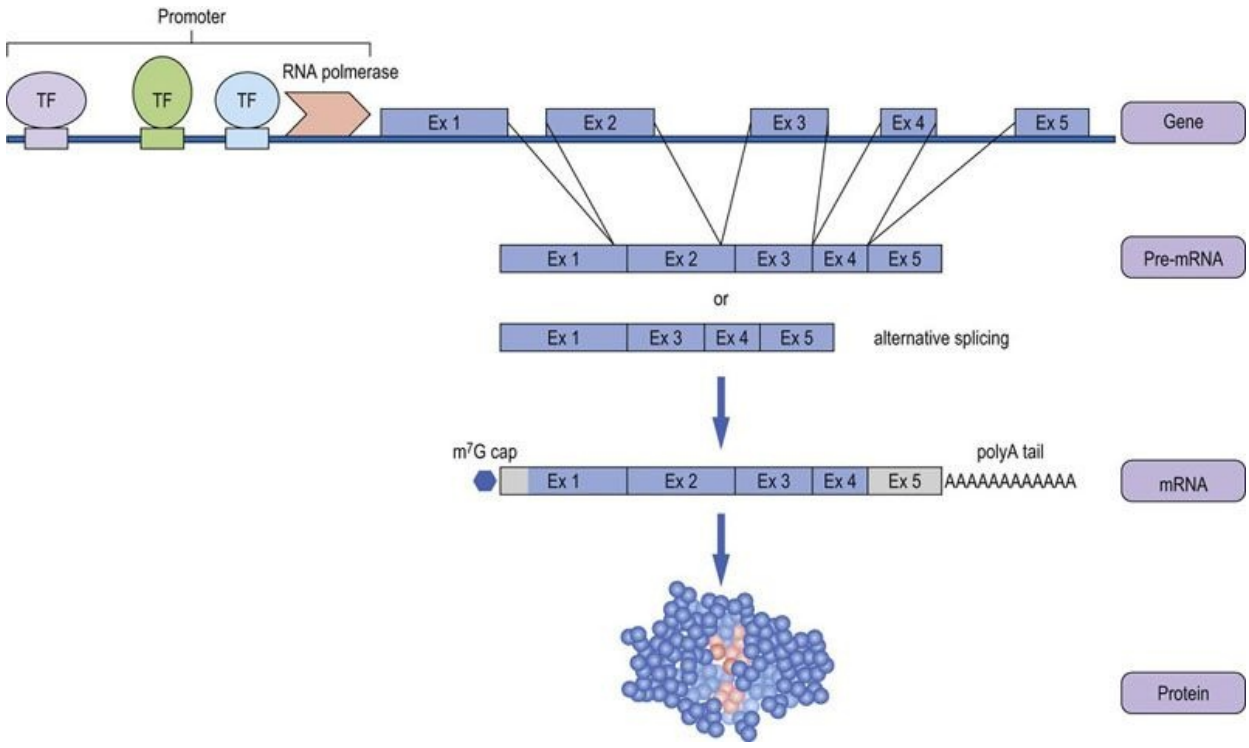


FIG. 36.6 Structure and transcription of a protein-coding gene.
Protein-coding genes are organized in exons and introns. Gene transcription is controlled by transcription factors (TFs), which bind to specific sequence elements in the gene promoter and regulate the ability of RNA polymerase to transcribe, *i.e.* copy, the DNA sequence of the gene into RNA. The transcribed RNA is processed into the mature

messenger RNA (mRNA) by removal of the introns, addition of a polyadenyl tail at the 3'-end and a 7-methylguanosine (m^7G) cap at the 5'-end, which mediates the binding to ribosomes and translation into protein. Untranslated exons and parts of exons are depicted in grey. Most genes produce alternatively-spliced mRNAs where different exons are included.

Some genes produce up to 20 differently spliced mRNAs, with an average estimate of ~5 transcripts per human gene

The further maturation of mRNA involves addition of a **polyA tail at the 3′-end** for nuclear export, and addition of a **7-methylguanylate cap at the 5′-end** which is required for efficient interaction with the ribosome. The mature mRNA regions typically contain other control elements in the untranslated regions that regulate binding to ribosomes and efficiency of translation. Protein translation is explained in Chapter 34. Hence, this chapter only discusses the principles of transcriptional regulation.

Gene promoters typically have two elements: (i) a **binding site for the RNA polymerase complex,** which reads the DNA gene sequence and synthesizes a complementary RNA from it; and (ii) one or more **binding sites for transcription factors** (TFs), which modulate the transcription rate. TFs are modular proteins consisting of separable DNA binding and transactivation domains that regulate the transcriptional activity of RNA polymerase (Chapter 35). The DNA, binding domains recognize short nucleotide sequences. According to their effects on transcription, they are classified as **enhancers** or **repressors**. TFs consisting of only DNA-binding domains act as repressors as they occupy binding sites without being able to stimulate transcription. Conversely, TFs that only contain transactivation domains act as coregulators that rely on interactions with other DNA-binding proteins to regulate transcription. Binding sites that confer responses to specific stimuli or hormones are often called **response elements**, *e.g.* estrogen response element (Chapter 35). Enhancers and repressors are often located in the gene promoter, but also in introns or even far upstream or downstream of a gene as they can act over long distances. These remote effects are mediated by the yet ill-understood three-dimensional arrangement of DNA that can spatially colacate factors bound to distant sites on a linear DNA molecule.

TFs can affect transcription directly by controlling the

function of RNA polymerase, or indirectly by affecting the chromatin structure

Transcription can only proceed when histones are acetylated and chromatin is in the open conformation permissive for transcription. Thus, **many TFs associate with histone-modifying enzymes** that change chromatin structure **to** either facilitate or repress transcription. TF function itself is controlled by intricate **cellular signal transduction networks** that process external signals, such as stimulation by growth factors, hormones and other extracellular cues, into changes of TF activities. Such changes are often conferred by phosphorylation, which can regulate the nuclear localization of TFs, their ability to bind to DNA, or their regulation of RNA polymerase activity. The presence of different TF-binding sites in a gene promoter further confers combinatorial regulation. As **TFs represent more than 10% of all human genes**, this multilayered type of control ensures that gene transcription can be finely adjusted in a highly versatile way to specific cellular states and environmental requirements.

Studying gene transcription by gene (micro)arrays and RNA sequencing

Methods for studying global transcription are now well established. **Gene (micro)arrays** contain several million DNA spots arranged on a slide in a defined order (Fig. 36.7). Modern arrays use synthetic oligonucleotides, which can be either prefabricated and deposited on the chip or synthesized directly on the chip surface. Usually several oligonucleotides are used per gene. They are carefully designed based on genome sequence information to represent unique sequences suitable for the unambiguous identification of specific RNA transcripts. Today's high-density arrays contain enough data points to survey the transcription of all human genes, map exon content and splice variants of mRNAs. Noncoding RNAs such as siRNAs and miRNAs also can be included. **The arrays are hybridized with complementary RNA (cRNA) probes** corresponding to the RNA transcripts isolated from the cells or tissues that are to be compared. The probes are made from isolated RNAs by copying them first into cDNA using reverse transcriptase, a polymerase that can synthesize DNA from RNA templates. The resulting complementary DNA (cDNA) is transcribed back into cRNA, as RNA hybridizes stronger to the DNA oligonucleotides on the array than cDNA would. During cRNA synthesis, modified nucleotides are

incorporated that are labeled with fluorescent dyes or tags, such as biotin, which can easily be detected after hybridization of the cRNA probes to the array. After hybridization and washing off unbound probes, the array is scanned and the hybridization intensities are compared using statistical and bioinformatic analysis. The results allow a relative quantitation of changes in transcript abundances between two samples or different timepoints. Thanks to a common convention for reporting microarray experiments, called Minimal Information for the Annotation of Microarray Experiments (MIAME), array results from different experiments can be compared, and public gene array databases are a valuable source for further analysis. Gene array analysis is already being used for clinical applications. For instance, patterns of gene transcription in **breast cancers** have been developed into tests for assessing the risk of recurrence and the potential benefit of chemotherapy.

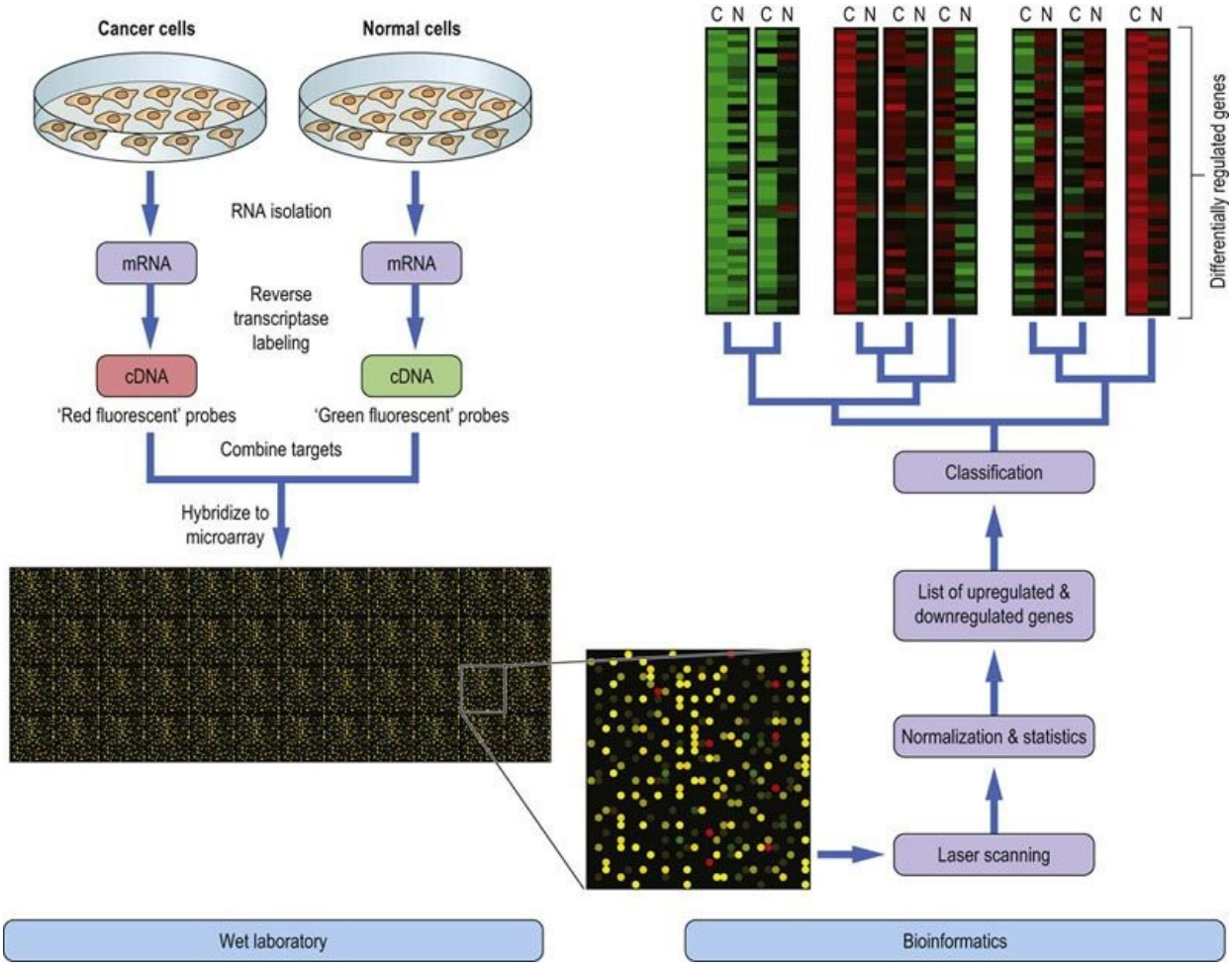


FIG. 36.7 The workflow of a gene (micro)array experiment.
A two-color array experiment comparing normal and cancer cells is shown as an example. See text for details. A common way to display results is heatmaps, where increasing intensities of reds and greens indicate up-and downregulated genes, respectively, while black means no change. C, cancer cells; N, normal cells. The figure is modified from http://en.wikipedia.org/wiki/DNA_microarray.

Transcriptome analysis also can be performed by **direct sequencing**, once the RNAs have been converted to cDNAs. The advances in rapid and cheap DNA sequencing methods permit every transcript to be sequenced multiple times. These **'deep sequencing' methods** not only unambiguously identify the transcripts and splice forms but also allow the direct counting of transcripts over the whole dynamic range of RNA expression, resulting in absolute transcript numbers rather than relative comparisons. Thus, the sequencing methods, dubbed **RNA seq**, are quickly becoming attractive alternatives to array-based transcriptomics methods.

ChIP-on-chip technique combines chromatin immunoprecipitation with microarray technology

Mapping of the occupancy of transcription-factor-binding sites can reveal which genes are likely to be regulated by these factors

Our ability to survey the transcription of all known human genes poses the question which TFs are controlling the observed transcriptional patterns. The human genome contains many thousands of binding sites for any given TF, but only a small fraction of these binding sites actually is occupied by TFs and involved in the regulation of gene transcription. Thus, **the systematic mapping of the occupancy of TF-binding sites can reveal which genes are actually regulated by which TFs**. The techniques developed for this (Fig. 36.8) combine **chromatin immunoprecipitation (ChIP)** with **microarray technology (chip)** or DNA sequencing, and are called **ChIP-on-chip or ChIP-seq**.

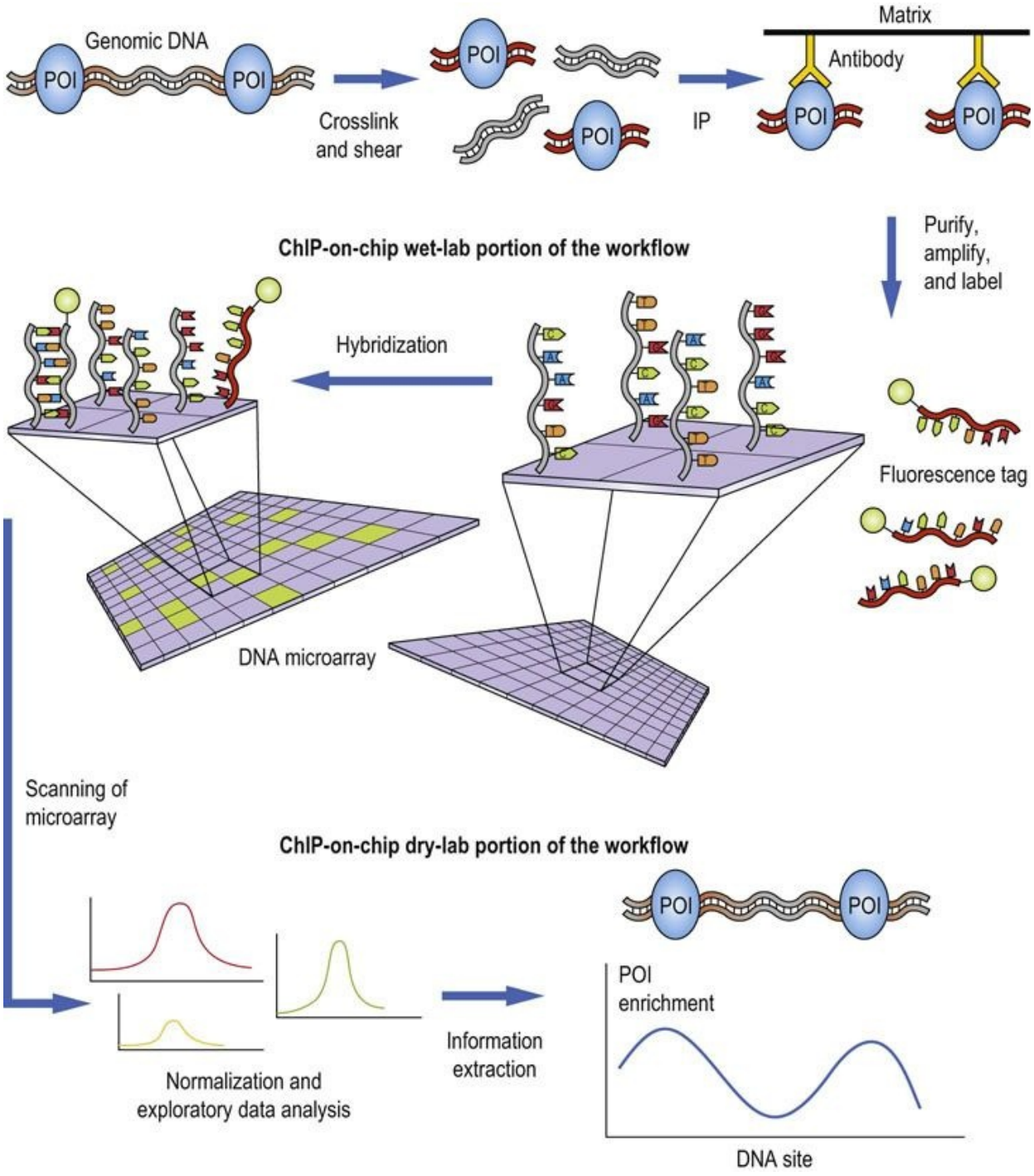


FIG. 36.8 ChIP-on-chip analysis.
See text for details. POI (protein-of-interest). The figure has been modified from http://en.wikipedia.org/wiki/Image:ChIP-on-chip_wet-lab.png.

ChIP involves the covalent crosslinking of proteins to the DNA they are bound to by formaldehyde treatment of living cells. Then, the DNA is purified

and fragmented into small (0.2–1 kb) pieces by ultrasound sonication. These DNA fragments are isolated by immunoprecipitating the crosslinked protein with a specific antibody. The associated DNA is then eluted and identified by PCR with specific primers that amplify the DNA region one wants to examine. This method assesses one binding site at a time and requires a hypothesis suggesting which site(s) should be examined. The identification of the associated DNA, however, can be massively multiplexed by using DNA microarrays for detection that represent the whole or large parts of the genome. Similarly, as discussed above, as alternative to gene microarrays, the associated DNA can be identified by sequencing.

ChIP-on-chip and ChIP-seq are powerful and informative techniques allowing the correlation of TF binding with transcriptional activity. **The ChIP techniques can be used to study any protein that interacts with DNA,** including proteins involved in DNA replication, DNA repair and chromatin modification. Success is critically dependent on the quality and specificity of the antibodies used, as the amounts of co-immunoprecipitated DNA are very small, and there is no other separation step than the specificity provided by the antibody.

Proteomics

Proteomics is the study of the protein complement of a cell; the protein equivalent of the transcriptome or genome

The word ‘proteome’ was coined by Marc Wilkins in a talk in Siena in 1994. Wilkins defined the proteome as the protein complement of a cell, the protein equivalent of the transcriptome or genome. Since that time the study of the proteome, called proteomics, has evolved into a number of different themes encompassing many areas of protein science.

Proteomics is possibly the most complex of all the -omics sciences but is also likely to be the most informative, since proteins are the functional entities in the cell, and virtually no biological process takes place without the participation of a protein. Among their many roles, they are responsible for the structural organization of the cell, as they make up the cytoskeleton, the control of membrane transport (Chapter 8) and energy generation. Hence, an understanding of the proteome will be necessary to understand how biology works.

Initially, proteomics concentrated on cataloguing the proteins contained in an organelle, cell, tissue or organism, in the process validating the existence of the predicted genes in the genome. This rapidly evolved into comparative proteomics, where the protein profiles from two or more samples were compared to identify quantitative differences that could be responsible for the observed phenotype: for example, from diseased versus healthy cells or looking at changes induced by drug treatment. Now, proteomics also includes the **study of post-translational modifications** of individual proteins, the make-up and dynamics of **protein complexes**, the mapping of networks of **interactions between proteins**, and the identification of **biomarkers in disease**. Quantitative proteomics has become a robust tool, and even absolute quantification is relatively routine now.

Proteomics poses several challenges

It quickly became apparent that the complexity of the proteome would be a major obstacle to achieving Wilkins' initial ideal of looking at all the proteins in a cell or organism at the same time.

While the number of genes in an organism is not overwhelming, the post-

translational modifications (PTMs) of proteins in eukaryotic systems, such as alternate splicing and the potential addition of over 40 different covalently attached chemical groups (including the well-known examples of phosphorylation and glycosylation), mean that there may be 10 or, in extreme cases, 1000 different protein species, all fairly similar, generated from each gene, and that the predicted 23,000 genes in the human genome could give rise to 500,000 or more individual protein species in the cell. In addition, there is a wide range of protein abundances in the cell, estimated to range from less than 10 to 500,000 or more molecules per cell, and a protein's function may depend on its abundance, PTMs, localization in the cell, and association with other proteins, and these may all change in a fraction of a second!

There is no protein equivalent of PCR that would allow for the amplification of protein sequences, so we are limited to the amount of protein that can be isolated from the sample

If the sample is small, *e.g.* a needle biopsy, a rare cell type, or an isolated signaling complex, ultrasensitive methods are needed to detect and analyze the proteins.

It is clear that proteomics is an extremely challenging undertaking. It is only since the introduction of new methods in **mass spectrometry** in the mid 1990s that an attempt could be made to analyze the proteome and new, higher-throughput and high-data content methods are being continually developed. The proteomes of prokaryotic, *e.g. Mycoplasma*, or simple eukaryotic species, *e.g.* yeast, have been deciphered in terms of identifying expressed proteins and many of their interactions. Even the complement of human proteins expressed in cell lines has been determined, but we are still far away from being able to identify all protein variants and PTMs.

Advanced concept box Post-translational modifications

During the process of transcription, translation and in the functioning of the cell, proteins can undergo a range of modifications. During transcription, introns are spliced out of the gene, and different splicing of the gene can result in a number of different mRNAs being produced, and hence a number of proteins that differ markedly in their sequence can emerge from the same

gene. After translation of the mRNA into protein, the protein can be 'decorated‘ with a bewildering array of additional chemical groups covalently attached to it, many of which regulate the activity of the protein. Some examples are given below:

■ **The addition of fatty acids** to cysteine residues, which anchor the protein to a membrane.

■ **Glycosylation:** the addition of complex oligosaccharides to an asparagine or serine residue, which is common in membrane proteins that have an extracellular component or are secreted. Many proteins involved in cell–cell recognition events are glycosylated, as are antibodies.

■ **Phosphorylation:** the addition of a phosphate group to serine, threonine, tyrosine or histidine residues. This is a modification that can be added or removed, allowing the system to respond very rapidly to a changing environment. It is fundamental to signaling events in the cell. It has been estimated that one-third of all eukaryotic proteins may undergo reversible phosphorylation.

■ **Ubiquitination:** the addition of a polyubiquitin chain that targets the protein for destruction by the proteasome. Ubiquitination also can regulate enzyme activities and subcellular localization. Ubiquitin is itself a small protein.

■ **Formation of disulfide bridges** between cysteine residues in the polypeptide backbone which are close together in space once the protein is folded. These play a number of roles, including adding additional structural stability, especially for exported proteins, and sensing the redox balance in the cell.

■ **Acetylation** of residues, most commonly the N-terminus of the protein or lysine. Acetylation of lysines on histones plays an important role in the gene transcription process, and drugs that target the proteins that acetylate or deacetylate histone are potential cancer therapeutics.

■ **Proteolytic cleavage:** most proteins have the N-terminal methionine removed that results from the ATG initiation codon of gene translation. In some proteins, cleavage of the polypeptide chain occurs, such as in the activation of zymogens in the clotting cascade, or significant parts of the initial polypeptide chain are removed completely, for example in the conversion of proinsulin

into insulin.

Proteomics in medicine

Despite the challenges, proteomics is a powerful tool in understanding fundamental biological processes, and has become well established

Like the other -omics technologies it has the advantage that it is possible to discover new information about a biological problem without having to have a clear understanding in advance of what might change. There are often more data generated from a good proteomics experiment than it is reasonable, or possible, to follow up.

Proteomics has been applied successfully to the study of basic biochemical changes in many different types of biological sample: cells, tissues, plasma, urine, cerebrospinal fluid (CSF) and even interstitial fluid collected by microdialysis

In cells isolated from cell culture, it is possible to ask complex fundamental biological questions. Deciphering the mitogenic signaling cascades, which involve specific association of proteins in multiprotein complexes, and understanding how these can go wrong in cancer is one widely studied area. It is possible to gain information from biological fluids on the overall status of an organism because, for example, blood would have been in contact with every part of a body. Diseases at specific locations may eventually show up as changes in the protein content of the blood, as leakage from the damaged tissue occurs. This area is now often described as **biomarker discovery**. Tissues are a more of a challenge. The heterogeneity of many tissues makes it difficult to compare tissue biopsies which may contain differing amounts of connective tissue, vasculature, *etc*. Improvements in the sensitivity of analysis are now being overcome by allowing small amounts of material recovered from tissue separation methods, such as laser capture microdissection or flow cytometry, to be used for the analysis. There is much effort being directed towards the ultimate

challenge; the analysis of individual cells. This is valuable, since current approaches average out changes in the analyzed sample and we lose all information on natural heterogeneity in biology: for instance, a recorded 50% change in the level of a protein could be 50% in all cells, or 100% in 50% of the cells in the sample.

Main methods used in proteomics

Proteomics relies on the separation of complex mixtures of proteins or peptides, quantification of protein abundances, and identification of the proteins

This approach is multi-step but modular, which is reflected in the many combinations of separation, quantitation and identification. Here, we focus on highlighting the principles rather than trying to be comprehensive.

Protein separation techniques

Strategies for protein separation are driven by the need to reduce the complexity, *i.e.* the number of proteins being analyzed, while retaining as much information as possible on the functional context of the protein, which includes the subcellular localization of the protein, its incorporation in different protein complexes, and the huge variety of PTMs. No method can reconcile all these requirements. Therefore, different methods were developed that exploit the range of physicochemical properties of proteins (size, charge, hydrophobicity, PTMs, etc.) for separating complex mixtures (Fig. 36.9).

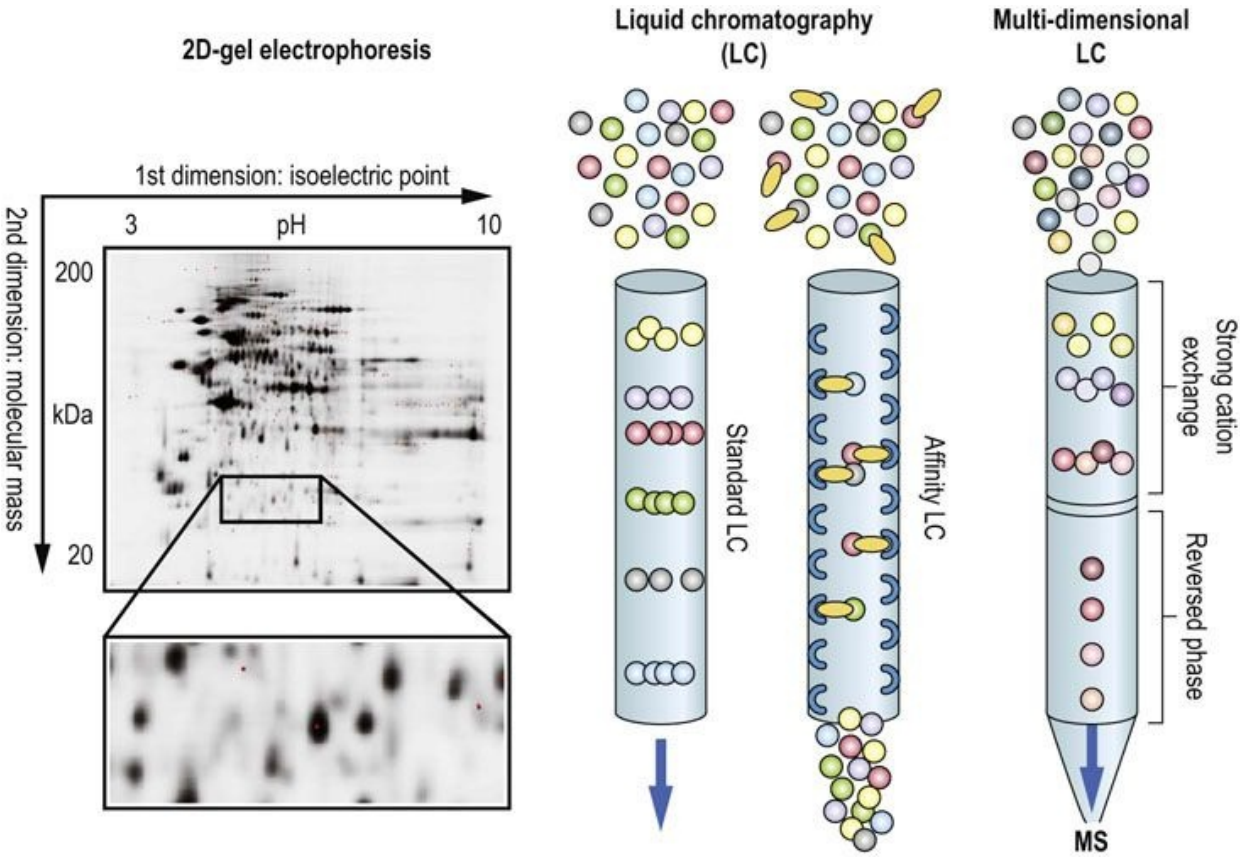


FIG. 36.9 Protein and peptide separation techniques.
The **left panel** shows a 2D gel where protein lysates were separated by isoelectric focusing in the 1st dimension, and according to molecular weight in the 2nd dimension. Protein spots were visualized with a fluorescent stain. The **middle panel** illustrates the principles of LC, where proteins or peptides are separated by differential physicochemical interactions with the resin as the flow through the column. A variation is affinity LC, where the resin is modified with an affinity group that retains molecules which selectively bind to these groups. The **right panel** demonstrates the setup for multidimensional LC, where a strong cation exchange column is directly coupled to a reversed-phase column enabling a 2-step separation by hydrophilicity and hydrophobicity. The eluate can be directly infused into a mass spectrometer (MS) for peptide identification.

A classic protein separation method is two-dimensional polyacrylamide gel electrophoresis (2DE, 2D-PAGE)

Proteins are separated by isoelectric focusing according to their electrical charge in the first dimension, and according to their size in the second dimension. Labeling the proteins with fluorescent dyes or using fluorescent stains made the method quantitative, but protein spots have to be picked from the gels

individually for subsequent identification by mass spectrometry (MS).

Therefore, 2D-PAGE is now much less frequently used, and has been replaced by **liquid chromatography** (LC), which can be directly coupled to **mass spectrometry** (MS). Thus, molecules eluting from the chromatographic column can be measured and identified in real time. As for technical reasons MS-based identification works better with smaller molecules, proteins are digested with proteases (usually trypsin) into small peptides before MS analysis. LC separates proteins or peptides on the basis of different physicochemical properties, most commonly the charge of the molecule or its hydrophobicity, using ion exchange or reversed-phase chromatography, respectively. This is achieved by having chemical groups attached to a particulate resin packed into a column and flowing a solution over this. Molecules will bind to the resin (the stationary phase) with differing affinities. Those with a high affinity will take longer to traverse the length of the column and hence will elute from the column at a later time. Molecules are therefore separated in time in the effluent that elutes off the column. Affinity chromatography uses specialized resins that strongly bind to certain chemical groups or biological epitopes and retain proteins carrying these groups. For instance, resins containing chelated Fe^{3+} or TiO_2 (**immobilized metal affinity chromatography,** IMAC) bind phosphate and are used to select phosphorylated peptides. LC also can be carried out in two dimensions. Adding a **strong cation exchange (SCX) chromatography** step before IMAC removes many non-phosphorylated peptides, enhancing the enrichment of phosphosphopeptides in the IMAC step.

The first 2D LC method with direct coupling of the two dimensions is called multidimensional protein identification technology (MudPIT)

In MudPIT the total protein content of the sample is first digested with trypsin, and the resulting peptides are fractionated by an SCX column, which separates peptides according to charge. Then, the peptide fractions are further separated by reversed-phase LC and directly injected into the MS. Modern fast-scanning, high-resolution MS instruments coupled to high-resolution separation LC now make it possible to dispense with the first dimension for all but the most complex samples. This method of MS-enabled, peptide-based protein identification is often referred to as **‘shotgun proteomics’**.

Protein identification by mass spectrometry (MS)

Mass spectrometry is a technique used to determine the molecular masses of molecules in a sample

MS can also be used to select an individual component from the mixture, break up its chemical structure and measure the masses of the fragments, which can then be used to determine the structure of the molecule. There are many different types of mass spectrometers available, but the underlying principles of mass spectrometry are relatively simple. The first step in the process is to generate charged molecules, ions, from the molecules in the sample. This is relatively easily achieved for many soluble biomolecules because their polar chemistry provides groups that are easily charged. For example, the addition of a proton (H^+) to the side-chain groups on the basic amino acids lysine, arginine or histidine gives a positively charged molecule. When a charged molecule is placed in an electric field, it will be repelled by an electrode of like sign and attracted by an electrode of opposite sign, accelerating the molecule towards the electrode of opposite charge. Since the force is equal for all molecules, larger molecules will accelerate less than small molecules (force = mass × acceleration), so small molecules will acquire a higher velocity. This is utilized to determine the mass. For example, after the molecules have been accelerated, the time then taken for them to travel a certain distance can be measured and related to the mass. This is called **time-of-flight mass spectrometry**.

A tandem mass spectrometer is effectively two mass spectrometric analyzers joined together sequentially, with an area between them where molecules can be fragmented

The first analyzer is used to select one of the molecules from a mixture based on its molecular mass, which is then broken up into smaller parts, usually by collision with a small amount of gas in the intermediate region (called the collision cell). The fragments that are generated are then analyzed in the second mass spectrometer (Fig. 36.10). As peptides tend to fragment at the peptide bond, the fragment peaks are separated by the masses of the different amino acids in the corresponding sequence. This result is, in principle, similar to the Sanger method of DNA sequencing, allowing the peptide sequence to be deduced. However, in contrast to Sanger DNA sequencing, peptide fragmentation is not uniform, and the spectra usually only cover part of the

sequence, leaving gaps and ambiguous sequence reconstruction. Therefore, the **peptide sequence is predicted based on a statistical matching of observed masses against a virtual digest and peptide fragmentation of proteins in a database** (Fig. 36.11). With today's highly accurate MS and well-annotated databases, these computational sequence predictions are very reliable. It also demonstrates that proteomics is intimately reliant on the quality and completeness of genome sequencing, and genome databases that are used to infer the encoded proteins.

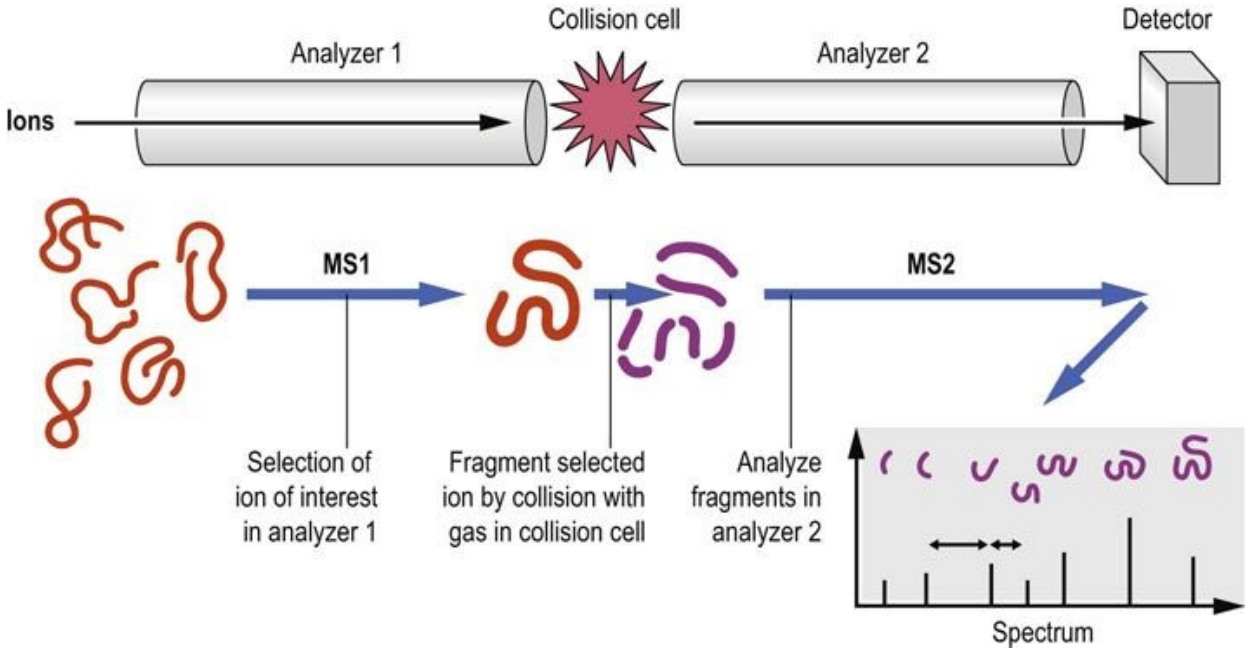


FIG. 36.10 The basic principles of tandem mass spectrometry. See text for details. MS, mass spectrometer.

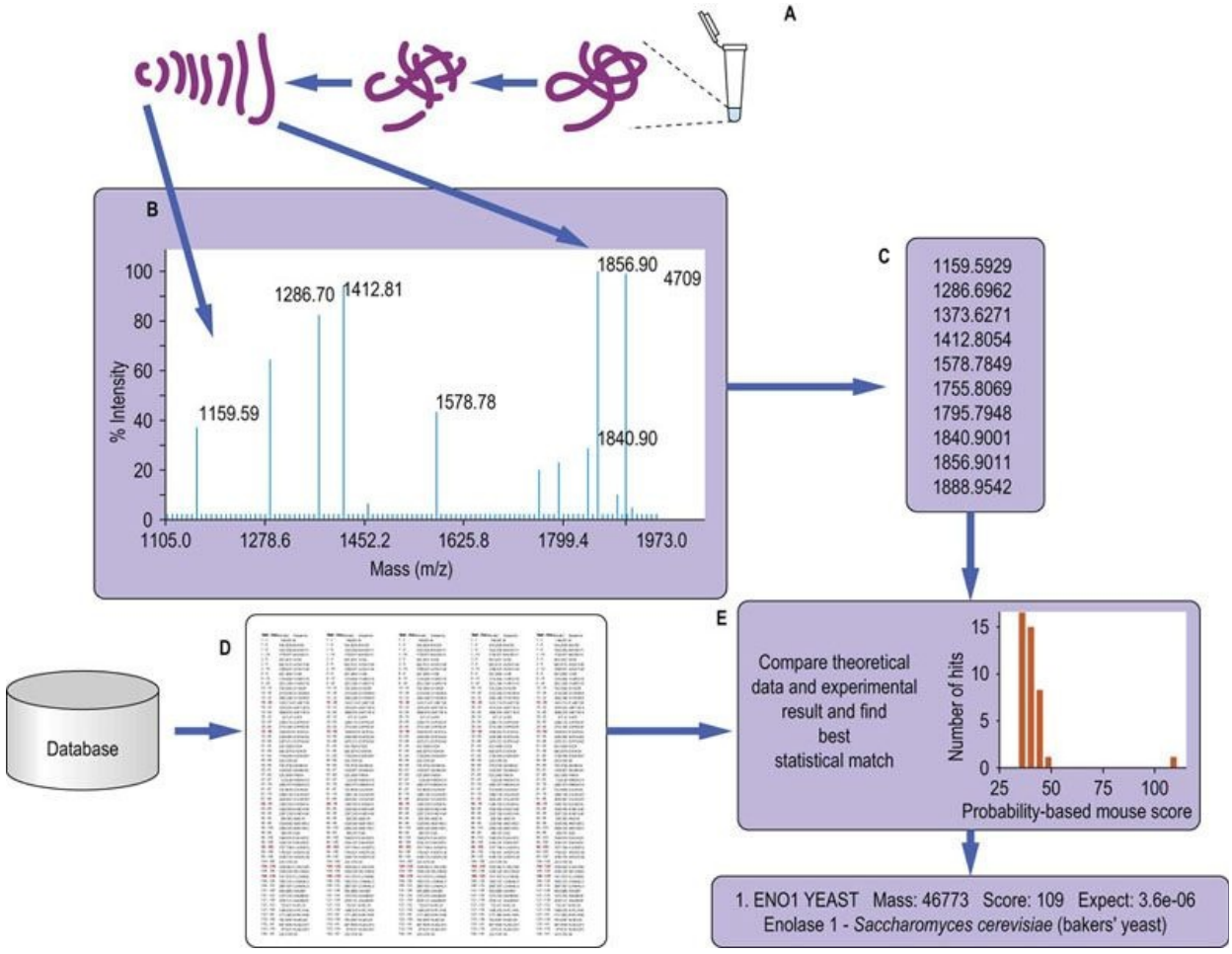


FIG. 36.11 Protein identification by mass spectrometry.
A typical workflow: **(A)** the sample is digested with a specific protease, usually trypsin, to give a set of smaller peptides which will be unique to the protein; **(B)** the mass of a subset of the resulting peptides is measured using MS; in tandem MS each peptide is fragmented and the mass of the fragments is measured as well; **(C)** a list of the observed experimental masses is generated from the mass spectrum; **(D)** a database of protein sequences is theoretically digested (and fragmented in the case of tandem MS) *in silico*, and a set of tables of the expected peptides generated; **(E)** the experimental data are compared to the theoretical digested database and a statistical score of the fit of the experimental to theoretical data is generated, giving a 'confidence' score, which indicates the likelihood of correct identification.

MS measures and fragments peptides as they elute from the LC, resulting in abundant proteins being identified many times, whereas low-abundant proteins are overlooked if the MS is overwhelmed by a flow of abundant peptides. This is a main reason why protein or peptide pre-fractionation increases the number of successfully identified proteins. This workflow requires measuring the whole protein complement even when one only is interested in a few specific proteins.

To enable the targeted identification of specific proteins, a technique was developed, called selected reaction monitoring (SRM) or multiple reaction monitoring (MRM)

This method uses MS1 to select a peptide ion out of a mixture, then fragment it and select defined fragment masses for detection in MS2 (Fig. 36.12). A software protocol for MS1 peptide selection and MS2 fragment detection gives unique protein identifications based on the measurement of a few selected peptides. This is a powerful method to streamline the identification of proteins from complex samples by systematically monitoring only the most informative peptide fragmentations. The peptide atlas (www.peptideatlas.org/) is a database of such informative fragments, and will greatly facilitate the systematic analysis of proteomes and sub-proteomes.

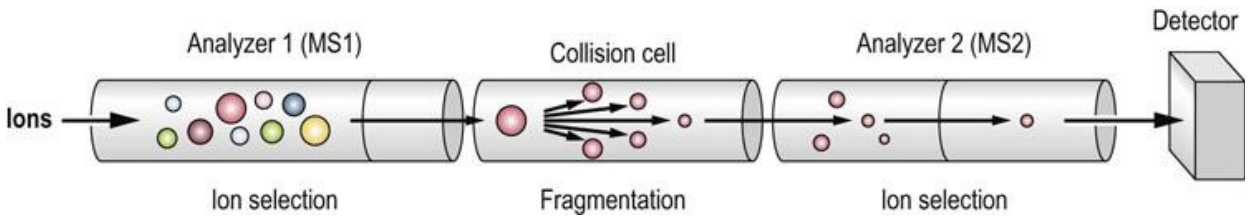


FIG. 36.12 The principles of selected reaction monitoring (SRM) or multiple reaction monitoring (MRM) experiments.
See text for details.

Quantitative mass spectrometry

MS can be made quantitative in a number of ways. If possible, samples can be grown in a selective medium that provides an essential amino acid in the natural form (the 'light' form) or isotopically labeled with a stable isotope (for example, ^{13}C or ^{2}H, the 'heavy' form) that makes all of the peptides containing this amino acid appear heavier in the mass spectrometer. This is called the **stable isotope labeling with amino acids in cell culture** (SILAC) method, and is one of the most widely used and robust of the labeling technologies (Fig. 36.13). The samples are mixed together and analyzed using the shotgun approach. The ratios of 'heavy' to the equivalent 'light' peptides are used to determine the relative quantities of the protein from which they were derived.

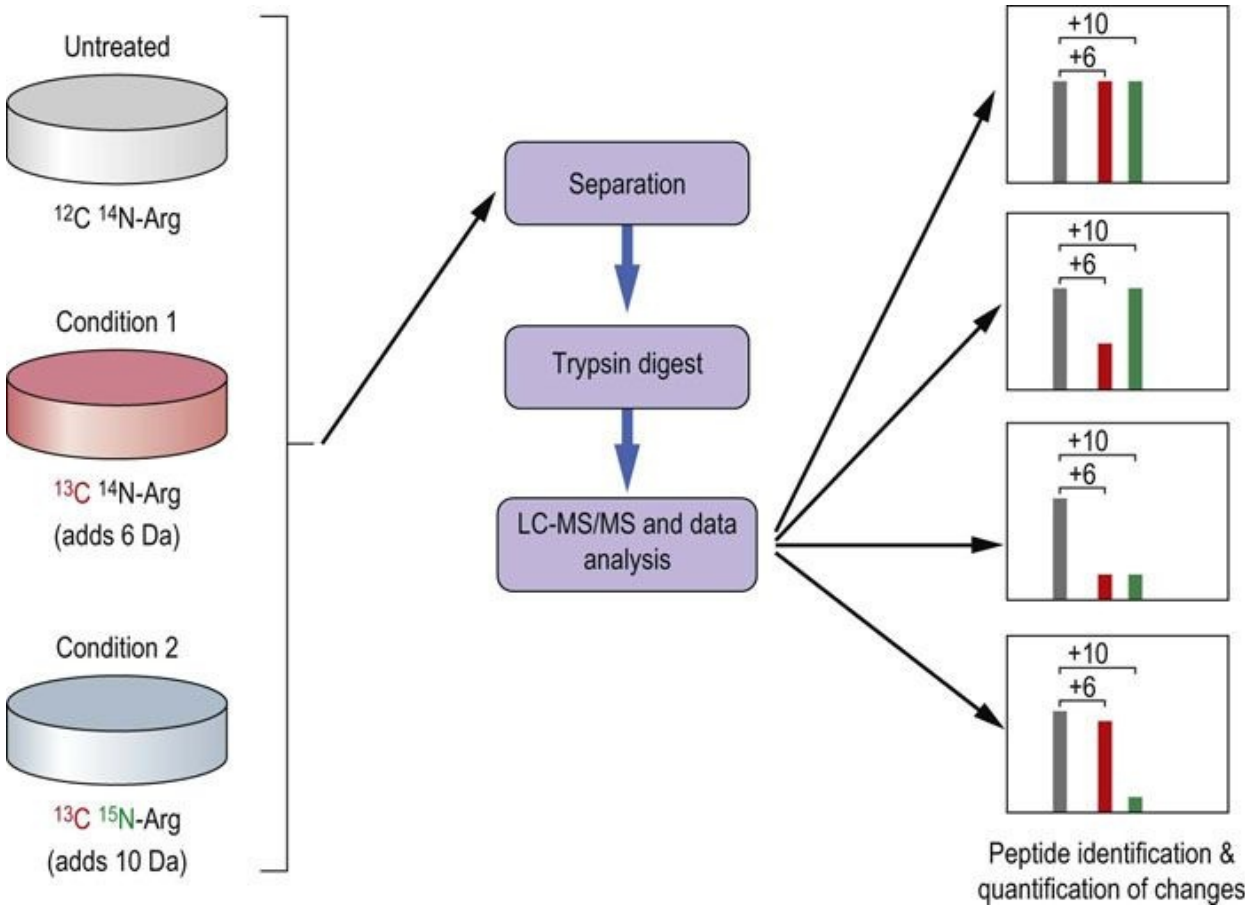


FIG. 36.13 Stable isotope labeling with amino acids in cell culture (SILAC) for quantitative mass spectrometry.
See text for details.

Alternative methods include chemically reacting the proteins in the sample (using, for example, the **isotope coded affinity tags, ICAT)**, or the peptides after digestion of the sample (e.g. **in isobaric tags for relative and absolute quantitation, iTRAQ)**, with a 'light' or equivalent isotopically labeled 'heavy' chemical reagent, then mixing the samples and analyzing them as for the SILAC approach. Direct comparison of 1D-LC runs based on the normalized signal intensity without the need for labeling is also possible due to improvements in the reproducibility of LC and software. In addition, counting of peptide ions in the mass spectrometer have led to so-called **label-free quantitation methods** that are rapidly improving and soon may allow accurate quantification without the need to label cells or proteins. The advantage of these methodologies is that the analysis is easily automated and that approaches can be used to get information on proteins that do not work well in the 2DE approach, such as

membrane proteins, small proteins and proteins with extreme pIs (e.g. histones). The disadvantage is that information on post-translational modifications is usually lost, and digesting the sample generates a much more complex sample for the separation step.

Non-mass spectrometry based technologies

While MS remains a main stay technique used for proteomics, a variety of other methods are becoming established. **Protein microarrays** are conceptually similar to those used for transcriptomics. They come in three versions (Fig. 36.14). In the **reverse phase protein array** (RPPA) lysates of cells or tissues are spotted on a microscope slide with a protein-friendly coating. These arrays are then probed with an antibody specific to a protein or a certain PTM. After washing to remove unbound antibodies, successful binding events are visualized by a secondary anti-antibody antibody, which carries a detectable label, usually a fluorescent dye. Thereby, a large number of samples or treatment conditions can be compared simultaneously. The success of this method is completely dependent on the specificity of the antibody, and constrained by the limited availability of high-quality mono-specific antibodies. In the **capture array**, antibodies are deposited onto the array, which then is incubated with a protein lysate. Detection of the captured proteins is by another antibody. Thus, the overall specificity is the overlap between the specificities of the capturing and detecting antibodies, ameliorating the limitation that each antibody should be absolutely specific. **Target arrays** contain a single species of purified protein in each spot. These arrays are used to find binding partners for specific proteins. They can be probed with another purified protein, or with a mixture of antibodies, *e.g.* from patient sera, to determine whether a patient has antibodies against particular proteins. Protein microarrays can be used to quantify the amount of the protein present in a sample and thus lend themselves for clinical diagnostics.

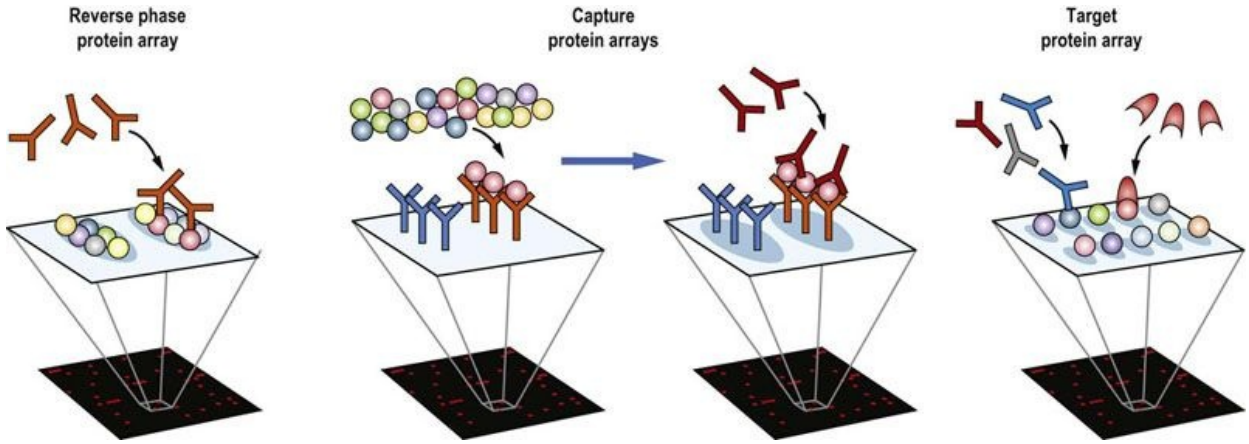


FIG. 36.14 Protein (micro)arrays.
See text for details.

The Human Protein Atlas (HPA) (www.proteinatlas.org/) aims to generate antibodies to every protein in the human proteome, and use these to visualize proteins and their subcellular localization in healthy and diseased human tissues

In 2012 the HPA comprised more than 14,000 proteins, *i.e.* approximately 70% of gene products, if splice forms and other variants are neglected. Efforts to include protein variants and post-translational modifications are underway. Thus, the HPA is becoming a major resource for proteome analysis.

Microscopy has also become a tool that is frequently used in spatial proteomics, to assess where proteins are localized in the cell and how this changes under different conditions. This has been enabled by the advances in the intracellular expression of proteins that are a fusion between the proteins of interest and green fluorescent protein (GFP) or its analogues. The cellular location of the protein can then be tracked by microscopy by following the fluorescent signal of the protein attached to it. There are now analogues of GFP that emit at a wide range of wavelengths, meaning that three, or even four, proteins can be followed in parallel.

Advanced concept box Nuclear magnetic resonance spectroscopy

Nuclear magnetic resonance (NMR) spectroscopy gives useful

structural information on molecules that can be used to identify them. Atomic nuclei behave like small magnets, so when they are put in a strong magnetic field they align with the field. Application of an appropriate energy (radiofrequency electromagnetic radiation) causes the nuclei to flip and align against the field. They then return to their ground state once the irradiation is turned off by flipping back, and as they do so they emit specific frequencies of radiation. These can be recorded and plotted. Each nucleus in a molecule that has a unique environment will emit a unique frequency, and nuclei bonded together or close together in space will interact with adjacent nuclei (coupling), and this can also be measured. This rich information on the molecule allows the structural elements to be determined, and the amplitude of the signals can be used to reasonably accurately quantify the amount of material. This is very useful in **metabolomics.** The main limitation is that the NMR spectrum quickly becomes congested with information from a complex sample, so high resolution (coming from very strong magnetic fields) is required, and the technique is relatively insensitive, having a limit of detection 3–4 orders of magnitude worse than mass spectrometry.

Metabolomics

Metabolites are the small chemical molecules, such as sugars, amino acids, lipids and nucleotides, present in a biological sample. The study of the metabolite complement of a sample is called **metabolomics**, while the quantitative measurement of the dynamic changes in the levels of metabolites as a result of a stimulus or other change is often referred to as **metabonomics**. The words metabolomics and metabonomics are often used interchangeably, although purists will claim that while both involve the multiparametric measurement of metabolites, metabonomics is dedicated to the analysis of dynamic changes of metabolite levels, whereas metabolomics focuses on identifying and quantifying the steady-state levels of intracellular metabolites. Metabolomics is the most commonly used generic term.

Metabolomics gives another level of information on a biological system

It provides information on the results of the activity of enzymes, which may not depend on the abundance of the protein alone, as this may be modulated by supply of substrates, the concentration of cofactors or products, and effect of other small molecules or proteins that modulate the activity of the enzyme (effectors). In some ways, metabolomics may be easier to perform than proteomics. In the metabolome, there is an amplification of any changes that occur in the proteome, as the enzymes will turn over many substrate molecules for each molecule of enzyme. The methods used to look for a metabolite in each organism will be the same, as many of the metabolites will be identical, unlike proteins, whose sequences are much less conserved between organisms. Thus, metabolic networks are more constrained, making them easier to follow.

However, the analysis of the metabolome is still complex as it is very dynamic, many metabolites give rise to a number of molecular species by forming adducts with different counter ions, and xenobiotics, molecules that do not come from the host but from foodstuffs, drugs, the environment or even the microflora in the gut, complicate the analysis greatly; the actual metabolome may be getting close to being as complicated as the proteome.

In a similar way, **lipidomics** has become a topic in its own right, studying the dynamics changes in lipids in diverse functions such as membranes, lipoproteins,

and as signaling molecules. In 2007, the Human Metabolome Project released the first draft of the human metabolome consisting of 2500 metabolites, 3500 food components and 1200 drugs. Currently, there is information on approximately 20,000 metabolites, approximately 1600 drug and drug metabolites, 3100 toxins and environmental pollutants, and around 28,000 food components.

The most commonly used methods for investigating metabolites are mass spectrometry, often coupled to LC, as used in proteomics, and **nuclear magnetic resonance (NMR) spectroscopy**. Identification of signals corresponding to specific metabolites can then be used to quantify these metabolites in a complex sample, and see how they change.

Metabolomics can be broken down into a number of areas

- **Metabolic fingerprinting:** taking a 'snapshot' of the metabolome of a system, generating a set of values for the intensity of a signal from a species, without necessarily knowing what that species is. Often there is no chromatographic separation of species. It is used for biomarker discovery.
- **Metabolite profiling:** generating a set of quantitative data on a number of metabolites, usually of known identity, over a range of conditions or times. It is used for metabolomics, metabonomics, and systems biology and biomarker discovery.
- **Metabolite target analysis:** measuring of the concentration of a specific metabolite or small set of metabolites over a range of conditions or times.

Biomarkers

Biomarkers are markers that can be used in medicine for the early detection, diagnosis, staging or prognosis of disease, or for determination of the most effective therapy

A biomarker is generally defined as a marker that is specific for a particular state of a biological system. The markers may be metabolites, peptides, proteins or any other biological molecule, or measurements of physical properties, *e.g.* blood pressure. The importance of biomarkers is rapidly increasing because the trend to personalized medicine will be impossible to sustain without a detailed and objective characterization of the patients afforded by biomarkers. Biomarkers can arise from the disease process itself or from the reaction of the

body to the disease. Thus, they can be found in body fluids and tissues. For ease of sample sourcing and patient compliance, most biomarker studies use urine or plasma, although saliva, interstitial fluid, nipple duct aspirates, and cerebrospinal fluid also have been used.

The most common methods for biomarker discovery have developed from those used in transcriptomics, proteomics and metabolomics, *i.e.* gene arrays, mass spectrometry, often coupled to chromatography, and NMR spectroscopy

Biomarker discovery is often done on small patient cohorts but, to be clinically useful, a robust statistical analysis of a large number of samples from healthy and sick individuals in well-controlled studies is required. Improvements in methods for statistical analysis, coupled to detection methods that can differentiate hundreds to tens of thousands of individual components in the complex sample, have improved the selectivity to the level where these aims are achievable. It is usually necessary to define a number of markers, *i.e.* a **biomarker panel**, that are indicative for a given disease in order to achieve selectivity rather than just detecting a general systemic response such as the inflammatory response or a closely related disease. In theory, it is not necessary to actually identify what the biomarker is, although doing so may give insight into the underlying biochemistry of the disease, and many regulatory authorities demand that the markers are identified before a method can be licensed. It may also allow subsequent development of cheaper and higher throughput assays.

Some well-known examples of biomarkers are the measurement of blood glucose levels in diabetes, prostate-specific antigen for prostate cancer, and HER-2 or BRCA1/2 genes in breast cancer

Biomarker research can also elucidate disease mechanisms and further markers or potential drug targets. For example, using a 2DE approach to determine which DNA repair pathways had been lost in **breast cancer** led to the discovery that cancers deficient in the BRCA1/2 genes are sensitive to the inhibition of another DNA repair protein, poly(ADP-ribose) polymerase 1, known as PARP-1. Inhibitors of PARP-1 are showing promise in clinical trials for the treatment of BRCA1/2 deficient tumors.

Data analysis and interpretation by bioinformatics and systems biology

The -omics experiments can generate many gigabytes, and even terabytes of information. However, **data are not information, and information is not knowledge**. Making use of this data is fundamentally dependent on computational methods. **Bioinformatics** is the term used for computational methods for the extraction of useful information from the complex datasets generated from-omics experiments: for instance, generating quantitative data on gene transcription from next-generation sequencing, or identifying proteins from the fragments generated in mass spectrometry. The annotation of these datasets, for instance with protein function or localization, and the hierarchical organization of the data, can be seen as static information. **Systems biology** takes this further and generates computational and mathematical models from our knowledge of biology and the refined data coming from bioinformatics analysis. These models are used to simulate biochemical and biological processes *in silico* (an expression meaning ‘performed on computers’) and reveal how complex systems, such as intracellular signaling networks, actually work.

Summary

■ The -omics approaches hold a huge potential for the risk assessment, early detection, diagnosis, stratification and tailored treatment of human diseases.
■ The -omics technologies are being introduced into clinical practice, with genomics and transcriptomics leading the way. This is mainly because DNA and RNA have defined physicochemical properties that are amenable for amplification, and the design of robust assay platforms compatible with the routines of clinical laboratories. For instance, PCR and DNA sequencing are used in forensic medicine to establish paternity and to determine the identity of DNA samples left at crime scenes. Genetic tests for the diagnosis of gene mutations and inherited diseases are now in place.
■ Transcriptomic-based microarray tests for breast cancer have been approved, and similar tests for other diseases are becoming available.
■ Proteomics and metabolomics require specialist equipment and expertise, which are difficult to put into the routine clinical laboratory. However, their information content exceeds that of genomics and with further progress in technology, their clinical applications will become reality.
■ The benefits of -omics technologies are obvious, mainly in risk assessment and the design of personalized treatments. However, there are also huge ethical implications pertaining to the use of this information, and new regulatory guidelines will be needed.

Active learning

1. Why does the information increase when moving from studying genes to studying proteins?
2. What is a gene and how is gene transcription regulated?
3. How can we study gene mutations?
4. How can we determine the sequence of DNA and proteins?

Further reading

Bensimon, A, Heck, AJ, Aebersold, R. Mass spectrometry-based proteomics and network biology. *Annu Rev Biochem*. 2012; 81:379–405.

Chapman, EJ, Carrington, JC. Specialization and evolution of endogenous small RNA pathways. *Nat Rev Genet*. 2007; 8:884–896.

Dong, H, Wang, S. Exploring the cancer genome in the era of next-generation sequencing. *Front Med*. 2012; 6:48–55.

Eckhart, AD, Beebe, K, Milburn, M. Metabolomics as a key integrator for 'omic' advancement of personalized medicine and future therapies. *Clin Transl Sci*. 2012; 5:285–288.

Fratkin, E, Bercovici, S, Stephan, DA. The implications of ENCODE for diagnostics. *Nat Biotechnol*. 2012; 30:1064–1065.

Kolch, W, Pitt, A. Functional proteomics to dissect tyrosine kinase signaling pathways in cancer. *Nat Rev Cancer*. 2010; 10:618–629.

Mischak, H, Allmaier, G, Apweiler, R, *et al.* Recommendations for biomarker identification and qualification in clinical proteomics. *Sci Transl Med*. 2010; 2:46ps42.

Moran, VA, Perera, RJ, Khalil, AM. Emerging functional and mechanistic paradigms of mammalian long noncoding RNAs. *Nucleic Acids Res*. 2012; 40:6391–6400.

Osman, A. MicroRNAs in health and disease – basic science and clinical applications. *Clin Lab*. 2012; 58:393–402.

Paik, YK, Hancock, WS. Uniting ENCODE with genome-wide proteomics. *Nat Biotechnol*. 2012; 30:1065–1067.

Spickett, CM, Pitt, AR, Morrice, N, *et al.* Proteomic analysis of phosphorylation, oxidation and nitrosylation in signal transduction. *Biochim Biophys Acta*. 2006; 1764:1823–1841.

Websites

Cancer Genome Project, US National Cancer Institut. www.cancer.gov/cancertopics/understandingcancer/CGAP

Human Metabolome Projec. www.metabolomics.ca/

International HapMap Projec. www.hapmap.org/

Introduction into Proteomic. www.childrenshospital.org/cfapps/research/data_admin/Site602/mainpageS602P0.html

Introduction into Metabolomic. www.docstoc.com/docs/2699139/Introduction-to-Metabolomics

Next-Generation Sequencin. http://en.wikipedia.org/wiki/DNA_sequencing#Next-generation_methods

NeXtprot protein databas. www.nextprot.org/

Online Mendelian Inheritance in Ma. www.ncbi.nlm.nih.gov/omim

The Human Genome Projec. www.ornl.gov/sci/techresources/Human_Genome/home.shtml

Wellcome Trust Sanger Centre (UK. www.sanger.ac.uk/

CHAPTER 37

Oxygen and Life

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

- Identify the major reactive oxygen species (ROS) and their sources in the cell.
- Identify targets of reactive oxygen damage and describe the effects of ROS on biomolecules.
- Identify the major antioxidant enzymes, vitamins, and biomolecules that provide protection against ROS.
- Describe the role of reactive oxygen in regulatory biology and immunologic defenses.
- Explain the role of ROS in development of chronic and inflammatory diseases.
- Describe the functions of the antioxidant response element in protection against nucleophiles and ROS.

Introduction

At body temperature, oxygen is a relatively sluggish oxidant

The element oxygen (O_2) is essential for the life of aerobic organisms. Although it is highly reactive in combustion reactions at high temperature, oxygen is relatively inert at body temperature; it has a high activation energy for oxidation reactions. This is fortunate, otherwise we might spontaneously combust. About 90% of our O_2 usage is committed to oxidative phosphorylation. Enzymes that use O_2 for hydroxylation and oxygenation reactions consume another 10%, and a residual fraction, <1%, is converted to **reactive oxygen species (ROS)**, such as superoxide and hydrogen peroxide, which are reactive forms of oxygen. ROS are important in metabolism – some enzymes use H_2O_2 as a substrate. ROS also play a role in regulation of metabolism and in immunologic defenses against infection. However, ROS are also a source of chronic damage to tissue biomolecules. One of the risks of harnessing O_2 as a substrate for energy metabolism is that we may, and do, get burned. For this reason, we have a range of antioxidant defenses that protect us against ROS.

This chapter will deal with the biochemistry of reactive oxygen, the mechanisms of formation and detoxification of ROS, and their role in human health and disease.

The inertness of oxygen

In most textbooks, oxygen is shown as a diatomic molecule with two bonds between the oxygen atoms. This is an attractive presentation from the viewpoint of electron dot structures and electron pairing to form chemical bonds, but it is incorrect. In fact, at body temperature, O_2 is a biradical, a molecule with two unpaired electrons (Fig. 37.1). These electrons have parallel spins and are unpaired. Since most organic oxidation reactions, *e.g.* the oxidation of an alkane to an alcohol or an aldehyde to an acid, are two-electron oxidation reactions, O_2 is generally not very reactive in these reactions. In fact, it is completely stable, even in the presence of a strong reducing agent such as H_2. When enough heat (activation energy) is applied, one of the unpaired electrons flips to form an electron pair, which then participates in the combustion reaction. Once started, the combustion provides the heat needed to propagate the reaction, sometimes explosively.

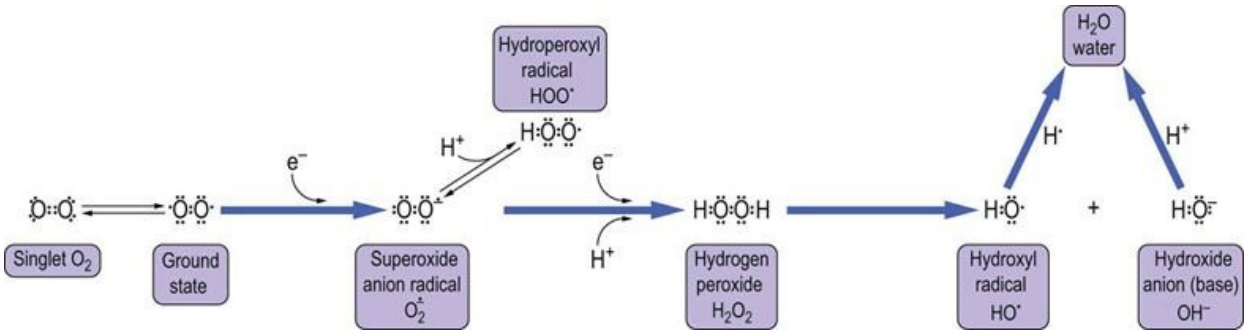


FIG. 37.1 Structure of oxygen and reactive oxygen species (ROS). Oxygen is shown at the far left as the incorrect double-bonded diatomic form. This form, known as singlet oxygen, exists to a significant extent only at high temperature or in response to irradiation. The diradical is the natural, ground-state form of O_2 at body temperature. ROS are partially reduced, reactive forms of oxygen. The first reduction product is the anion radical, superoxide ($O_2^{\bullet-}$), which is in equilibrium with the weak acid, hydroperoxyl radical (pK_a,~4.5). Reduction of superoxide yields hydroperoxide O_2^{-2}, in the form of H_2O_2. Reduction of H_2O_2 causes a hemolytic cleavage reaction that releases hydroxyl radical ($OH^{\bullet}$) and hydroxide ion (OH^-). Water is the end product of complete reduction of O_2.

Oxygen is activated by transition metal ions, such as iron or copper, in the active of metalloenzymes

Metabolic reactions are conducted at body temperature, far below the temperature required to activate free oxygen. In biological redox reactions involving O_2, the oxygen is always activated by redox active metal ions, such as **iron and copper**; these metals also have unpaired electrons and form reactive metal–oxo complexes. All enzymes that use O_2 in vivo are metalloenzymes and, in fact, even the oxygen transport proteins, hemoglobin and myoglobin, contain iron in the form of heme (Chapter 5). These metal ions provide one electron at a time to oxygen, activating O_2 for metabolism. Because iron and copper, and sometimes manganese and other ions, activate oxygen, these redox-active metal ions are kept at very low (submicromolar) free concentrations in vivo. Normally, they are tightly sequestered (compartmentalized) in storage or transport proteins, and they are locally activated at the active sites of enzymes where oxidation chemistry can be contained and focused on a specific substrate. Free redox-active metal ions are dangerous in biological systems because, in free form, they activate O_2, and ROS formed in these reactions cause oxidative damage to biomolecules. Damage to proteins is often site specific, occurring at sites of metal binding to proteins, indicating that metal–oxo complexes participate in ROS-mediated damage in vivo.

Clinical box Iron overload increases risk for diabetes and cardiomyopathy

Patients with hematologic disorders such as hereditary hemochromatosis, thalassemias and sickle cell disease, or who receive frequent blood transfusions, gradually develop **iron overload**, a condition that increases the risk for development of cardiomyopathy and diabetes. The heart and β-cells are rich in mitochondria. The development of secondary disease in iron overload is considered the result of iron-mediated enhancement of mitochondrial ROS production in these tissues. Mutations in the mitochondrial genome may lead to progressive mitochondrial dysfunction, compromising cardiac and β-cell function.

Reactive oxygen species and oxidative stress

ROS are reactive, strongly oxidizing forms of oxygen

Oxidative stress is defined as a condition in which the rate of generation of ROS exceeds our ability to protect ourselves against them, resulting in an increase in oxidative damage to biomolecules (Fig. 37.2). Oxidative stress is a characteristic feature of inflammatory diseases in which cells of the immune system produce ROS in response to challenge. Oxidative stress may be localized, for instance in the joints in arthritis or in the vascular wall in atherosclerosis, or can be systemic, *e.g.* in systemic lupus erythematosus (SLE) or diabetes.

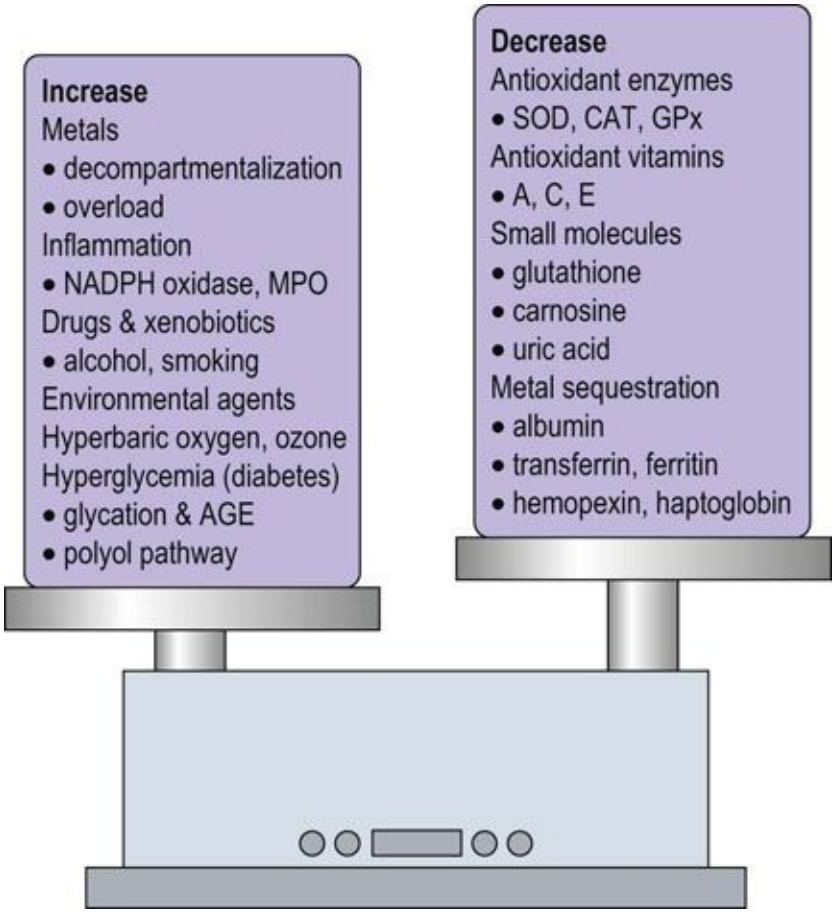


FIG. 37.2 Oxidative stress: an imbalance between prooxidant and antioxidant systems. As described in this chapter, numerous factors contribute to the enhancement and inhibition of oxidative stress. AGE, advanced glycation end product; CAT, catalase; GPx, glutathione peroxidase; MPO, myeloperoxidase; SOD, superoxide dismutase.

Among the ROS, **H_2O_2** is present at highest concentration in blood and tissues, albeit at micromolar or lower concentrations. H_2O_2 is relatively stable; it can be stored in the laboratory or medicine cabinet for years but decomposes in the presence of redox-active metal ions. The **hydroxyl radical** ($OH^{\bullet}$) is the most reactive and damaging species; its half-life, measured in nanoseconds, is diffusion limited, *i.e.* determined by the time to collision with a target biomolecule. **Superoxide** ($O_2^{\bullet-}$) is intermediate in stability and may actually serve as either an oxidizing or reducing agent, forming H_2O_2 or O_2, respectively. At physiologic pH, the **hydroperoxyl radical** ($HOO^{\bullet}$, $pK_a < 4.5$), the protonated form of superoxide (see Fig. 37.1), represents only a small fraction of total $O_2^{\bullet-}$ (about 0.1%), but this radical is intermediate in reactivity, between $O_2^{\bullet-}$ and $OH^{\bullet}$. $HOO^{\bullet}$ and H_2O_2 are small, uncharged molecules and readily diffuse through cell membranes.

ROS are formed by three major mechanisms in vivo: by reaction of oxygen with decompartmentalized metal ions (Fig. 37.3); as a side reaction of mitochondrial electron transport (Fig. 37.4); or by normal enzymatic reactions, *e.g.* formation of H_2O_2 by fatty acid oxidases in the peroxisome (Chapter 15). Secondary ROS are also formed by enzymatic reactions, *e.g.* myeloperoxidase in the macrophage catalyzes the reaction of H_2O_2 with Cl^- to produce another ROS, hypochlorous acid (HOCl). HOCl, which is the major oxidizing species in chlorine-based bleaches, is also part of the bactericidal machinery of the macrophage (see below).

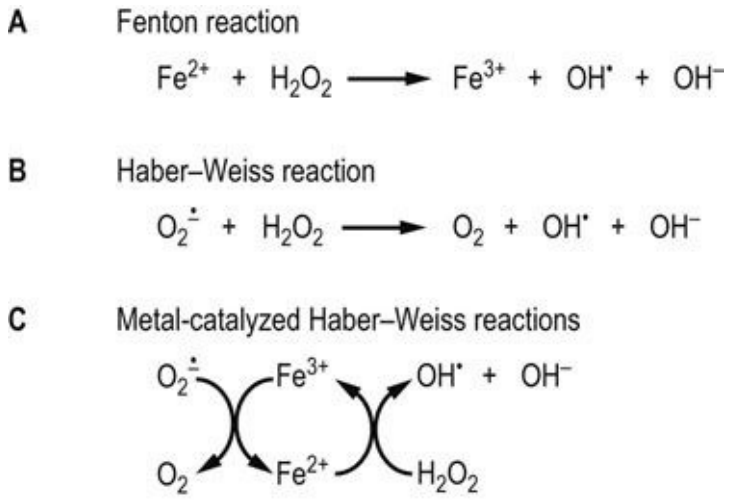


FIG. 37.3 Formation of ROS by the Fenton and Haber–Weiss reactions.
(A) Fenton first described the oxidizing power of solutions of Fe^{2+} and H_2O_2. This reaction generates the strong oxidant $OH^{\bullet}$. Cu^+ catalyzes the same reaction. (B) The Haber–Weiss reaction describes the production of $OH^{\bullet}$ from $O_2^{\bullet-}$ and H_2O_2. (C) Under physiologic conditions, the Haber–Weiss reaction is catalyzed by redox-active metal ions.

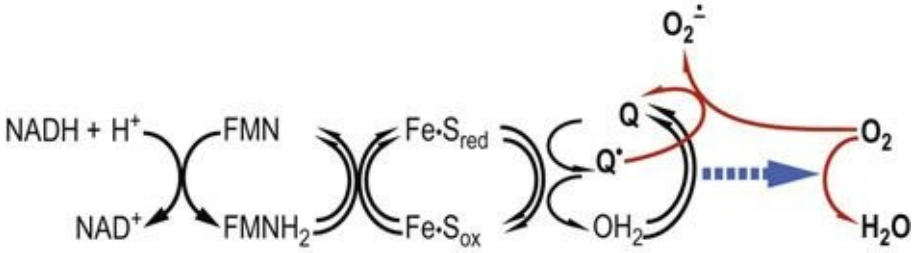


FIG. 37.4 Formation of superoxide by mitochondria.
After the oxidation of NADH, the electron transport chain catalyzes single-electron redox reactions. The semiquinone radical, an intermediate in the reduction of Q to QH_2 by Complex I or II, is sensitive to oxidation by molecular oxygen and is considered a major source of superoxide radicals in the cell. This reaction is enhanced at high oxygen tension and also at low oxygen tension: *e.g.* during ischemia-reperfusion injury (p. 489) when NADH is increased and the electron transport chain is saturated with electrons and the membrane potential is high.

Advanced concept box

Radiotherapy: medical application of reactive oxygen

Radiation therapy uses a focused beam of high-energy electrons or γ-rays from an X-ray or cobalt-60 source to destroy tumor tissue. The radiation produces a flux of hydroxyl radicals (from water) and organic radicals at the site of the tumor. The localized oxidative stress causes damage to all biomolecules in the tumor cell, but the damage to DNA is critical – it prevents tumor cell replication, inhibiting tumor growth. Irradiation is also used as a method of sterilization of food, destroying viral or bacterial contaminants or insect infestations and preserving food products during long-term storage.

Exposure to ionizing radiation from nuclear explosions or accidents, or breathing or ingestion of radioactive elements, such as radon gas or strontium-90, also causes oxidative damage to DNA. Cells that survive the damage may have mutations in DNA that eventually lead to development of cancers. Leukemias are particularly prominent because of the rapid division of bone marrow cells.

Clinical box Toxicity of hyperoxia

Supplemental oxygen therapy may be used for treatment of patients with hypoxemia, respiratory distress, or following exposure to carbon monoxide. Under normobaric conditions, the fraction of oxygen in air can be increased to nearly 100% using a facial mask or nasal cannula. However, patients develop chest pain, cough and alveolar damage within a few hours of exposure to 100% oxygen. Edema gradually develops and compromises pulmonary function. The damage results from overproduction of ROS in the lung. Rats can be protected from oxygen toxicity by gradually increasing the oxygen tension over a period of several days. During this time, antioxidant enzymes, such as superoxide

dismutase, are induced in the lung and provide increased protection against oxygen toxicity.

The lung is not the only tissue affected by hyperoxia. Premature infants, especially those with acute respiratory distress syndrome, often require supplemental oxygen for survival. During the 1950s, it was recognized that the high oxygen tension used in incubators for premature infants increased the risk for blindness, resulting from retinopathy of prematurity (retrolental fibroplasia).

Clinical box Ischemia/reperfusion injury: a patient with myocardial infarction

A patient suffered a severe myocardial infarction, which was treated with tissue plasminogen activator, a clot-dissolving (thrombolytic) enzyme. During the days following hospitalization, the patient experienced palpitations, irregular rapid heartbeat, associated with weakness and faintness. The patient was treated with antiarrhythmic agents.

Comment.

Ischemia, meaning limited blood flow, is a condition in which a tissue is deprived of oxygen and nutrients. Damage to heart tissue during a myocardial infarction occurs not during the hypoxic or ischemic phase, but during reoxygenation of the tissue. This type of damage also occurs following transplantation, and cardiovascular surgery. ROS are thought to play a major role in reperfusion injury. When cells are deprived of oxygen, they must rely on anaerobic glycolysis and glycogen stores for ATP synthesis. NADH and lactate accumulate, and all the components of the mitochondrial electron transport system are saturated with electrons (reduced), because the electrons cannot be transferred to oxygen. The mitochondrial membrane potential is increased (hyperpolarized), and when oxygen is reintroduced, great

quantities of ROS are rapidly produced, overwhelming antioxidant defenses. ROS flood throughout the cell, damaging membrane lipids, DNA and other vital cellular constituents, leading to necrosis. Antioxidant supplements and drugs are being evaluated for use during recovery from myocardial infarction and stroke, during surgery, and for protection of tissues prior to transplantation.

Reactive nitrogen species and nitrosative stress

Peroxynitrite is a strongly oxidizing reactive nitrogen species

Nitric oxide synthases (NOS) catalyze the production of the free radical nitric oxide ($NO^{\bullet}$) from the amino acid L-arginine. There are three isoforms of NOS: nNOS in neuronal tissue, where $NO^{\bullet}$ serves a neurotransmitter function; iNOS in the immune system, where it is involved in regulation of the immune response; and eNOS in endothelial cells, where $NO^{\bullet}$, known as endothelium-derived relaxation factor (EDRF), has a role in the regulation of vascular tone.

In a side reaction at sites of inflammation, $NO^{\bullet}$ reacts with $O_2^{\bullet-}$ to form the strong oxidant, peroxynitrite ($ONOO^-$), a reactive nitrogen species (RNS). Like ROS, which produce oxidative stress, RNS produce nitrosative stress by reaction with biomolecules. ONOOH has many of the strong oxidizing properties of $OH^{\bullet}$ but has a longer biological half-life. It is also a potent nitrating agent, producing nitrotyrosine in proteins, nitrated phospholipids in membranes and nucleotides in DNA. Simultaneous production of $NO^{\bullet}$ and $O_2^{\bullet-}$, with the concomitant increase in $ONOO^-$ and a decrease in $NO^{\bullet}$, is thought to limit vasodilatation and exacerbate hypoxia and oxidative stress in the vascular wall during ischemia-reperfusion injury, setting the stage for vascular disease. ONOOH degrades, in part, by homolytic cleavage to produce two very reactive species, $OH^{\bullet}$ and $NO_2^{\bullet}$. $NO_2^{\bullet}$ is also formed by eosinophil peroxidase or myeloperoxidase-catalyzed oxidation of $NO^{\bullet}$ by H_2O_2.

The nature of oxygen radical damage

The hydroxyl radical is the most reactive and damaging ROS

The reaction of ROS with biomolecules produces characteristic products, described as biomarkers of oxidative stress. These compounds may be formed directly in the oxidation reaction with the ROS, or by secondary reactions between oxidation products and other biomolecules. The hydroxyl radical reacts with biomolecules primarily by hydrogen abstraction and addition reactions. One of the most sensitive sites of free radical damage are cell membranes, which are rich in readily oxidized polyunsaturated fatty acids (PUFA). Peroxidative damage to the plasma membrane affects the integrity and function of the membrane, compromising the cell's ability to maintain ion gradients and membrane phospholipid asymmetry. As shown in Figure 37.5, when $OH^{\cdot}$ abstracts a hydrogen atom from a PUFA, it initiates a chain of lipid peroxidation reaction, producing secondary oxidation products, lipid peroxides and lipid peroxyl radicals. The lipid oxidation products formed in this reaction degrade to form reactive compounds, such as **malondialdehyde** (MDA) and **hydroxynonenal** (HNE). These compounds react with proteins to form adducts and crosslinks, known as **advanced lipoxidation end products (ALE)**. Increased levels of MDA and HNE adducts to lysine residues have been measured in lipoproteins in plasma and the vascular wall in atherosclerosis and in amyloid plaque in Alzheimer's disease, implicating oxidative stress and damage in the pathogenesis of these diseases.

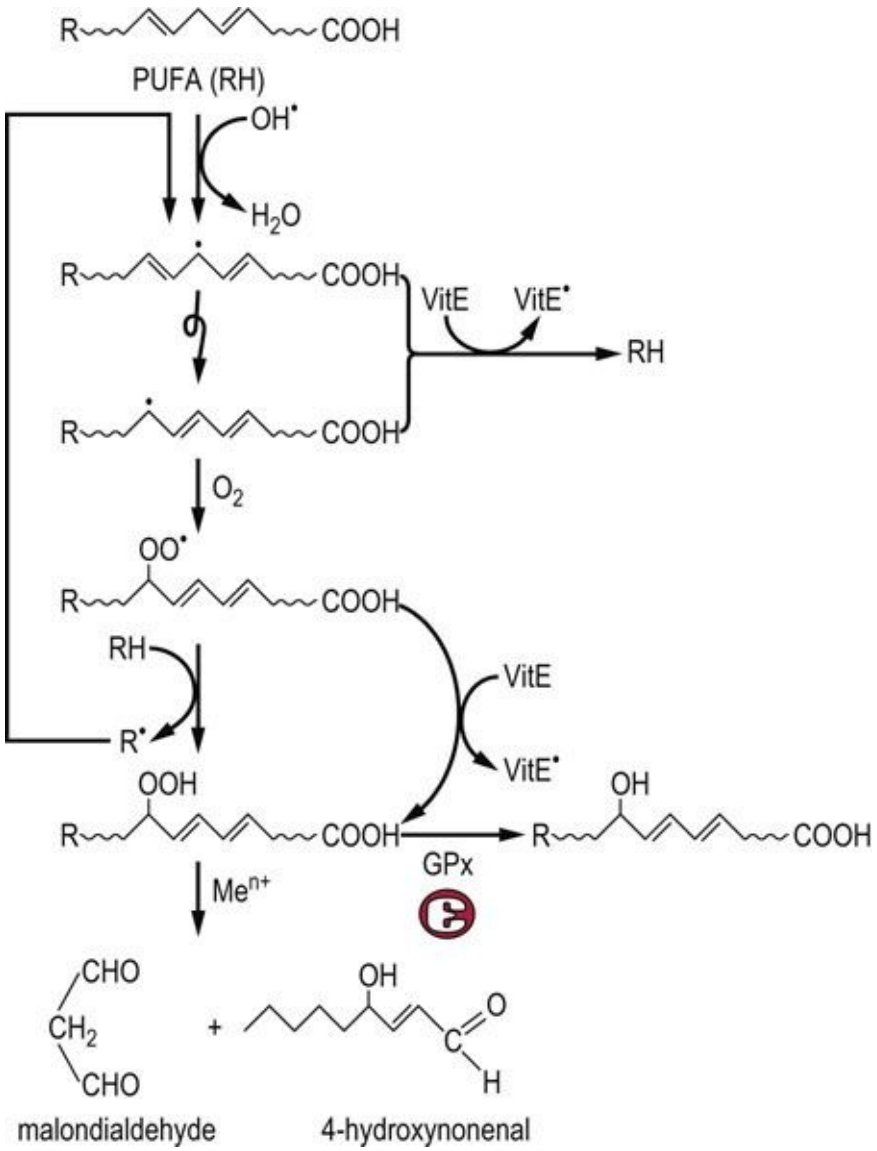


FIG. 37.5 Pathway of lipid peroxidation.

$OH^{\bullet}$ attacks PUFA, forming a carbon-centered lipid radical. The radical rearranges to form a conjugated dienyl radical. This radical reacts with ambient O_2, forming a hydroperoxyl radical, which then abstracts a hydrogen from a neighboring lipid, forming a lipid peroxide and $R^{\bullet}$, which starts a chain reaction. This reaction continues until the supply of PUFA is exhausted, unless a termination reaction occurs. Vitamin E (below) is the major chain-terminating antioxidant in membranes; it reduces both the conjugated dienyl and hydroperoxyl radicals, quenching the chain or cycle of lipid peroxidation reactions. Lipid peroxides may also be reduced by glutathione peroxidase (GPx), forming inert lipid alcohols Otherwise, they decompose to form a range of '**reactive carbonyl species**' such as malondialdehyde and hydroxynonenal, which react with protein to form advanced lipoxidation end products (ALE), which are biomarkers of oxidative stress. The reaction scheme shown here for PUFA also occurs with intact phospholipids and cholesterol esters in lipoproteins and cell membranes.

Hydroxyl radicals also react by addition to phenylalanine, tyrosine and nucleic acid bases to form hydroxylated derivatives and crosslinks (Fig. 37.6). Other ROS and RNS leave tell-tale tracks, such as nitro-and chlorotyrosine, formed from ONOOH and HOCl, respectively, and methionine sulfoxide, formed by reaction of H_2O_2 or HOCl with methionine residues in proteins (see Fig. 37.6). Nitrotyrosine, like ALEs, is increased in atherosclerotic and Alzheimer's plaques.

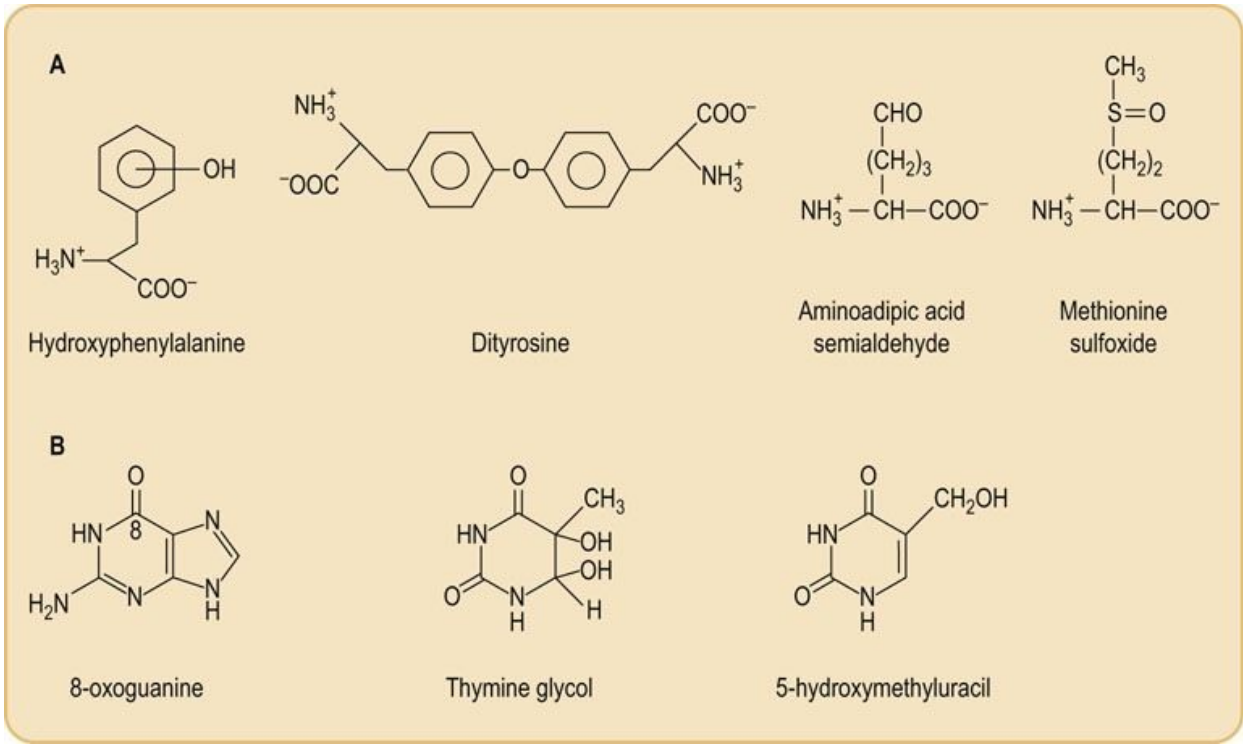


FIG. 37.6 Products of hydroxyl radical damage to biomolecules.
(A) Amino acid oxidation products: *o-, m-* and *p*-tyrosine and dityrosine from phenylalanine; amino adipic acid semialdehyde from lysine; methionine sulfoxide. Other products include chlorotyrosine (from HOCl), nitrotyrosine (from $ONOO^-$ and $NO_2^{\bullet}$), dihydroxyphenylalanine produced by hydroxylation of tyrosine, and aliphatic amino acid hydroperoxides, such as leucine hydroperoxide. (B) Nucleic acid oxidation products: 8-oxoguanine is the most commonly measured indicator of DNA damage.

ROS also react with carbohydrates to form reactive carbonyl compounds that react with protein to form **crosslinks and adducts, known as advanced glycation end products (AGE)**. AGEs are increased in tissue proteins in diabetes as a result of hyperglycemia and oxidative stress, and the increase in

chemical modification of proteins by AGEs and ALEs is implicated in the development of diabetic vascular, renal, and retinal complications (see Chapter 21).

Antioxidant defenses

There are several levels of protection against oxidative damage

ROS damage to lipids and proteins is repaired largely by degradation and resynthesis. Oxidized proteins, for example, are preferred targets for proteasomal degradation, and damaged DNA is repaired by a number of excision-repair mechanisms. The process is not perfect. Some proteins, such as collagens and crystallins, turn over slowly, so that damage accumulates and function may be impaired, *e.g.* age-dependent browning of lens proteins, crosslinking of collagen and elastin, and loss of elasticity or changes in permeability of the vascular wall and renal basement membrane (see Chapter 43). The association between chronic inflammation and cancer indicates that chronic exposure to ROS causes cumulative damage to the genome in the form of nonlethal mutations in DNA.

Advanced concept box Sentinel function of methionine

Methionine (Met) residues in protein may be oxidized to methionine sulfoxide (MetSO) by H_2O_2, HOCl or lipid peroxides. Met is generally on the surface of proteins and rarely has a role in the active site or mechanism of action of enzymes. However, there is evidence that it serves as an ‘antioxidant pawn’, protecting the active site of enzymes. Half of the Met residues of glutamine synthetase can be oxidized without affecting the enzyme's specific activity. These residues are physically arranged in an array that ‘guards’ the entrance to the active site, protecting the enzyme from inactivation by ROS. MetSO can be reduced back to methionine by methionine sulfoxide reductases, providing a catalytic amplification of the antioxidant potential of each methionine residue.

Clinical box Methionine oxidation and emphysema

α_1-Antitrypsin (A1AT) is a plasma protein, synthesized and secreted by liver. It is a potent inhibitor of elastase and protects tissues from damage by the neutrophil enzyme secreted during inflammation. Deficiency of this protein (about 1 in 4000 worldwide) is commonly associated with emphysema, progressive lung disease, and also hepatic damage resulting from accumulation of protein aggregates. The lung damage is attributed to failure of A1AT to inhibit elastase released by alveolar macrophages during phagocytosis of airborne particulate matter.

Therapy includes replacement therapy by weekly intravenous infusion of a purified plasma concentrate or recombinant protein. Cigarette smoking and exposure to mineral dust (coal, silica) exacerbate pathology in patients with A1AT deficiency, but are also independent risk factors for emphysema and pulmonary fibrosis. Cigarette smoke and microparticulate materials activate lung macrophages, leading to release of proteolytic enzymes and increased production of ROS as a result of inflammation. The ROS cause oxidation of a specific Met residue in A1AT, irreversibly inhibiting the anti-elastase activity of this protein. Increased levels of inactive, Met(O)-containing A1AT are present in plasma of chronic smokers.

Advanced concept box Selenium, an antioxidant micronutrient

Selenocysteine is an unusual amino acid in proteins, found in only 25 proteins in the human proteome. It is encoded by UGA, which is normally a STOP codon, under direction of a SElenoCysteine Insertion Sequence (SECIS), a 50-nucleotide stem-loop structure in the mRNA. The 25-member selenoproteome includes five

glutathione peroxidase isozymes, three thioredoxin reductases, methionine sulfoxide reductase, one of three enzymes that reduces methionine sulfoxide back to methionine, and three iodothyronine deiodinases. About a third of the selenoproteins have no known function, but appear to be involved in antioxidant defense mechanisms. Selenium is essential for life, in part because of severe hypothyroidism and oxidative stress in its absence. Selenium deficiency in adults is associated with cardiomyopathy in Keshan disease, osteoarthropathy (cartilage degeneration) in Kashin–Beck disease, and with symptoms of hypothyroidism, including chronic fatigue and goiter.

Our first line of defense against oxidative damage is sequestration or chelation of redox-active metal ions

Advanced concept box The antioxidant response element

Cells adapt to oxidative stress by induction of antioxidant enzymes. Many of these are controlled by the antioxidant response element (ARE), also known as the electrophile response element. The central regulator of the ARE is the transcription factor Nrf2, which is retained in inactive form in the cytoplasmic compartment by binding to a cysteine-rich protein, Keap1. Under normal conditions, Keap1 directs ubiquitination and proteasomal degradation of Nrf2. During oxidative stress, modification of the sulfhydryl groups of Keap1 by nucleophiles, such as the lipid peroxidation products, hydroxynonenal or acrolein, causes dissociation of Nrf2 from Keap1. Nrf2 then translocates to the nucleus and activates ARE-dependent genes. Keap1 also reacts with exogenous nucleophiles, including carcinogens that would otherwise react with DNA.

ARE-dependent enzymes include catalase (CAT) and superoxide dismutase (SOD), and enzymes that catalyze the

oxidation and conjugation of carcinogens and oxidants for excretion. One of these enzymes, hepatic glutathione *S*-transferase, catalyzes conjugation with GSH. The conjugates are then excreted in urine as **mercapturic acid**, which are *S*-substituted *N*-acetyl-cysteine derivatives.

Endogenous chelators include a number of metal-binding proteins that sequester iron and copper in inactive form, such as transferrin and ferritin, the transport and storage forms of iron. The plasma protein haptoglobin binds to hemoglobin from ruptured red cells, and delivers the hemoglobin molecule to the liver for catabolism. Plasma hemopexin binds heme, the lipid-soluble form of iron, which catalyzes ROS formation in lipid environments; it delivers the heme to the liver for catabolism. Albumin, the major plasma protein, has a strong binding site for copper and effectively inhibits copper-catalyzed oxidation reactions in plasma. Carnosine (β-alanyl-L-histidine) and related peptides are present in muscle and brain at millimolar concentrations; they are potent copper chelators and may have a role in intracellular antioxidant protection.

Despite these manifold and potent metal chelation systems, ROS are formed continuously in the body, both by enzymes and spontaneous metal-catalyzed reactions. In these cases, there are a group of enzymes that act to detoxify ROS and their precursors. These include **superoxide dismutase (SOD), catalase (CAT), and glutathione peroxidase (GPx)** (Fig. 37.7). SOD converts $O_2^{\bullet}$ to the less toxic H_2O_2. There are two classes of SOD: an MnSOD isozyme which is found in mitochondria, and CuZnSOD isozyme which is widely distributed throughout the cell. An extracellular, secreted glycoprotein isoform of CuZnSOD (EC-SOD) binds to proteoglycans in the vascular wall and is thought to protect against $O_2^{\bullet}$ and $ONOO^-$ injury. CAT, which inactivates H_2O_2, is found largely in peroxisomes, the major site of H_2O_2 generation in the cell.

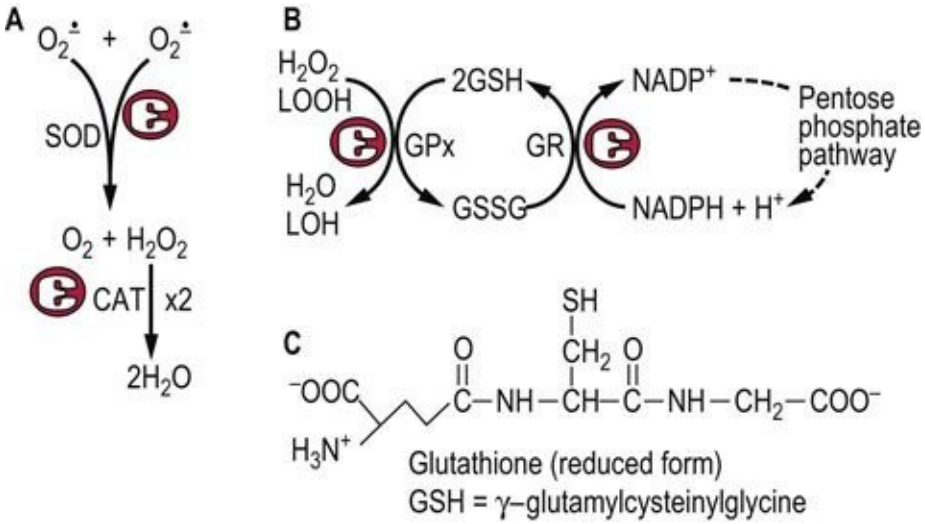


FIG. 37.7 Enzymatic defenses against ROS.
(A) Superoxide dismutase (SOD) and catalase (CAT) are dismutases, catalyzing oxidation and reduction of two separate substrate (H_2O_2) molecules; both are highly specific for their substrates, $O_2^{\cdot}$ and H_2O_2, respectively. (B) Glutathione peroxidase (GPx) reduces H_2O_2 and lipid peroxides (LOOH), using GSH as a co-substrate. The GSH is recycled by glutathione reductase (GR) using NADPH from the pentose phosphate pathway. (C) Structure of GSH.

GPx is widely distributed in the cytosol, in mitochondria and the nucleus. It reduces H_2O_2 and lipid hydroperoxides to water and a lipid alcohol, respectively, using reduced glutathione (GSH) as a co-substrate. GSH is a tripeptide (Chapter 2) that is present at 1–5 mM concentration in all cells. The GSH is recycled by an NADPH-dependent enzyme, GSH reductase. The NADPH, provided by the pentose phosphate pathway, maintains about a 100 : 1 ratio of GSH : GSSG in the cell. GPx is actually a family of selenium-containing isozymes; a phospholipid hydroperoxide glutathione peroxidase will reduce lipid hydroperoxides in phospholipids in lipoproteins and membranes, while other isozymes are specific for free fatty acid or cholesterol ester hydroperoxides. There is also an isoform of GPx in intestinal epithelial cells, which is thought to have a role in detoxification of dietary hydroperoxides, *e.g.* in fried foods.

Vitamin C is the outstanding antioxidant in biological systems

Three antioxidant vitamins, A, C and E, provide the third line of defense against oxidative damage. These vitamins, primarily vitamin C (ascorbate) (Fig. 37.8) in

the aqueous phase and vitamin E (α-and γ-tocopherol; Fig. 37.9) in the lipid phase, act as chain-breaking antioxidants (see Fig. 37.5). They act as reducing agents, donating a hydrogen atom ($H^{\bullet}$) and quenching organic radicals formed by reaction of ROS with biomolecules. The vitamin C and E radicals produced in this reaction are unreactive, resonance-stabilized species; they do not propagate radical damage and are enzymatically recycled, *e.g.* by dehydroascorbate reductase (see Fig. 37.8). Vitamin C reduces superoxide and lipid peroxyl radicals, but also has a special role in reduction and recycling of vitamin E. In response to severe oxidative stress. vitamin C recycles vitamin E, so that vitamin E is maintained at constant concentration in the lipid phase until all the vitamin C is consumed (see Fig. 37.9). These antioxidants work together to inhibit lipid peroxidation reactions in plasma lipoproteins and membranes. Vitamin A (carotene; Chapter 11) is also a lipophilic antioxidant. Although best understood for its role in vision, it is a potent singlet oxygen scavenger and protects against damage from sunlight in the retina and skin.

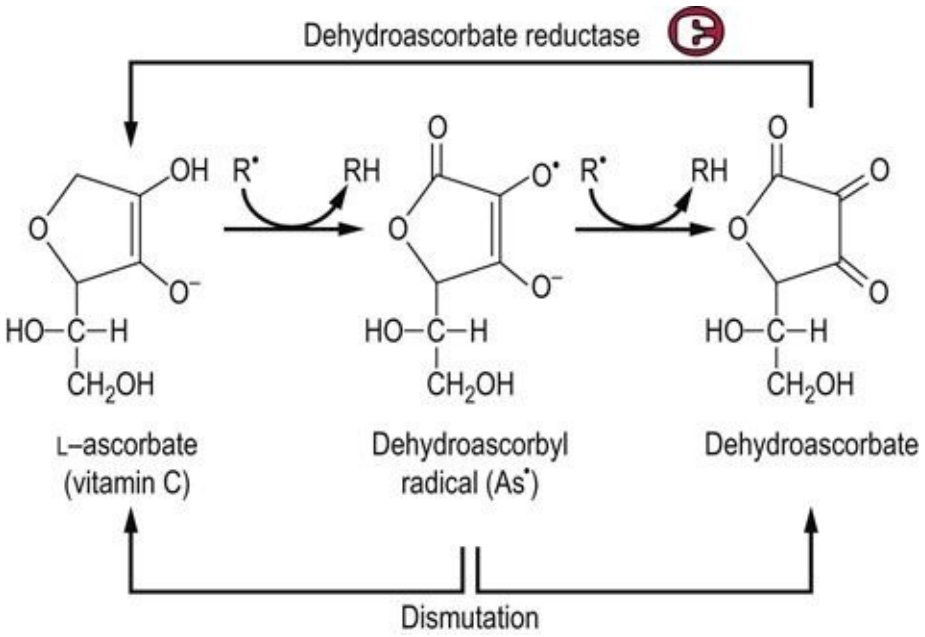


FIG. 37.8 Antioxidant activity of ascorbate.
Vitamin C exists as the enolate anion at physiologic pH. The enolate anion spontaneously reduces superoxide, organic ($R^{\bullet}$) and vitamin E radicals, forming a dehydroascorbyl radical ($As^{\bullet}$). Dehydroascorbyl radical may dismutate to ascorbate and dehydroascorbate. Dehydroascorbate is also recycled by dehydroascorbate reductase, a GSH-dependent enzyme present in all cells.

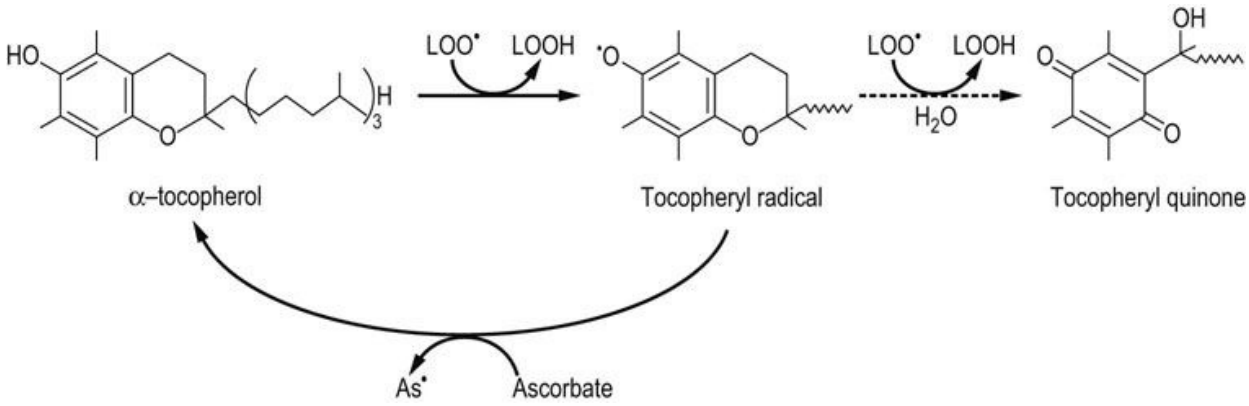


FIG. 37.9 Antioxidant activity of vitamin E.
The term vitamin E refers to a family of related tocopherol and tocotrienol isomers with potent lipophilic antioxidant and membrane-stabilizing activity. Tocopherols reduce lipid hydroperoxyl radicals and also inactivate singlet oxygen. α-Tocopherol is the most effective form in humans and the major form of vitamin E in the diet. It consists of a chromanol ring structure with a polyisoprenoid side chain which helps to anchor the vitamin in membranes; the isoprene units are unsaturated in tocotrienol. The α, β, γ, and δ Isomers differ in the pattern of methyl groups on the benzene ring (see Figure 11.3). The major commercial form of vitamin E is α-tocopherol acetate, which is more stable than free tocopherol during storage. The tocopheryl radical, the major product formed during antioxidant action of vitamin E, is recycled by ascorbate. Tocopheryl quinone is also formed in small quantities.

Glutathionylation of proteins – protection against ROS under stress

Despite the multiplicity of defensive mechanisms, there is always some evidence of ongoing oxidative damage in tissues. Under physiologic conditions, when proteins are exposed to O_2, their sulfhydryl groups gradually oxidize to form disulfides, either intramolecularly or intermolecularly with other proteins. This is multi-step process. First, a protein sulfhydryl group is oxidized to a **sulfenic acid** (PrSOH) by an ROS, such as H_2O_2 or HOCl; then the sulfenic acid reacts with another PrSH to form a crosslinked protein PrS-SPr. These crosslinked proteins are reduced by glutathione to form oxidized glutathione and regenerate the native protein with free sulfhydryl groups. The reaction sequence is

$$PrSH + ROS \rightarrow PrSOH$$

$$PrSOH + PrSH \rightarrow PrS\text{-}SPr + H_2O$$

$$PrS\text{-}SPr + 2\ GSH \rightarrow 2\ PrSH + 2\ GSSG$$

During oxidative stress, there is a significant increase in S-glutathionylated proteins (PrS-SG) in the cell. In this case, the reaction sequence is

$$PrSH + ROS \rightarrow PrSOH$$

$$PrSOH + GSH \rightarrow PrS\text{-}SG + H_2O$$

$$PrS\text{-}SG + GSH \rightarrow PrSH + GSSG$$

S-**Glutathionylation** is reversed by nonenzymatic reduction by GSH (above) or by enzymes using thiol protein cofactors (thioredoxin, glutaredoxin). This pathway inhibits the formation of crosslinked protein aggregates, such as **Heinz bodies**, which are hemoglobin precipitates that develop in red cells in glucose-6-phosphate dehydrogenase deficiency, characterized by decreased levels of GSH (Chapter 12). *S*-glutathionylation is thought to have a dual role, not only in protecting cysteine against irreversible oxidation to the sulfinic or sulfonic acid during oxidative and/or nitrosative stress but also in modulating cellular metabolism (redox regulation). Target proteins include a wide range of enzymes with active site or regulatory –SH groups, such as glyceraldehyde-3-phosphate dehydrogenase in glycolysis and protein kinases in signaling cascades, as well as chaperones and transport proteins. *S*-glutathionylation also appears to protect proteins from ubiquitin-mediated proteasomal degradation during oxidative stress.

The beneficial effects of reactive oxygen species

ROS are essential for many metabolic and signaling pathways

Clinical test box Peroxidase activity for detection of occult blood

Peroxidases, such as the glutathione peroxidase (GPx), are enzymes that catalyze the oxidation of a substrate using H_2O_2. Hemoglobin and heme have a pseudoperoxidase activity in vitro. In the guaiac-based test for occult fecal blood, a stool sample is applied to a small card containing guaiac acid. Hemoglobin in a stool specimen oxidizes phenolic compounds in guaiac acid to quinones. A positive test is indicated by a blue stain along the edge of the fecal smear. Incompletely digested hemoglobin and myoglobin from animal meat and some plant peroxidases may cause false positives. Similar peroxidase-dependent assays are used to identify bloodstains at crime scenes.

Advanced concept box The glyoxalase pathway: a special role for glutathione

A small fraction of triose phosphates produced during metabolism spontaneously degrades to **methylglyoxal** (MGO), a reactive dicarbonyl sugar. MGO is also formed during metabolism of glycine and threonine, and as a product of nonenzymatic oxidation of carbohydrates and lipids – it is a significant precursor of advanced glycation and lipoxidation end products (AGE/ALEs)

(see Chapters 21 and 43). MGO reacts primarily with arginine residues in proteins, but also with lysine, histidine and cysteine, leading to enzyme inactivation and protein crosslinking.

MGO is inactivated by enzymes of the glyoxalase pathway, a GSH-dependent system found in all cells in the body. The glyoxalase pathway (Fig. 37.10) consists of two enzymes that catalyze an internal redox reaction in which carbon-1 of MGO is oxidized from an aldehyde to a carboxylic acid group and carbon-2 is reduced from a ketone to a secondary alcohol. The end product, D-lactate, does not react with proteins; D-lactate is distinct from L-lactate, the product of glycolysis, but may be converted into L-lactate for further metabolism.

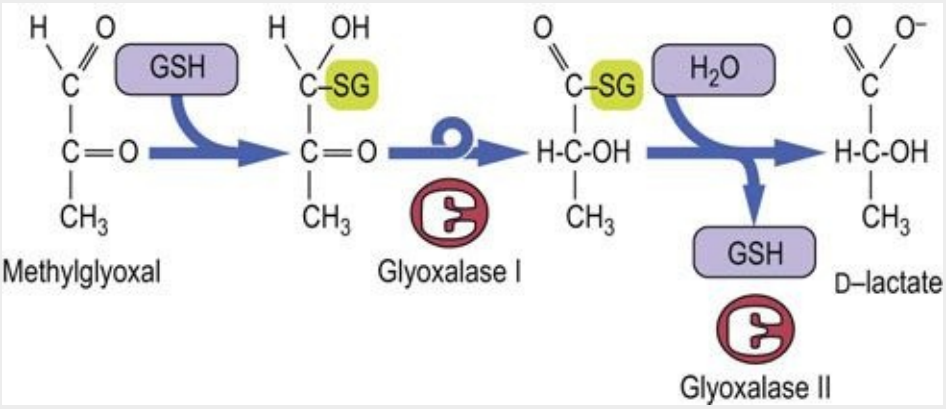


FIG. 37.10 The glyoxalase system.
Glyoxalase I catalyzes the formation of a thiohemiacetal adduct between GSH and MGO and its rearrangement to a thioester. Glyoxalase II catalyzes the hydrolysis of the thioester, forming D-lactate and regenerating GSH. Unlike GPx, this pathway does not consume GSH.

Levels of MGO and D-lactate are increased in blood of diabetic patients, because levels of glucose and glycolytic intermediates, including triose phosphates, are increased intracellularly in diabetes. The glyoxalase system also inactivates glyoxal, and other dicarbonyl sugars produced during nonenzymatic oxidation of carbohydrates and lipids. Glyoxalase inhibitors are being evaluated for chemotherapy because cancer cells appear to be more sensitive to glyoxalase inhibitors, perhaps because of their increased

reliance on glycolysis.

Advanced concept box The respiratory burst in macrophages

As outlined in Fig. 37.11, macrophages launch a sequence of ROS-producing reactions during the burst of oxygen consumption accompanying phagocytosis. **NADPH oxidase** in the macrophage plasma membrane is activated to produce $O_2^{\bullet-}$, which is then converted to H_2O_2 by superoxide dismutase. The H_2O_2 is used by another macrophage enzyme, myeloperoxidase (MPO), to oxidize chloride ion, ubiquitous in body fluids, to hypochlorous acid (HOCl). H_2O_2 and HOCl mediate bactericidal activity by oxidative degradation of microbial lipids, proteins, and DNA. The macrophage has a high intracellular concentration of antioxidants, especially ascorbate, to protect itself during ROS production, but its relatively short life span, 2–4 months, suggests that it is not immune to oxidative damage.

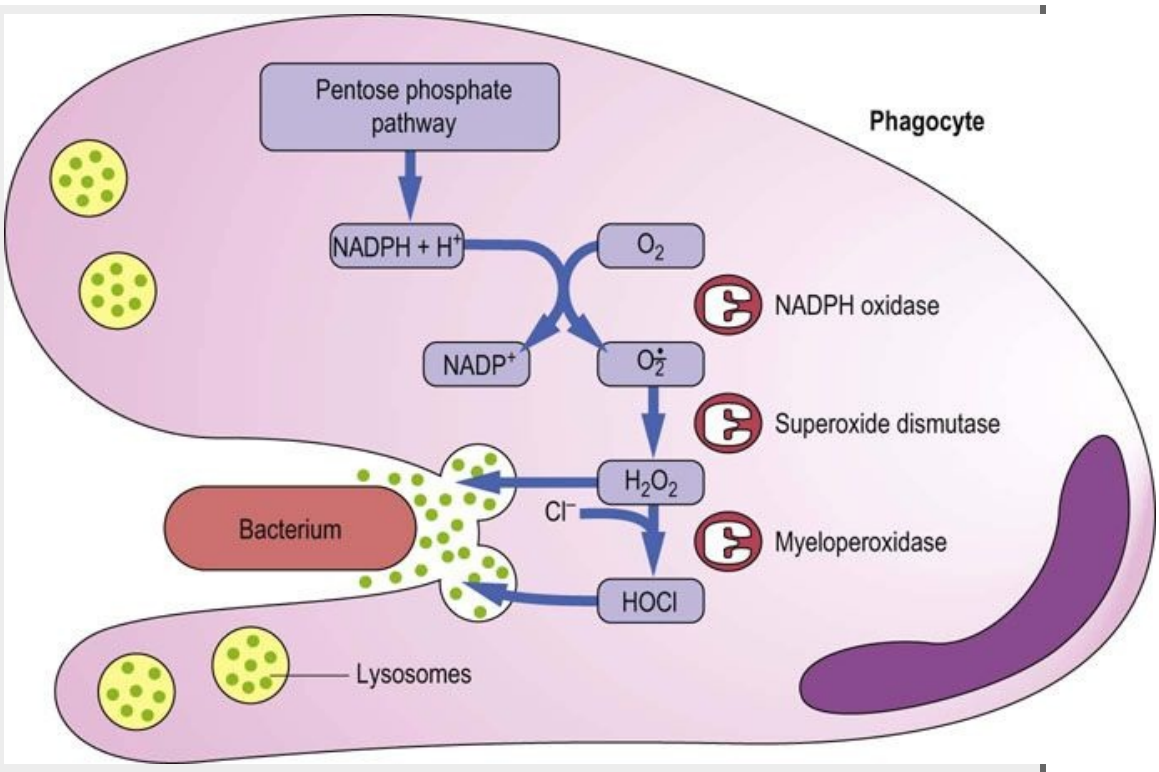


FIG. 37.11 Generation and release of ROS during phagocytosis. A cascade of reactions generating ROS is initiated during phagocytosis to kill invading organisms. Hydrolytic enzymes are also released from lysosomes to assist in degradation of microbial debris.

The consumption of O_2 by NADPH oxidase is responsible for the '**respiratory burst**', the sharp increase in O_2 consumption for production of ROS, which accompanies phagocytosis. One of the end products of this reaction sequence, HOCl, is also the active oxidizer in chlorine-containing laundry bleaches. Intravenous infusion of dilute HOCl solutions was actually used for treatment of bacterial sepsis in battlefield hospitals during World War I, before the advent of penicillin and other antibiotics. Chronic granulomatous disease (CGD) is an inherited disease resulting from a genetic defect in NADPH oxidase. The inability to produce superoxide leads to chronic life-threatening bacterial and fungal infections.

Advanced concept box
Antioxidant defenses in the red blood cell (RBC)

The RBC does not use oxygen for metabolism, nor is it involved in phagocytosis. However, because of the high O_2 tension in arterial blood and the heme iron content of RBCs, ROS are formed continuously in the RBC. Hb spontaneously produces superoxide ($O_2^{\bullet-}$) in a minor side reaction associated with binding of O_2. The occasional reduction of O_2 to $O_2^{\bullet-}$ is accompanied by oxidation of normal (ferro) Hb to methemoglobin (ferrihemoglobin), a rust-brown protein that does not bind or transport O_2. **Methemoglobin** may release heme, which reacts with $O_2^{\bullet-}$ and H_2O_2 in Fenton-type reactions to produce hydroxyl radical ($OH^{\bullet}$) and reactive iron-oxo species. These ROS initiate lipid peroxidation reactions that can lead to loss of membrane integrity and cell death.

The RBC is well fortified with antioxidant defenses to protect itself against oxidative stress. These include catalase (CAT), superoxide dismutase (SOD) and glutathione peroxidase (GPx), as well as a methemoglobin reductase activity that reduces methemoglobin back to normal ferrohemoglobin. Normally, less than 1% of Hb is present as methemoglobin. However, persons with congenital **methemoglobinemia**, resulting from methemoglobin reductase deficiency, typically have a dark and cyanotic appearance. Treatment with large doses of ascorbate (vitamin C) is used to reduce their methemoglobin to functional hemoglobin.

GSH, present at ~2 mmol/L in the RBC, not only supports antioxidant defenses but also is an important sulfhydryl buffer, maintaining –SH groups in hemoglobin and enzymes in the reduced state.

While this chapter has focused thus far on the dangerous aspects of reactive oxygen, it is worth closing with some recognition of the beneficial effects of ROS. Among these are the regulatory functions of NO, the role of ROS in activation of the ARE, the requirement for ROS in the bactericidal activity of macrophages, and the use of ROS as substrates for enzymes, *e.g.* H_2O_2 for the hemeperoxidases involved in iodination of thyroid hormone. There is also increasing evidence that ROS, particularly H_2O_2, are important signaling molecules involved in regulation of metabolism. The tissue concentration of H_2O_2 is estimated to be in the submicromolar range; estimates vary widely, from 1 to 700 nmol/L. However, significant changes in H_2O_2 concentration occur in response to cytokines, growth factors and biomechanical stimulation. The fact that these signaling events are inhibited by peroxide scavengers or by overexpression of catalase implicates H_2O_2 in the signaling cascade. Insulin signaling, for example, appears to involve H_2O_2 as part of the mechanism for reversible inactivation of some protein tyrosine phosphatases, at the same time that protein tyrosine kinases are activated through the insulin receptor (Chapter 21). As the evidence for the signaling role of H_2O_2 has become convincing, there is increasing interest in research on the regulatory role of superoxide.

Summary

■ Reactive oxygen species (ROS) are the sparks produced by oxidative metabolism, and oxidative stress may be viewed as the price we pay for using oxygen for metabolism.
■ ROS and RNS, such as superoxide, peroxide, hydroxyl radical, and peroxynitrite, are reactive and toxic, sometimes difficult to contain, but their production is important for regulation of metabolism, turnover of biomolecules and protection against microbial infection.
■ ROS and RNS cause oxidative damage to all classes of biomolecules: proteins, lipids and DNA.
■ There are a number of protective antioxidant mechanisms, including sequestration of redox-active metal ions, enzymatic inactivation of major ROS, inactivation of organic radicals by small molecules, such as GSH and vitamins, and, when all else fails, repair and/or turnover, and, in extremis, apoptosis.
■ Biomarkers of oxidative stress are readily detected in tissues in inflammation, and oxidative stress is increasingly implicated in the pathogenesis of age-related, chronic disease.
■ Despite their damaging actions, ROS are also essential for the normal functions of the immune system, and for many enzymes and cell signaling pathways.

Active learning

1. Review the evidence that atherosclerosis is an inflammatory disease resulting from overproduction of ROS in the vascular wall.
2. Discuss the evidence that hyperglycemia in diabetes induces a state of oxidative stress that leads to renal and vascular complications.
3. Review the data on use of antioxidants in therapy for atherosclerosis and diabetes. Based on these studies, how strong is the evidence that these diseases are the result of increased oxidative stress?
4. Discuss recent advances in the use of antioxidants for organ and tissue protection during surgery and transplantation.

Further reading

Burgoyne, JR, Oka, SI, Ale-Agha, N, *et al.* Hydrogen peroxide sensing and signaling by protein kinases in the cardiovascular system. *Antioxid Redox Signal*. 2013; 18:1042–1052.

Ferrari, CK, Souto, PC, França, EL, *et al.* Oxidative and nitrosative stress on phagocytes' function: from effective defense to immunity evasion mechanisms. *Arch Immunol Ther Exp*. 2011; 59:441–448.

Greenough, MA, Camakaris, J, Bush, AI. Metal dyshomeostasis and oxidative stress in Alzheimer's disease. *Neurochem Int*. 2013; 62:540–555.

Riccioni, G, D'Orazio, N, Salvatore, C, *et al.* Carotenoids and vitamins C and E in the prevention of cardiovascular disease. *Int J Vitam Nutr Res*. 2012; 82:15–26.

Tkachev, VO, Menshchikova, EB, Zenkov, NK. Mechanism of the Nrf2/Keap1/ARE signaling system. *Biochemistry (Mosc)*. 2011; 76:407–422.

Zhang, H, Forman, HJ. Glutathione synthesis and its role in redox signaling. *Semin Cell Dev Biol*. 2012; 23:722–728.

Zhang, Y, Tocchetti, CG, Krieg, T, *et al.* Oxidative and nitrosative stress in the maintenance of myocardial function. *Free Radic Biol Med*. 2012; 53:1531–1540.

Websites

Antioxidants and cancer. www.cancer.gov/cancertopics/factsheet/antioxidantsprevention.

Free radical lectures. http://web.mst.edu/~nercal/documents/chem464/lectures/Lec01_FreeRadicals.pdf.

Oxidative stress and disease. www.oxidativestressresource.org/.

Reactive oxygen species and antioxidant vitamins. http://lpi.oregonstate.edu/f-w97/reactive.html.

Virtual Free Radical Schoo. www.sfrbm.org/sections/education/frs-presentations

CHAPTER 38

The Immune Response

J. Alastair Gracie and †Alex Farrell

Learning objectives

After reading this chapter you should be able to:

- Compare and contrast the basis of both innate and adaptive immune responses.
- Characterize the cellular and humoral elements of both innate and adaptive immune responses.
- Compare and contrast antigen recognition by cells of both the innate and adaptive immune responses.
- Describe the key features of an inflammatory response.
- Outline the functions of cytokines, chemokines and adhesion molecules during immune responses.
- Describe the main function of T cell subsets in the cell-mediated adaptive immune response.
- Describe the basis of antibody diversity.
- Appreciate the consequence of aberrant immune responses which may result in immunodeficiency, hypersensitivity or autoimmunity.

Introduction

The immune system has evolved to produce a coordinated response to protect the host from, and remove, ongoing infection

Key to this is the ability to distinguish self from nonself, whilst attempting to maintain the homeostasis of the body. Immunity can be categorized as being either **innate (nonspecific)** or **adaptive (acquired/specific)**. Inappropriate responses to either self or nonself can result in immune-mediated conditions such as **autoimmunity** or **hypersensitivity**. In addition, the importance of a healthy and effective immune response can be seen in individuals who have one of the many **immunodeficiency states**. They can present with a rainge of ailments, from minor recurring infections to life-threatening illnesses, depending on severity of condition. The immune system often seems too complicated. In effect the complex collection of cells and molecules has evolved to eliminate the almost infinite infectious organisms we might meet, whilst causing minimal damage to self, and ultimately resolving. Key to this is the ability to recognize the pathogen. The following sections will describe how the mechanisms involved in innate and adaptive responses differ.

Innate immune response

When activated, the innate response is seen as an inflammatory response

Innate immunity is the body's immediate response and first line of defense. The innate immune response protects an organism from attack, using physicochemical barriers, such as the skin and mucosal epithelia, and their associated secreted products, *e.g.* sweat, mucus and acid. It is often referred to as ‘natural’ or ‘nonspecific’, as it appears to be pre-existing. **When activated, the innate response is often seen as an inflammatory response.** Inflammation is the body's response to injury or tissue damage. Its purpose is to limit, and then repair, the damage brought about by any injurious agent. It involves the interaction of the microvasculature, circulating blood cells, other immune cells in the tissues, and their secreted effector molecules. Endothelial activation, increased vascular permeability and vasodilation allow the normally circulating leukocytes to migrate into tissue where, along with other tissue resident immune cells, they mount an effective and rapid response to try to eliminate the pathogen (Table 38.1). This will often include release of toxic mediators and phagocytosis, a process first described over 100 years ago by Mechnikov, who observed cells ‘eating’ the pathogens.

Table 38.1

Cells involved in inflammation

	Circulating	Tissue based
Polymorphonuclear leukocytes		
Neutrophil	Yes	Migrate as required
Eosinophil	Yes	Migrate as required
Basophil	Yes	Migrate as required
Mast cell		Yes
Mononuclear phagocytes	Monocyte	Macrophage
Lymphocytes (primarily part of adaptive response)	Yes	Migrate as required
Endothelial cells		Yes

Inflammatory mediators contribute to the immune response

Innate immune cells, when activated, can synthesize and secrete a wide variety of different soluble chemical substances termed inflammatory mediators. Some may be directly toxic to the pathogen, whilst others (cytokines) may be released in an attempt to signal to, recruit, and activate other immune cells, in order to help in the response. The liver also produces a number of these mediators, present in the blood, including acute phase reactants such as C-reactive protein (CRP) (Chapter 4) and components of the complement system described below (Fig. 38.1).

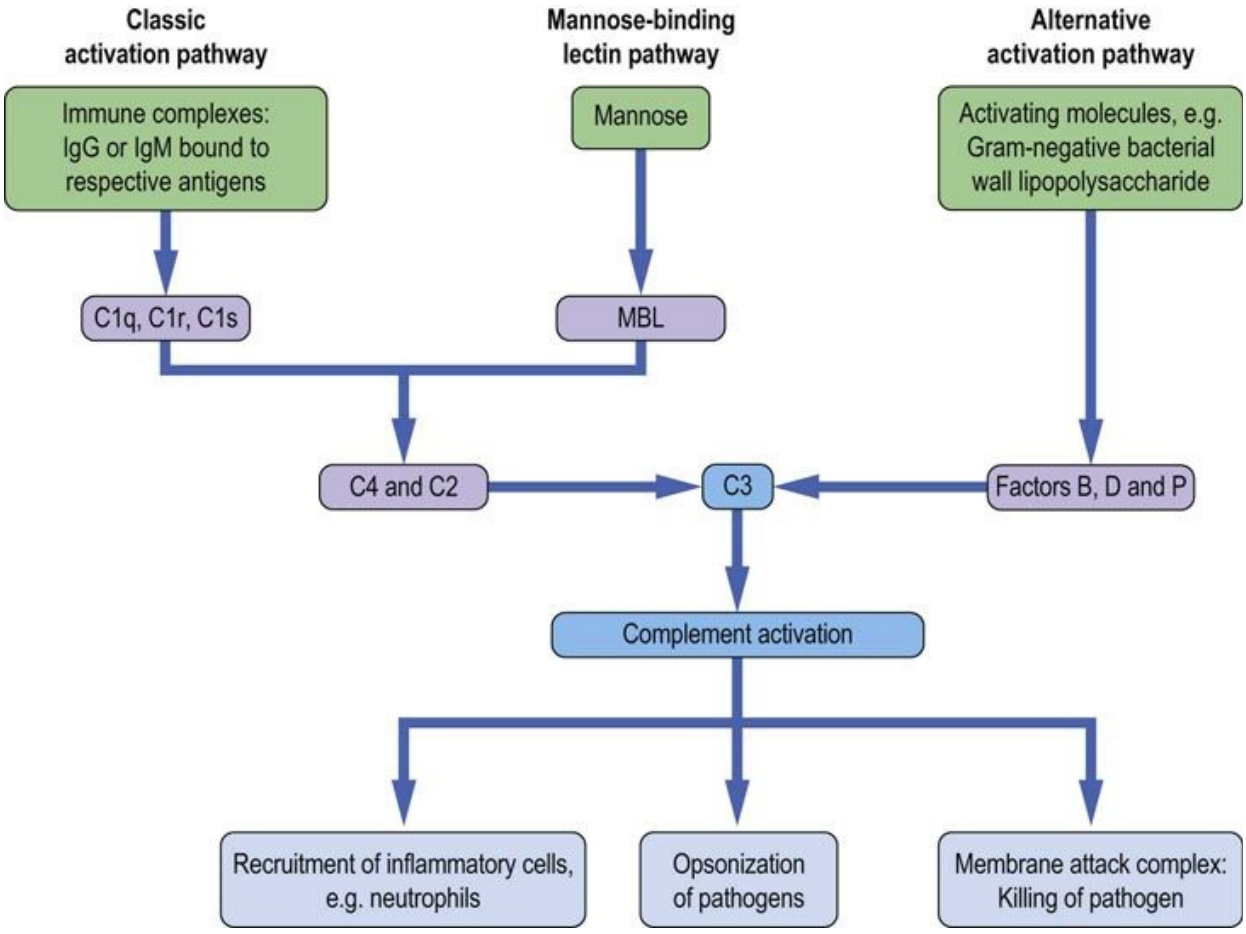


FIG. 38.1 The complement cascade.
Activating stimuli include surfaces that trigger complement activation, and to which the activated component can attach itself. Activation of complement recruits innate cells during early phase of an immune response. The late polymeric macromolecule (the membrane attack complex) can insert itself into the activating surface (the cell wall in the case of bacteria), breaching its integrity and causing osmotic lysis. MBL, mannose-binding lectin.

Cytokines

Cytokines are soluble mediators of inflammatory and immune responses

Cytokines are produced by a variety of cell types, including those of the innate and adaptive immune response. Covering a large number and different families, cytokines are small (usually less than 20 kDa) peptides or glycoproteins active at concentrations between 10^{-9} and 10^{-15} mol/L. In general, macrophages are often

their main producers during innate responses, and T cells during adaptive responses. However, many cell types, including all the cells of the immune system as well others, such as fibroblasts, epithelial cells and adipocytes, can secrete cytokines. By interacting with specific receptors on the surfaces of their target cells, the large number of cytokines now identified exhibit many effects. The majority act locally to their site of production (paracrine action) or on the very cells that produced them (autocrine action). A few, however, are capable of acting on cells more distant from their site of production (endocrine action). The cytokine network displays both redundancy and pleiotropy, with several having overlapping effects and an ability to act on numerous cell types. Cytokines have been grouped into subfamilies, based on structure and function, as discussed briefly below. Cytokine receptors are not restricted to cells of the immune system, being found widespread on disparate cell types. For more detail on cytokine signaling, see Chapter 40.

Cytokines may be classified into families by their principal effect:

- **Colony-stimulating factors:** as the name suggests, these are involved in the development and differentiation of immune cells from bone marrow precursors.
- **Interferons (IFNs):** while IFN-α and IFN-β have a role in inhibiting viral replication, IFN-γ regulates immune responses. The latter is made primarily by T cells and activates macrophages.
- **Interleukins (ILs):** currently there are in excess of 30 interleukins recognized, participating in regulating both innate and adaptive immune responses. They are made by a number of immune (and other) cell types, as the name suggests, the principal mode of action is communication between leukocytes.
- **Tumor necrosis factor (TNF) family:** this is a mixed collection of cytokines whose effects range from promoting inflammation (TNF-α and TNF-β) to stimulating osteoclasts and bone resorption (osteoprotegerin, OPG).
- **Chemokines:** these are a family of cytokines that bring about chemokinesis – movement in response to chemical stimuli. Interest has increased dramatically in the receptors for these mediators since some appear to act as co-receptors for infection (in particular HIV infection of $CD4^+$ T lymphocytes).

Previously it was fashionable to describe cytokines as either pro-or anti-inflammatory. It is now clear that this can be confusing given their pleiotropic effects. Comparing cytokine production in innate and adaptive responses and their cells of origin can now be considered.

- **During innate responses** macrophages, dendritic cells and natural killer (NK) cells are major producers of TNF-α, IL-1, IL-6, IL-8 and many chemokines, IL-

12, IL-15 and IL-18, IFN-γ (NK cells). These can all be thought of as important intercellular communicators, inducing inflammation and immune responses.

■ **If an adaptive or cell-mediated response develops**, T cells, especially $CD4^+$ T cells, become a major producer of cytokines. Their effects generally either promote or control further responses and they include IL-2, IL-4, IL-5, IL-10, IL-13, IL-17, IL-22, and TGF-β.

The complement system

Complement is activated in a series of sequential steps

The complement system, consisting of a series of proteins, both circulating and cell-membrane-bound, plays an important role in antimicrobial host defense. There are three pathways of complement activation. As part of the innate response, and in the absence of antibody, the **alternative** and **lectin** pathways activate complement during infection. For example, recognition of lipopolysaccharide found on Gram-negative bacterial cell walls will trigger the alternative pathway and mannose and other particular carbohydrates found in the cell wall of fungi, bacteria and viruses will trigger the lectin pathway. In addition, antibody produced during the adaptive response to infection can bind microbial antigens and activate complement via the **classic pathway**. The sequential activation of the cascade by proteolytic cleavage results in an autoamplifying response, producing a number of effector molecules involved in elimination of the microbial infection, as shown in Figure 38.1. The different pathways converge to produce a common outcome, whereby the late components combine with each other to form a multimolecular complex that can breach the integrity of the surface of infecting organisms by insertion into their membrane (the **membrane attack complex,** MAC). Fragments produced as a result of the complement cascade have distinct biological activities, which include the facilitation of phagocytosis (termed **opsonization**), the attraction of cells (**chemotaxis**), and stimulating the degranulation of immune cells (**anaphylatoxin** activity).

Cells participating in the innate response

Neutrophils and monocytes are recruited to sites of infection

Neutrophils and monocytes which are normally found circulating in the

bloodstream are recruited to sites of infection by the process of extravasation. Through the interaction of receptors on the phagocyte and counter-ligands on vascular endothelium, cells attach, arrest and move from the circulation to the infected tissue. Neutrophils are the most abundant leukocytes in the bloodstream, numbering 4000–10,000/mm^3. This increases rapidly during infection through recruitment from the bone marrow, and numbers often reach 20,000/mm^3. Neutrophils are generally the first cells to respond to infection, phagocytosing microbes in the circulation and moving rapidly into infected tissue. They are short-lived (normally a few hours to days) and die rapidly after reaching the tissue and exerting their effect by the process of apoptosis (Chapter 42).

Monocytes transform into macrophages, which are‘the dustbin of the immune response’

Monocytes are found in much lower numbers within the blood, 500–1000/mm^3, and, by contrast with neutrophils, are longer-lived. Similarly to neutrophils, they can also migrate into tissue, and on doing so they differentiate into macrophages. Macrophages have a number of key functions, including phagocytosis of infecting microbes, antigen presentation and general removal of dying or damaged host cells. Indeed, the macrophage has often been termed the ‘dustbin of the immune response’. Most organs of the body and connective tissue have resident macrophages, whose job it is to survey their environment for signs of infection. Their rapid identification of infection with the resultant release of numerous cytokines and chemokines, initiates the inflammatory response.

Neutrophils and macrophages recognize attacking microbes through their receptors

In order to mount an efficient response to infection, neutrophils and macrophages must realize that the body is under attack. They do so through a number of **germline-encoded cell surface and intracellular receptors** which, unlike the receptors used by cells of the adaptive immune response, are not produced by somatic recombination of their genes (Table 38.2). As a result, the response elicited by such receptors is amnesic, meaning that the cells will respond similarly on reinfection. The receptors involved in microbial recognition, and often termed **pattern recognition receptors (PRR)**, identify structures that are shared by various microbes, and which are generally not present on host cells, such as nucleic acids, lipids, sugars, proteins, or a

combination of molecules. Often the structures recognized by these receptors, called pathogen-associated molecular patterns (PAMPs), are conserved structural features required by the pathogen for survival or infectivity, and are generally common to particular microbial families.

Table 38.2

Comparison of antigen receptors of innate and adaptive immunity

Receptor characteristic	Innate immunity	Adaptive immunity
Triggers an immediate response	Yes	No
Receptors germline encoded	Yes	No
Specificity same across lineage	Yes	No
Broad spectrum of recognition	Yes	No
Encoded by multiple gene segments	No	Yes
Gene rearrangement occurs	No	Yes
Each receptor has unique specificity	No	Yes

There are four main categories of pattern recognition receptors, classified according to location and function

The first is mannose-binding lectin (MBL), which is actually not a cell-associated molecule, but rather a free circulating plasma protein, which on recognizing and binding to pathogen PAMP can activate the complement cascade via the lectin pathway (Fig. 38.1). The remaining are either surface-bound receptors that promote phagocytic function of the cell, or membrane-bound signaling receptors that found either on the outer cell membrane of the cell or within the cytoplasm.

Receptors are used by neutrophils and macrophages to promote phagocytosis

Microbes can be coated by soluble mediators of the immune response, such as complement components or antibodies, in the process termed **opsonization**. This makes the phagocytic process of microbial uptake by neutrophils and macrophages more efficient. Surface receptors used by neutrophils and macrophages to promote phagocytosis and killing of microbes include the mannose and scavenger receptors, complement receptors and Fc receptors. Mannose and scavenger receptors allow direct microbial recognition by phagocytes, whilst the complement and Fc receptors which recognize the opsonized microbe coated with complement components and antibodies, respectively, will also enhance phagocytic activity (Fig. 38.2).

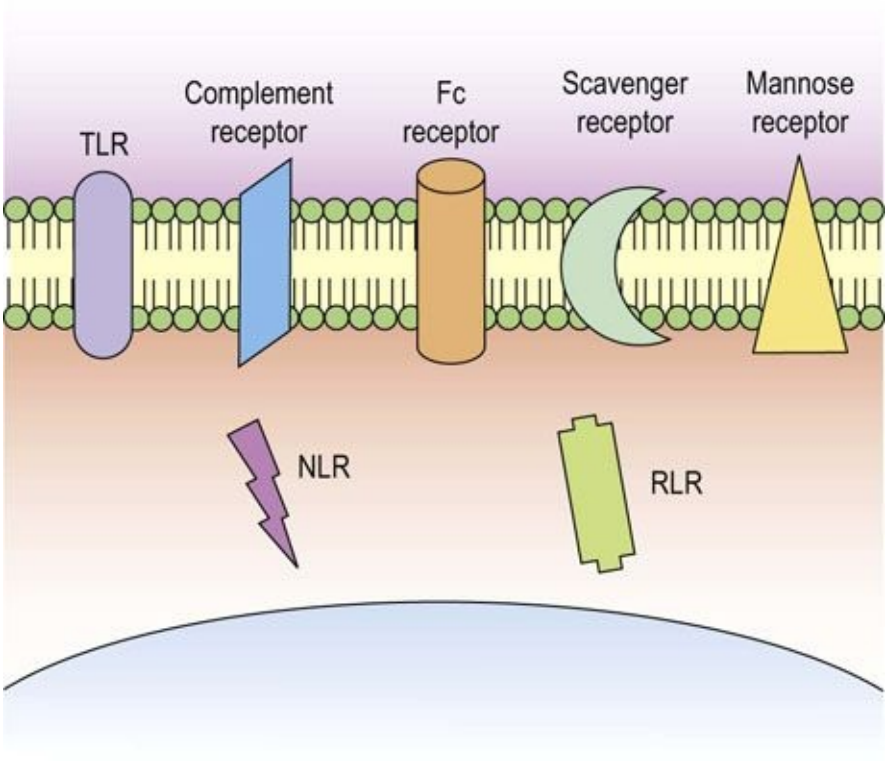


FIG. 38.2 Phagocytes utilize numerous receptors to detect pathogens.
Cells of the innate immune system express numerous receptor types, both on the cell membrane and intracellularly, to detect pathogen and initiate an effective immune response. TLR, toll-like receptor, NLR, NOD-like receptor, RLR, RIG-1-like receptor.

Signaling receptors are used by innate immune cells to trigger many of their functions

One of the best characterized signaling PRR families is the evolutionary conserved Toll-like receptor (TLR) system in mammals, named after a homologous receptor system used by the *Drosophila* fruit fly for protection from infection. In man, there are 10 expressed TLR genes (13 in mice), their products forming homo-or heterodimers with other family members, thus increasing the repertoire for recognition. TLR4, for example, has been shown to be the receptor recognizing lipopolysaccharide (LPS) found on the surface of Gram-negative bacteria such as *E. coli* but not present on mammalian cells. The effects of TLR activation can vary depending on need, and may include increased cytokine/chemokine production, enhanced phagocytosis, upregulation of costimulatory molecules on cell surface, cell migration and, if appropriate,

processing and presentation of pathogen antigens to T cells in order to activate an adaptive immune response. Table 38.3 summarizes TLR function and cellular distribution. TLRs can be expressed either on external cell membrane or on intracellular vesicles and function primarily to recognize extracellular pathogens. Whilst certain intracellular TLRs (TLR3/7/9) can detect viral RNAs and DNAs, they interact primarily with extracellular pathogen products entering by the endocytic pathway.

Table 38.3

Toll-like receptor (TLR) ligands and cellular distribution

TLR	Cellular expression	Ligands	Pathogen species
TLR1-TLR2 heterodimer	Monocytes, DCs	Zymosin	Fungi
TLR2-TLR6 heterodimer	NK cells Eosinophils Basophils	Lipoproteins Lipoteichoic acid β-glucans Lipomannans	Bacteria Gram (+ve) bacteria Bacteria and fungi Mycobacteria
TLR3	NK cells	Double-stranded RNA	Viruses
TLR4	Macrophages, DCs, eosinophils, mast cells	LPS	Gram (–ve) bacteria
TLR5	Gut epithelium	Flagellin	Bacteria
TLR7	DCs, NK cells, eosinophils, B cells	Single-stranded RNA	Viruses
TLR8	NK cells	Single-stranded RNA	Viruses
TLR9	DCs, NK cells, eosinophils, B cells, basophils	Unmethylated CpG (DNA)	Bacteria
TLR10	DCs, NK cells, eosinophils, B cells	Unknown	Bacteria

DC, dendritic cell, the main type of antigen-presenting cell; LPS, lipopolysaccharide; CpG, cytosine-guanine dinucleotide, NK cells, natural killer cells.

NOD-like receptors are located in the cytoplasm

Other pattern recognition receptors including the more recently described **NOD-like receptors** (NLR), are located within the **cytoplasm**. They act as intracellular sensors, ultimately triggering the NFκB pathway, which results in similar responses to those activated following TLR engagement. In the presence of certain invading pathogenic stimuli (bacterial, fungal, or viral), TLRs and NLRs cooperate, activating a cytoplasmic multiprotein complex called the **inflammasome**. The resulting caspase-1 activation leads to the processing and release of mature forms of pro-inflammatory cytokines, including IL-1 and IL-18.

A final family of intracellular signaling PRRs, the RIG-1-like receptors (RLR), which detect viral RNAs and stimulate antiviral responses through the

production of interferon (Fig. 38.2).

Activation of PRRs such as TLR/NLR not only induce efficient innate responses to deal with microbial infections but also may induce responses, which in turn lead to the full activation of the adaptive immune response.

Adaptive immune response

Specificity of the response is achieved through receptors that recognize antigen

Adaptive immune responses are brought into play if our innate defenses are unsuccessful, *e.g.* due to the persistence of the triggering agent. The response is initiated when the lymphocytes recognize components of the infectious agent. The infectious agents are called **antigens** and **their binding to receptors on lymphocytes triggers an adaptive response.** Receptors on B cells and T cells differ and see quite different forms of antigen.

Thymic education and self-tolerance help distinguish between self and nonself

Crucial for successful adaptive responses is the **ability to distinguish between self and nonself.** The immune system does this through the processes of thymic education and self-tolerance. This ensures that appropriate immune responses to infection develop. Failure of this process and inappropriate activation of the immune response by self-antigens can result in developing **autoimmunity**, presenting as *e.g.* **rheumatoid arthritis** or systemic **lupus erythematosus**.

Adaptive immune response needs time to develop and remembers what it sees

When an adaptive immune response is first initiated, relatively few cells and components are likely to be available that could react specifically with any chosen antigenic substance. There is a delay while these increase to a level which can ensure elimination of the antigen, or at least reduce the antigen to a level that would be manageable by the innate immune response.

The adaptive immune response employs a mechanism to remember a specific encounter, so that if the same foreign antigen is encountered again, it can be dealt with more quickly and effectively. This is called **immunologic memory.** Thus, in comparison to innate immunity, the adaptive response exhibits both specificity for and memory to the foreign or nonself antigen.

Cells primarily responsible for adaptive response are the

lymphocytes

Adaptive immunity is mediated, similarly to the innate response, by cellular and humoral elements. The cells primarily responsible are **the lymphocytes**. There are two major types of lymphocytes circulating in blood and present within the lymphoid tissues:

- **T cells**, which are responsible for a number of cellular responses.
- **B cells**, which are responsible for humoral responses, *i.e.* antibody production.

In addition to the lymphocytes, **professional antigen-presenting cells (APCs)** are required for the initial efficient activation of the T lymphocytes.

T and B lymphocytes

T and B lymphocytes carry particular collections of cell surface markers that can assist in assigning their lineage

The effector cells primarily involved in the adaptive immune response are the T and B lymphocytes. In total, lymphocytes are present in the peripheral blood at 1.3–4.0 × 10^9/L. Of these, approximately 50–70% are T cells and 10–20% are B cells.

A third population termed **'natural killer' (NK) cells**, so-called because they demonstrate the ability to kill neoplastic or virally infected cells without prior exposure or sensitization. They are atypical lymphocytes and are generally considered to be part of the innate response. Identification of T and B cells is based on immunophenotypic or functional studies. They carry particular collections of cell surface markers that can assist in assigning their lineage. The distinction between T and B cells is most easily made with reference to their antigen receptors (Fig. 38.3).

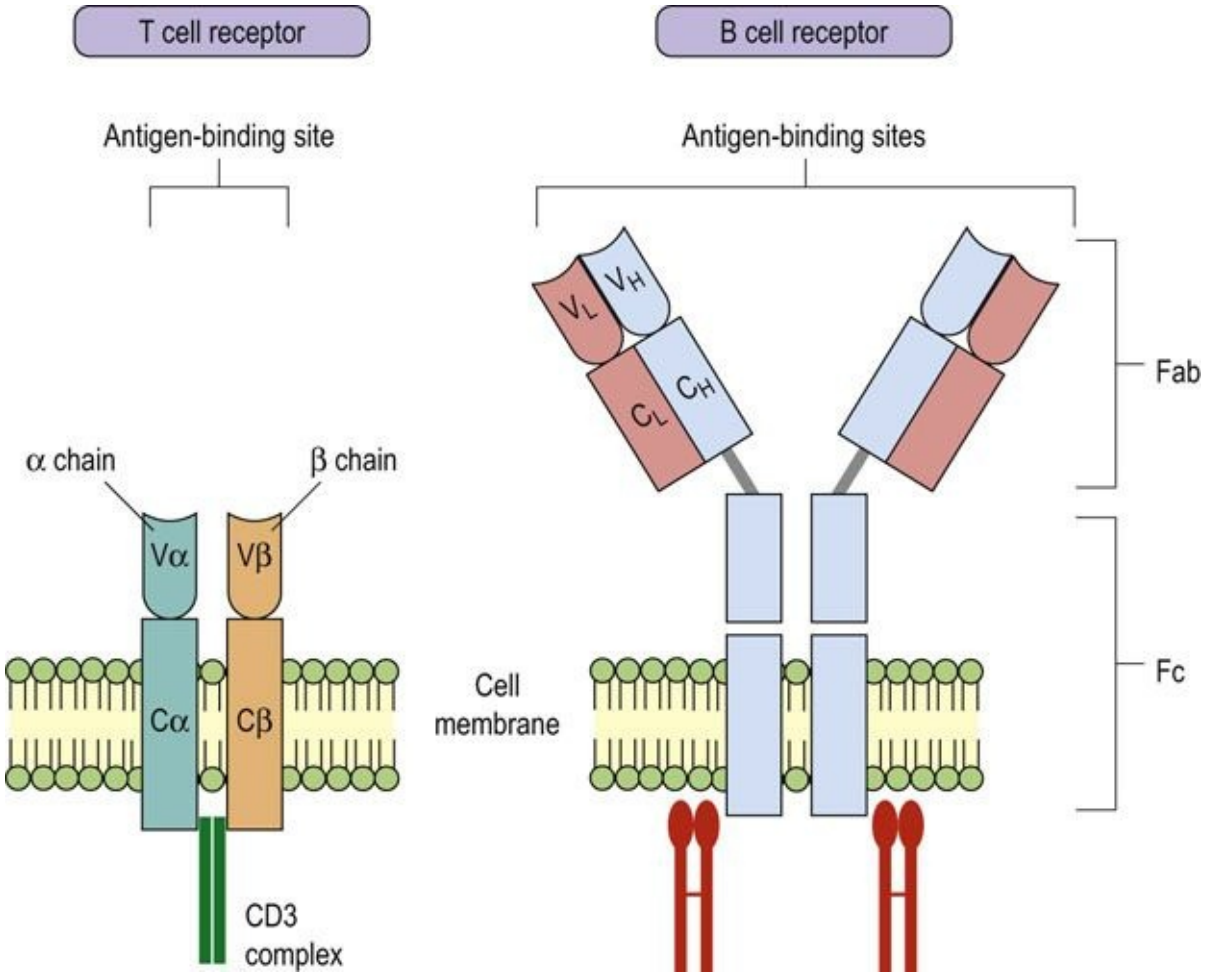


FIG. 38.3 Similarity in structure of T and B cell antigen receptors.
The receptors used by T cells and B cells to detect antigen share structural similarity. V, variable regions; C, constant regions.

Insert new para:The surface markers on lymphocytes and other immune cells are classified according to the **cluster of differentiation (CD) system** based on antibody-binding characteristics of these molecules.

B and T lymphocytes are activated by binding of antigen and by costimulatory molecules

The antigen recognition receptor on the B cell is a **surface immunoglobulin** termed ‘sIg’. On binding to its antigen, it brings about the cell's activation, and subsequent proliferation and differentiation. In addition to sIg, B, cells express several other markers, the best characterized of which include CD19, CD20 and the **major histocompatibility complex (MHC) class II molecules**.

The T cell antigen receptor is termed the T cell receptor (TCR) and it is complexed with CD3

Two other CD markers whose expression appears to be mutually exclusive on T cells are the CD4 and CD8, and they are useful in further categorizing the T cell function.

NK cells are currently identified by the expression of the combination of CD16 and CD56

These markers are often used in flow cytometric technology using fluorescent monoclonal antibodies to identify cell types.

Another group of surface molecules, the costimulatory molecules, are found on the surface of T and B cells

Following exposure to antigen, CD28 on T cells will bind CD80/CD86 on the APC, while CD40 on B cells will bind CD40L on T cells, resulting in full activation. Without such costimulation, T and B cells would not be fully activated following exposure to antigen and could become anergic, *i.e.* nonresponsive.

Molecules involved in antigen recognition

Antigen is recognized by specific receptors on T and B cells

The ability to recognize the enormous number of possible antigenic configurations is achieved by differences in amino acid sequence of these receptors, which gives rise to differences in their shape or conformation. The antigen and its specific receptor have a 'hand-in-glove' relationship. Both T and B cell antigen receptors show marked variability in the sequence of amino acids that come into contact with the antigen, while other parts of these molecules are relatively constant with regard to their amino acid sequences.

Unlike the antigen receptors found on innate cells, which are germline encoded, the receptors found on the T and B cells are generated by random recombination of receptor genes during cell maturation. These antigen recognition receptors are clonally distributed. As a result, each clone will exhibit unique specificity for a particular antigen, thus generating the enormous pool of cells capable of responding to all antigens.

As mentioned previously, T and B cells differ in what they recognize as 'foreign'. The sIg antigen receptor found on B cells is capable of recognizing macromolecules (proteins, polysaccharides, lipids, etc.), whereas T cell receptors recognize small peptides of proteins previously processed by the APC.

Although the number of T and B cell clones, each recognizing different antigens, is enormous, the engagement by appropriate antigen will generally induce a similar response, *i.e.* **signal transduction**. This may lead to full cell activation resulting, for B cells, in antibody production, and for T cells, in the proliferation and promotion of the cellular adaptive immune response.

The T cell antigen receptor

The T cell antigen receptor (TCR) resembles the binding portion of an immunoglobulin molecule

The TCR is a heterodimer made up of two nonidentical polypeptide chains termed α and β (see Fig. 38.3). In addition, a small unique T cell population

found primarily within the gut expresses alternative TCRs, their chains being termed γ and δ. Each chain of the TCR comprises two domains – one **constant** and one **variable** amino acid sequence. The antigen-binding site of the TCR is in the cleft formed by the adjoining single *N*-terminal variable domains of the constituent α (Vα) or β (Vβ) chains. The effector function of the constant domain in each of the antigen receptor chains is signal transduction. The two chains come into close contact via the covalent bonds between the variable domains and noncovalent hydrophobic interactions between the opposing faces of the constant domains. Structurally, the TCR resembles the binding portion of an immunoglobulin molecule, the antigen receptor found on B cells, but it is quite distinct, being the result of different gene products.

Major histocompatibility complex

The MHC is responsible for how T cells ‘see’ an antigen against a background of self

For an immune response to be initiated, antigen cannot simply bind to the nearest T cell but must be ‘formally’ presented to the immune system. This occurs when APCs express processed antigenic peptides bound within grooves of MHC molecules on their cell surface. The MHC class I and II molecules also provide a differential mechanism for processing antigens that originate from within cells, *e.g.* viruses, and those that arise from the extracellular environment, *e.g.* many bacterial antigens. The different class MHC molecules lead such antigens through different pathways to interact with the immune system, in particular with the T cells, on the basis that each will be better dealt with by differing effector systems: class I leads to $CD8^+$ T-cytotoxic responses, and class II instructs $CD4^+$ helper T to provide appropriate help to B cells for an antibody-mediated response.

The MHC complex of genes is grouped into three regions, termed class I, II and III

The MHC complex of genes is found on the short arm of chromosome 6 and is grouped into three regions, termed class I, II and III, with the same nomenclature being applied to the respective polypeptide products (Fig. 38.4). Key to the success of the adaptive immune response is the **polygenic and polymorphic nature of the MHC**. By this we mean that there are various different MHC class I and II genes, and for any one, multiple variants, or alleles, can be expressed. Class I and II molecules are directly involved with immune recognition and cellular interactions, whereas class III molecules are involved in the inflammatory response by coding for soluble mediators, including complement components of the innate response and TNF.

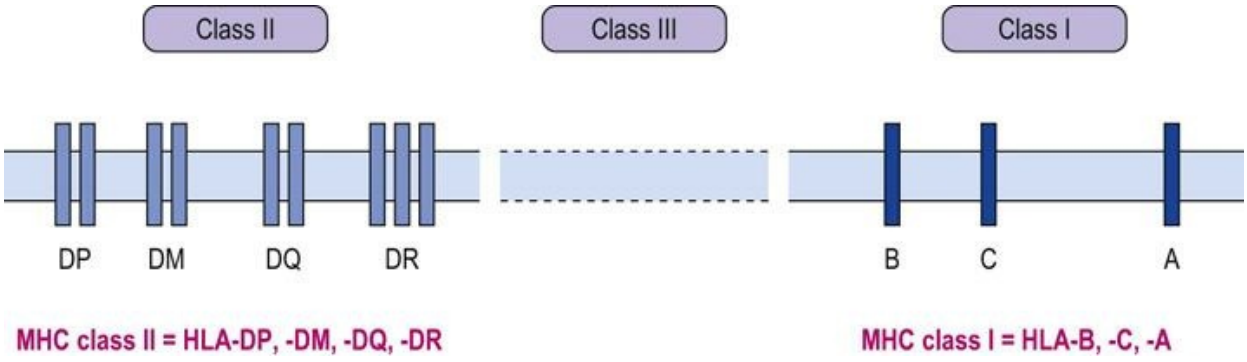


FIG. 38.4 Genetic organization of the MHC and expressed products.
Genes of the MHC in humans are located on chromosome 6. The gene products are the human leukocyte antigens (HLA).

MHC class I genes are organized into several loci, the most important of which are those termed HLA-A, HLA-B and HLA-C

Alleles are transmitted and expressed in Mendelian codominant fashion. Owing to their closeness on the chromosome, they are inherited en bloc as parts of a haplotype and are expressed on the surface of all nucleated cells. The α-chains they encode have three domains, one of which is structurally similar to those found in immunoglobulin molecules, but the other two show significant differences. The α-chains combine with β2-microglobulin to give rise to a functional class I molecule (Fig. 38.5).

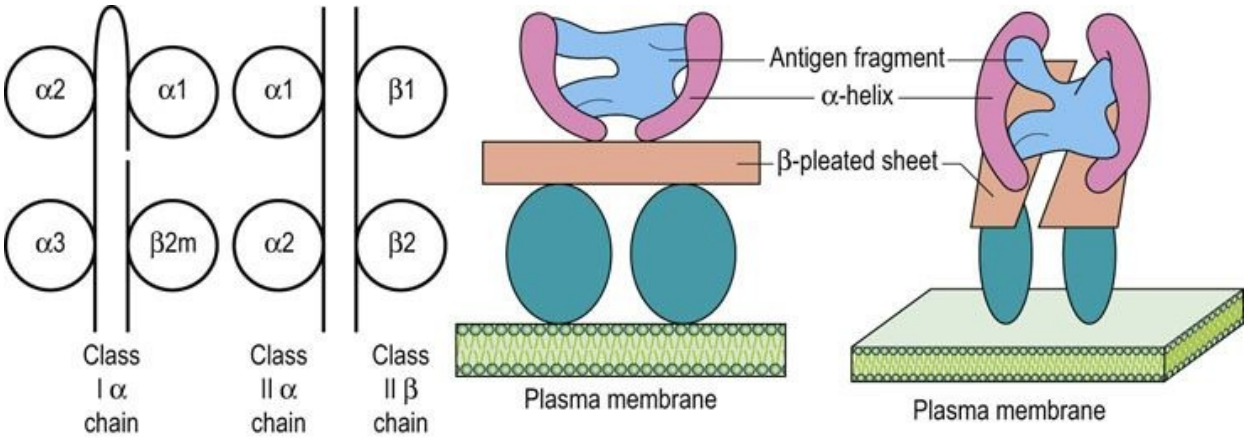


FIG. 38.5 Class I and II MHC (HLA) structure.
On the left there are class I and class II MHC molecules. In class I molecules β2-

microglobulin (β2m) provides the fourth domain. On the right, the protein conformation and folding of the MHC molecules shape the binding groove for antigenic peptides.

Class II genes are HLA-DR, HLA-DQ, HLA-DM and HLA-DP

The class II subregion genes, termed HLA-DR, HLA-DQ, HLA-DM and HLA-DP, are organized into α-and β-loci, giving rise to α-and β-polypeptide chains, respectively. Both are of approximately the same molecular weight and combine to form a heterodimer with a tertiary structure similar to a class I molecule, with a peptide groove into which the processed antigenic fragment is inserted during antigen presentation (Fig 38.5). Unlike class I expression, class II is far more restricted, being expressed mainly on professional APCs, such as dendritic cells, as well as on macrophages and B cells.

Many (currently in excess of 1000) allelic variants can be identified in each of the loci associated with antigen presentation. There are six major loci, each having between 10 and 60 functionally recognizable alleles, and as each parent passes on one set or haplotype on each chromosome, it is easy to appreciate that the likelihood of another individual in the same species having an identical set is remote.

The B cell antigen receptor

The B cell antigen receptor (BCR) is a membrane form of the immunoglobulin molecules found circulating in serum

Immunoglobulins are Y-shaped molecules made up of four polypeptide chains (Chapter 4, Fig. 4.5) – a pair of heavy chains each of approximate molecular weight 150 kDa and a pair of light chains each of approximate molecular weight 23 kDa. The arms interact with antigen and their structure is based on immunoglobulin domains with constant and variable sequences of amino acids in both the heavy and the light chains. It is the variably sequenced amino-terminal domains of both the heavy and the light chains that form a pocket that constitutes the antigen-binding site; the ‘fragment antigen binding’ (**Fab**) portion sits at the end of the arms (Fig. 4.5). The remaining relatively constant amino acid sequence domains of the chains are termed **constant heavy (C_H)** or **constant light (C_L)** and form the stem (**Fc portion**) that has a number of functions, including binding complement components and binding to Fc receptors on various leukocytes including macrophages, NK cells, neutrophils,

mast cells and B cells.

There is an almost infinite range of possibilities for antibody specificities

The receptor repertoire, which most likely has in excess of 10^{11} different specificities, occurs as a result of the process whereby the various genes are involved in making the molecule combine. The variable region of a light chain is the product of two different genes (V = variable, and J = joining). This in turn combines with the gene product for the constant region, giving the complete transcribed and translated light chain protein. For the heavy chain, the level of complexity is increased with the addition of the D (diversity) gene product, forming part of the variable area in addition to V and J gene segments. Again these will combine to the C region segments, but for heavy chains multiple C genes products make up the completed protein. Multiple copies of each of the gene segments in germline DNA, which are in turn used randomly, as well as the possibility for polymorphisms in individuals, result in the almost infinite range of possibilities for antibody specificities. Mature B cells have the capacity to accumulate small point mutations in the DNA encoding the heavy and light chains of immunoglobulin, termed **somatic hypermutations**, which add further variation to specificity of the recognition process.

Lymphoid tissues

Primary (central) lymphoid tissues

Lymphocytes originating from common bone marrow-derived hematopoietic stem cells are initially found within primary lymphoid tissue, where they undergo early development and differentiation.

Maturation of most B cells occurs within the bone marrow

Initially, progenitor B cells rearrange their immunoglobulin genes. They do so in an antigen-independent process by interacting with stromal cells within the bone marrow. The resulting immature B cell expresses **surface IgM** as an antigen receptor. If they interact too strongly with environmental antigens at this stage they are removed by the process of negative selection, thus limiting the chance of autoreactivity. Following exit into the periphery, the B cells will express both **surface IgM and IgD** and can be activated by antigen engagement. These cells will proliferate, some becoming antibody-secreting plasma cells and others long-lived memory cells.

T lymphocyte progenitors travel to the thymus where they develop into T lymphocytes

The thymus is a multilobed structure found in the midline of the body just above the heart. At the macroscopic level, there is an outer cortex and an inner medullary area within each lobule. T cell development progresses in the thymus as the immature T cells migrate from the cortex to the medulla. The immature T cells interact with thymic epithelia and dendritic cells. These cells are thought to be responsible for the processes of positive and negative selection that take place as part of the ‘thymic education of T cells’. During this processes the T cells are assessed for their ability to interact with self-MHC and, if appropriate, they receive survival signals. Cells which show excessive reactivity to self, receive signals leading to their deletion whilst still within the thymus. This removes **autoreactive** cells, which, if released into the periphery, could potentially induce autoimmunity. The development of both early T and B cells in the primary lymphoid tissues is independent of extrinsic antigen stimulation.

Secondary (peripheral) lymphoid tissues

The secondary lymphoid tissues comprise lymph nodes, spleen, and the mucosa-associated lymphoid tissues (MALT)

These tissues are functionally organized throughout the body and have in common a degree of compartmentalization, with specific areas for T cells and B cells, and areas of overlap where they interact and respond to antigen. It is at these sites that immune reactions actually develop. For example, on exiting the thymus, the naïve T cells will recirculate via the bloodstream, and enter the lymph nodes by appropriate upregulation of adhesion molecules and chemokine receptors, which allows them to localize in the T cell areas of the tissue.

Within the lymph node, the T cell area is the paracortex and the B cell area are the follicular areas of the medulla

Here follicular structures of two types can be found: the unstimulated primary follicle, and stimulated secondary follicles, characterized by the presence of germinal centers. **Lymph, which drains from the tissues to the lymph nodes will carry antigens, which in turn can be sampled by the APCs for presentation to the lymphocytes.** In addition, peripheral APCs which have encountered pathogen will have migrated to their nearest draining lymph node, processing the antigen on the way in the hope of activating the T cell with appropriate specificity. On activation, the T cell will again alter chemokine receptor expression and leave the lymph node to recirculate to the site of infection where it can induce an effector response.

The spleen, contains nonlymphoid tissue (the red pulp) as well as lymphoid areas, the white pulp

Within the white pulp, follicular B cell areas are evident and the T cell areas lie between them in the interfollicular space. The spleen is used by the immune response for the presentation of blood-borne antigens.

MALT comprises the lymphoid elements adjacent to the mucosal surfaces

They are found at the entrance to the respiratory tract and gut, and include the

tonsils and adenoids. Further down the digestive tract, unencapsulated aggregates of lymphoid cells referred to as Peyer's patches are found, overlaid by specialized areas of epithelium for sampling the antigenic environment. Similar to the lymph nodes and spleen, these tissues are important for **initial antigen sampling and presentation**, in particular for antigens which enter the body through a breach of the epithelium or via the gut.

Antigen-presenting cells

APCs are specialized cells which display microbial antigens on their surface to allow T cell activation

Dendritic cells are the major APC and are found throughout the body. The skin and different organs have their resident population of such cells. Dendritic cells can migrate throughout the body from tissue to circulation, and en route may enter the specialized secondary lymphoid organs such as lymph nodes (where they may activate lymphocytes for an adaptive response). On uptake of antigen, APCs can process and re-express it, in the context of the MHC, on the cell surface, to allow **presentation to the T cell**. Dendritic cells are termed 'professional APC', as in addition to being able to present the antigen, they also possess a number of other cell surface molecules, *e.g.* CD80/86, which provide the additional signals, so-called 'costimulation'. These signals are required by a naïve T cell for complete activation. In addition, they may release certain cytokines *e.g.* IL-12, which will influence T cell activation and differentiation. Other cells which also are capable of presenting antigens, and hence can be considered APCs, include macrophages and B cells.

Adhesion molecules

Adhesion molecules mediate adhesion between cells

Cellular interactions during an immune response are dependent on the expression of the molecules and ligands that mediate adhesion between cells or between cells and the extracellular matrix. These are termed 'adhesion molecules'. They are found on a wide variety of cell types, not only cells of the immune system but also, for instance, on vascular endothelium (Chapter 18). A major determinant of their expression is the prevailing cytokine environment and the surrounding connective tissue matrix. Typically, they are **transmembrane glycoproteins**. They deliver intracellular signals, and during immune responses are primarily involved in promoting cell–cell interactions and cell migration. The migration includes the movement of innate cells from blood to tissue during infection, as well as aiding lymphocytes to enter and leave lymph nodes as they circulate, looking for activation signals resulting from antigen presentation in these peripheral organs. Adhesion molecules involved in immunity are grouped

into three major families:

■ **Integrins.** These are heterodimeric proteins expressed on leukocytes, such as lymphocyte function-associated antigen 1 (LFA-1), or the macrophage adhesion molecule 1 (MAC-1).

■ **Immunoglobulin supergene family adhesion molecules.** These are often expressed on endothelial cells, *e.g.* intercellular adhesion molecule 1 (ICAM-1;CD54), or platelet/cell adhesion molecule 1 (PECAM-1; CD31).

■ **Selectins.** They are expressed on leukocytes and endothelial cells, *e.g.* L-selectin or P-selectin.

■ **Mucin-like vascular addressins.** These are often found on leukocytes and endothelium. They bind selectins.

Reaction with, the response to, and the elimination of, antigens

On binding to the antigen, the cell differentiates into progeny with an effector function or a memory function

On successful antigen binding, the activated lymphocyte undergoes repeated division or proliferation. Differentiation follows, which can lead to either the development of an **effector function** or the generation of **memory**.

Clonal selection creates clones of identical cells with unique antigen specificity

Clonal selection is the process whereby the immune response creates clones of identical cells, each clone having unique antigen specificity. With this clonal repertoire, the antigen determines which specific lymphocyte will be activated. The process of antigen drainage and lymphocyte recirculation to the peripheral lymphoid tissue ensures that **antigen is inspected by many lymphocytes** and can select for proliferation and differentiation the cell that bears a **specific and reciprocal antigen receptor**. Clonal selection ensures not only an adequate number of effector cells to deal with the threat at the time of initial stimulation but also a suitable number of part-primed memory cells that will be able to complete their activation more rapidly on subsequent antigen exposure. Figure 38.6 shows events for B cells, but T cells also undergo a similar process, resulting in proliferating clones of primed effector cells, and memory cells being generated for subsequent responses.

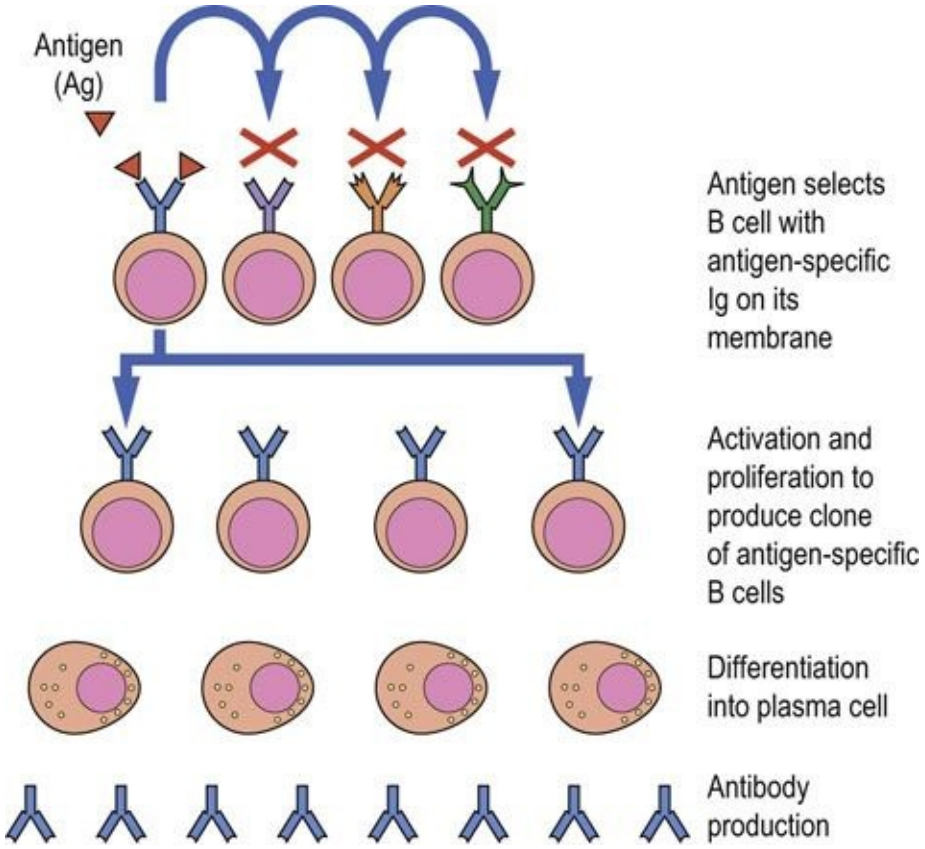


FIG. 38.6 Clonal selection in B cells.
Antigen-specific (Ag-specific) surface immunoglobulin (sIg) on the B cell membrane has a shape reciprocal to the antigen. Antigen-immunoglobulin binding leads to activation and proliferation to produce a clone of antigen-specific B cells. Each member of the specifically activated clone then undergoes differentiation into a plasma cell, which produces and secretes large quantities of a single homogeneous immunoglobulin, with specificity identical to the sIg that triggered the response in the first instance.

Immunologic memory distinguishes the adaptive immune response from the innate response

How immunologic memory is generated is still the subject of much research. On re-exposure to the same antigen, the adaptive immune response, due to the reactivation of long-lived memory cells, mounts a more rapid and more effective response, compared to the primary response. **The long-lasting protection offered by vaccination is a result of immunologic memory.** There are clear differences between the way naïve and memory lymphocytes respond to antigen. For example naïve and effector cells are relatively short-lived but memory lymphocytes persist for years and as a result often give lifelong protection after

initial exposure. Additionally, there are more memory cells compared to naïve cells specific for the same antigen.

Adaptive response is an integrated response

Adaptive immune response is mediated by cellular and humoral elements, T cells being considered responsible for cellular immunity and B cells for humoral immunity. It is important to consider the adaptive response as being an integrated response, not occurring in isolation. For example many of the events and functions of T cells will impact on how efficient B cells respond and make appropriate antibodies. Similarly, B cells can in turn activate T cells. Interestingly, this integrated response reflects back on innate immunity too. For example, the cells of the innate immune response have evolved to be more efficient when antibodies have been produced, and may respond to cytokines released by the lymphocytes.

T cell response

Distinct populations of T cells exist. All T cells, once they have left the thymus, express either CD4 or CD8 on their surface. This phenotypic distinction also has major consequences for effector function: $CD4^+$ T cells are often called **T helper cells (TH)**, whilst $CD8^+$ cells are **cytotoxic T cells (CTL)**. TH cells can be further subdivided. They were originally divided into TH1 and TH2 cells, but in recent years evidence has shown this to be too simplistic. There is much current interest in the subset being termed TH17 due to their release of the cytokine IL-17. T cells are also responsible for regulating the activities of the other arms of the immune response. They achieve this by direct cell–cell contact or by the secretion of soluble mediators that interact with the relevant cells, *i.e.* other T cells, B cells, cells involved in the innate immune response such as macrophages, or cells of other tissues. The different subsets of T cells are shown schematically in Figure 38.7.

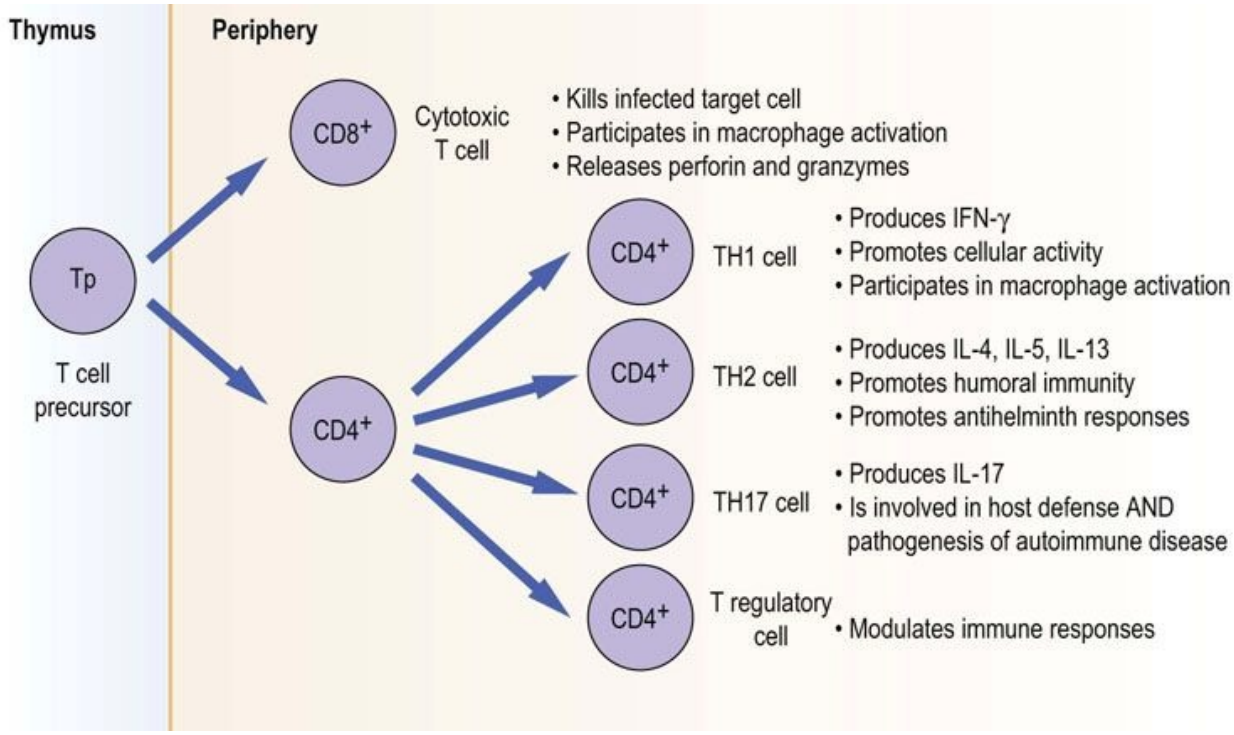


FIG. 38.7 Functional T cell subsets.
T cell precursor cells within the thymus develop into cells with different effector functions.

T helper cell subsets: TH1/TH2, TH17 and T regulatory (Treg)

The effector functions of $CD4^+$ cells appear to be largely in 'helping' other immune responses. We have said above that T cells need to be presented with antigen in the context of MHC on the surface of an APC: $CD4^+$ T cells this is done by MHC II molecules.

These T cells see antigen peptides which have been processed by the APC and expressed within the binding grooves of the MHC molecule. They also receive costimulatory signals when receptors on the T cell surface bind to counter-ligands on the surface of the APC. An example of this is the interaction between CD28 on the T cell and CD80/86 on the APC. Once activated, the T cell will differentiate, proliferate and perform appropriate effector functions.

TH1/TH2 cells

The subdivision into TH1 and TH2 subsets was originally made on the basis of their apparent function. TH1 cells appeared to function as promoting cellular responses. Once activated, they could release IFN-γ, which in turn further promotes macrophage activity. In addition, they may release TNF-α, which, through endothelial activation and subsequent upregulation of adhesion molecules and chemokines, would promote further leukocyte recruitment. In addition, they may provide help to B cells, enhancing antibody production.

TH2 cells help cellular responses in a different way. They too appear to help B cells make antibody, and produce IgE through their release of IL-4. They also preferentially stimulate eosinophil-driven inflammation through IL-5 production. Together these promote the major antihelminth response. By releasing IL-4 and IL-13, TH2 cells limit the TH1 activation of macrophages, and similarly TH1 products will inhibit TH2 responses. Therefore the effector functions of TH cells appear to be determined by the cytokine environment they produce and the response required. Finally expression of distinct transcription factors appears to be crucial in driving either TH1 or TH2 development with T-bet transcription factor being preferentially expressed by TH1 cells and GATA-3 is by the TH2 subset.

TH17 cells

Recently there has been much interest in another subset of TH cell which does

not fit with the original TH1/TH2 paradigm. The TH17 cell was originally identified in animal models of a number of autoimmune diseases, including multiple sclerosis, rheumatoid arthritis and inflammatory bowel disease. Understanding the potential role of this subset, particularly in human disease, has been a major research focus. It is known that in the presence of IL-6 and TGF-β, but the absence of IL-4 and IL-12, the default will be for developing CD4 T cells to become TH17. They will also require IL-21, itself a T cell product, and IL-23, produced by APCs. TH17 cells also appear to make IL-22. Clearly this is a complicated network and the finer points are still under investigation. Whilst the understanding from animal models has implicated TH17 in driving the pathology of several autoimmune diseases, it would appear that the natural, physiologic role for TH17 may be in host defense against certain bacterial pathogens, including *Klebsiella pneumoniae* and fungi such as *Candida albicans*. The IL-17 and IL-22 they produce likely act on local stromal and epithelial cells of infected tissue, promoting local production of chemokines (IL-8), which, in turn, recruits innate effector cells such as neutrophils. This again shows the close interaction between innate and adaptive responses for effective immunity.

T regulatory cells

Originally cells which can control a cell-mediated response were identified as **'suppressor cells'**. However, the original description of a $CD8^+$ cell providing this function is no longer considered valid. The cell types now studied are the so-called T regulatory cells (Treg). This appears to be a heterogeneous group. The most studied is a $CD4^+$ T cell which appears to be able to control the action of other immune cells through a combination of soluble mediator release (IL-10 and TGF-β) and direct cell–cell contact. These cells are thought to also express the transcription factor FoxP3, which appears to be crucial for T regulatory cell development. As discussed earlier, the thymus plays an important role in deleting autoreactive T cells before they exit into the periphery by the mechanism of central tolerance. This process is not 100% efficient, and it is now recognized that T regulatory cells play an important role in the process of peripheral tolerance (i.e. the holding in check cells within the circulation, which if allowed would be autoreactive and cause autoimmunity). The translation to clinical applications of preventing autoimmune disease is currently under investigation. The potential of the TReg cells may also be applicable to induce tolerance to organ grafts.

$CD8^+$ cytotoxic T cells (CTL) kill infected cells

The other major population of T cells based on their surface expression of CD8 is known as the cytotoxic T cell. Their role is **primarily to kill infected cells** (e.g. by virus). $CD8^+$ T cells recognize peptides of the antigen associated with MHC I on the surface of the infected cell. By doing this, they are able to limit infection. It is likely that initially, naïve $CD8^+$ T cells require help from $CD4^+$ T cells, as well as being activated by APCs to become effector cells. Once in the periphery, effector CTLs will become activated when they encounter virally infected cells. They will bind tightly to the infected cell using adhesion molecules: the main method of killing infected cells is by the calcium-dependent release of serine proteases known as granzymes, and perforins proteins capable of perforating cell membranes, from their granules. The result of delivery of these enzymes to the infected cell is activation of caspase-driven apoptosis (Chapter 42). Apoptotic cells are subsequently removed by innate phagocytic cells such as macrophages – the ‘dustbin’.

The adaptive humoral immune response

Humoral immune response is characterized by the release of antibodies from fully matured plasma B lymphocytes

Humoral or antibody-mediated specific immunity is directed at **extracellular infection**, especially bacteria and their products, extracellular parasites, and also at the extracellular phase of viral infection. Antibodies also play a major role in the immunopathogenesis of many autoimmune or aberrant responses resulting in hypersensitivity. The humoral immune response is characterized by the release of antibodies from fully matured plasma cells of the B lymphocyte lineage. As antibodies recognize many types of molecules, including polysaccharides and lipids, this response is particularly efficient against extracellular pathogens. The antibody binding to structural surface components of microbes blocks the adhesion of these bacteria or viruses, and prevents the harmful effects of their toxins in a process termed **neutralization**. However, simple antibody binding, in most situations, will not guarantee elimination of the antigen. To promote the response, the non-antigen-binding fragment of the molecule (Fc portion) is able to activate other components of the innate system, through complement activation or by binding to receptors on phagocytes. The diversity of effector functions is achieved by the genetic recombinations of the heavy-and light-chain genes, as described previously.

B cell subsets are involved in the humoral immune response

Similar to the cellular response, which is mediated by a number of T cell subsets, the humoral response uses distinct B cell subsets. As noted earlier, T cells, particularly the TH cells, interact with B cells both directly and indirectly via cell surface receptors and cytokines, respectively. This happens to such a degree that effective B cell responses are often described as being T cell-dependent. The B cells termed **B-2** are found in the follicles of the secondary lymphoid organs. They typically respond to protein antigens and produce the high-affinity antibodies typical of effective humoral responses. Within the marginal zone of the spleen, there is another population of B cells, which typically responds to polysaccharide antigens delivered via the bloodstream. They tend to secrete IgM

but can class-switch to produce IgG. Another population, termed **B-1**, which express similar receptors, account for around 5% of all B cells, is found in mucosal tissue and the peritoneum. They express surface IgM as their antigen receptor and little surface IgD (the pattern opposite to that of the classical B-2 follicular B cells). B-1 cells predominantly show an IgM response, typically to nonprotein antigens, undergo little somatic hypermutation and exhibit little memory development.

Clinical box A young man who developed sudden stridor and widespread nettle rash

anaphylactic shock

A young man was brought into the emergency room in a state of shock with stridor (a high -pitched sound on inspiration) and widespread urticaria (nettle rash). A companion told the admitting medic that the patient had developed difficulties in breathing shortly after eating a snack. Allergy to peanuts was suspected and a diagnosis of **anaphylaxis** was made. An intramuscular injection of epinephrine was given promptly and also treatment with intravenous antihistamine, corticosteroid, and cardiorespiratory support. The man recovered.

Comment.

While the physiologic role of the IgE response is considered to be protection against parasite infestation, this response is seen to be subverted in those who experience **atopic diseases** and **anaphylaxis**.

The major fraction of IgE is bound via Fc receptors to mast cells in the tissues. When antigen binds and crosslinks its specific IgE on the mast cells, it triggers their degranulation and release of preformed mediators (principally histamine). When the **mast cell degranulation** is localized to one site, it usually gives rise to only localized reactions such as allergic rhinitis and asthma. If the degree of sensitization with the antigen-specific IgE and/or the antigenic burden is greater, systemic degranulation can occur, with

consequent **anaphylactic shock**. Significant vasodilatation takes place, reducing the blood pressure. This is accompanied by large increases in vessel wall permeability, leading to substantial swelling, which particularly affects the skin and other loose connective tissue such as those in the larynx. Smooth muscle spasm also occurs, leading to bronchoconstriction with consequent respiratory difficulty and wheezing. These features are accompanied by increased secretory activity of seromucous glands in the respiratory and gastrointestinal tract as well as itching of the skin.

Antibodies illustrate the capability of the immune system for diversity

For T cell-dependent responses, re-exposure to antigen will induce a secondary antibody response. The higher levels of the produced antibody will have increased affinity and avidity for the particular antigen, as a result of the processes of heavy-chain class switching and affinity maturation. The normal human immune system is capable of producing a limitless number of highly specific antibodies with the ability to recognize any and all nonself elements with which it comes into contact. Failure of effective immune response control can result in the production of antibodies against self-antigens, termed **'autoantibodies'**, and these are characteristic of a number of **autoimmune diseases**, including, among others, **systemic lupus erythematosus (SLE)** and **rheumatoid arthritis (RA)**.

The terms antibody, gamma globulin and immunoglobulin are synonymous

Five classes of immunoglobulin are recognized – IgG, IgA, IgM, IgD, and IgE – with subclasses being recognized for IgG (IgG1, 2, 3, and 4) and for IgA (1 and 2). When studied at the individual molecular level, no other proteins show such amino acid sequence variation between individual members of the same class or subclass. This is most evident in the N-terminal domains of both heavy and light chains which are responsible for the antigen recognition portion of the molecule.

Antibodies are capable of discriminating between the molecules that characterize the outer capsular coverings of differing bacterial species which may vary by a single amino acid or a monosaccharide residue. This is a consequence of the dimensions of the area recognized by the antibody molecule being 10–20 Å (10^{-10}m), and thus being capable of being influenced by the alteration in three-dimensional conformation brought about by the change of a single residue.

Antibodies are good examples of how function is intimately related to structure

Antibodies (immunoglobulins) are Y-shaped molecules (see Fig. 4.5). The ends of the arms interact specifically with the pathogen (antigen) and the stem provides additional or effector functionality. This secondary or effector function endows the antibody with an ability to initiate immune responses that help eliminate the pathogen to which it is directed. An example of this is the activation of complement. The effector functions of antibodies are summarized in Table 38.4.

Table 38.4

The effector functions of antibodies

Type	Functions
IgG	Neutralization Opsonization for neutrophils and macrophages Passive immunity for fetus via transplacental passage Complement activation via classic pathway Antibody-dependent, cell-mediated cytotoxicity Natural killer function: cell killing of antibody-bound cells achieved by the receptors for the Fc portion Major isotype used in a secondary antibody response
IgA	Defense of mucosal surfaces; the most predominant immunoglobulin produced by MALT Neutralization
IgM	Neutralization Most effective classic complement pathway activator Predominant isotype in primary antibody responses
IgD	Possible role in signal transduction and B cell maturation Significance of circulating IgD is undefined
IgE	Major role is defense of mucosal surfaces against multicellular microorganisms

Activation of the complement system is one of the most important antibody functions

Activation of the complement system (see Fig. 38.1) is one of the most important antibody effector functions in the adaptive immune response. This is achieved by using a set of components termed the **‘classic activation pathway’,**

which comprise C1q, C1r, C1s, C4, and C-2. Sequential activation of these components leads to the activation of the pivotal and critically important C-3 component, which is an absolute requirement for full complement activation. Once this is achieved, the terminal membrane attack complex, which comprises the components C5, C6, C7, C8, and C9, is activated. This complex eventually generates the polymeric ring structure that inserts into the cell membrane of bacteria and is responsible for cell lysis. This classic pathway is triggered by C1q binding an IgG or IgM that is already bound to its specific antigen.

Two other pathways of activation exist, both constituting parts of the nonspecific immune response; they are probably older in evolutionary terms and have been described above.

Figure 38.8 summarizes the key elements of the adaptive immune response, showing the interrelationships between cellular and humoral components.

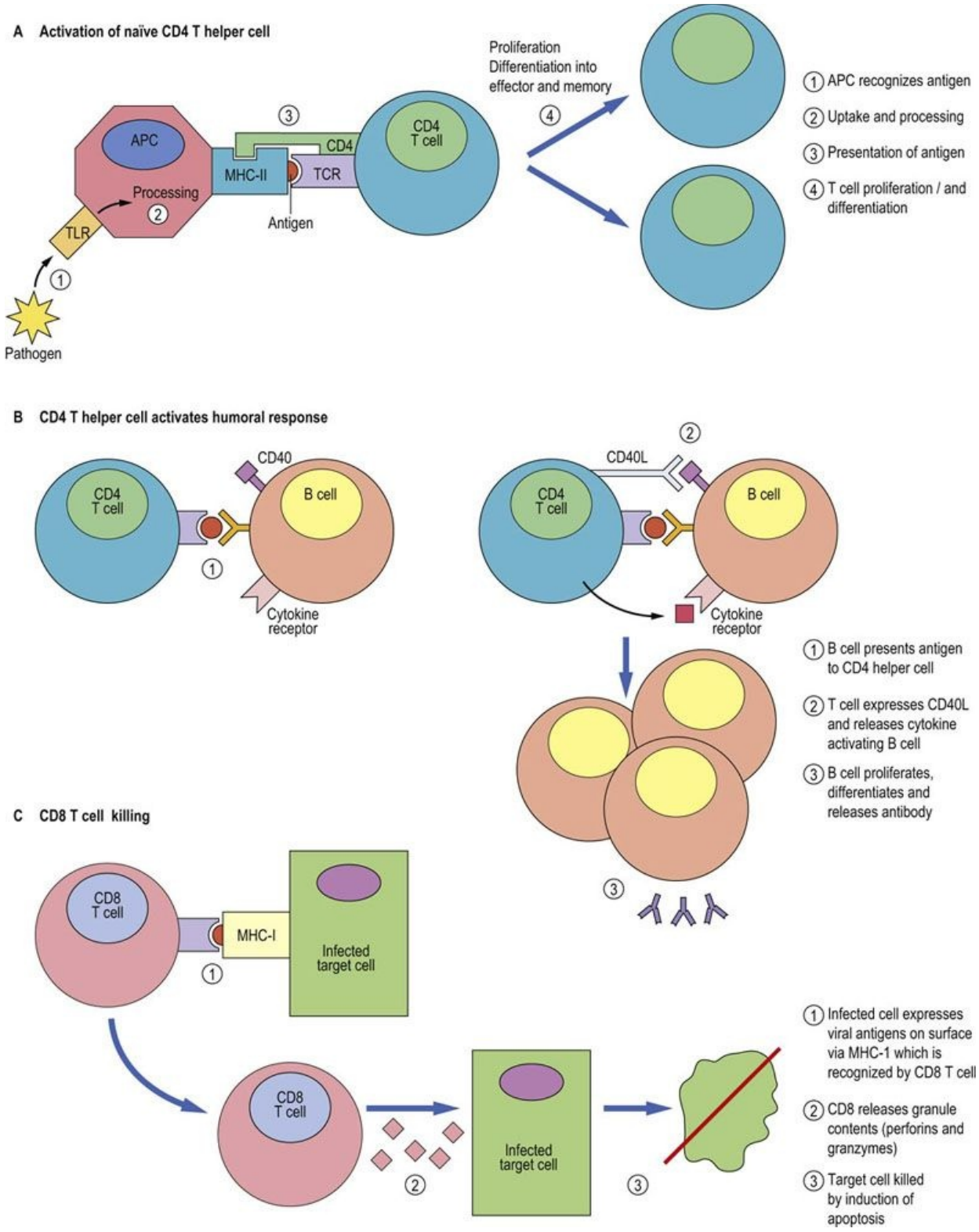


FIG. 38.8 Summary of the adaptive immune response.
Interrelationships between cellular and humoral components of the specific immune response. APCs activate naïve CD4 T cells, which in turn can activate B cells. Cytotoxic

T cells kill infected target cells.

Clinical box A 2-year-old with recurrent infections

immunodeficiency state

A 2-year-old child presented with a history of recurrent *Candida albicans* and chest infections. Investigations revealed a decreased number of neutrophils, IgG and IgA. Assessments of lymphocyte proliferative response showed decreased expression of CD40L (CD154) on T cells. A diagnosis of **X-linked hyper IgM** syndrome was made and treatment with intravenous immunoglobulin was commenced.

Comment.

Immunodeficiency states are classified as either primary or secondary. **Primary immunodeficiencies** are inherited conditions and with over 100 different conditions described can affect all parts of the immune system, both innate and adaptive. **Secondary immunodeficiencies** are often the consequence of infection (e.g. HIV) or other underlying conditions or environmental factors (e.g. malnutrition).

T cell help is required for effective B cell responses. Particular interactions are required for the switch of isotype from the IgM response that is typical of a primary antibody response to the more mature IgG and/or IgA isotypes seen during secondary antibody responses produced to subsequent antigenic challenges. CD40L on the T cell is required to interact with the CD40 on B cells, giving ‘help’ to achieve this. In its absence, antibody production is limited to IgM and the affected individual is immunocompromised, owing to the lack of the other important isotypes so critical to the integrity of the immune response. The problem of infection more typically associated with cellular problems suggests that the T cell defect has functional consequences for this arm of the immune response.

Vaccination

Vaccination has been probably the single most beneficial application developed to harness the immune response

The process of vaccination illustrates well the interactions of the humoral and cellular arms of the adaptive immune response, and the features that characterize it best – **specificity and memory**. On first encounter with antigen, the immune system and antigen interact to select lymphocytes with the receptors specific for that antigen. These undergo activation, proliferation and differentiation into effector memory cells, a process that may take up to 14 days to complete (Fig. 38.8). However, the process of memory cell generation now leaves a population of cells semi-primed for that specific antigen. On subsequent exposure, the response is more rapid in view of the partly activated state of the memory cells. It is also more effective as a consequence of a degree of maturation of the response, due to the differentiation of the lymphocytes that has already taken place.

With reference to antibody responses, the primary challenge elicits a predominantly IgM response. On subsequent challenge, the B lymphocytes undergo further maturation, differentiation and isotype switching through 'help' from appropriate T cells, and more rapidly produce a predominantly IgG response. This provides additional effector functions to that obtained with just IgM. It is this heightened and more specific response that can reduce both the severity and the duration of any damage caused by the offending antigen.

Autoimmunity is normally prevented by thymic education; a breakdown in the processes involved may lead to autoimmune disease

While the immune system's activities are mostly beneficial, there are several situations in which they can have deleterious effects. These are best considered as aberrations of the quality, quantity or direction of the response.

One particular aspect of these disorders, that of **autoimmunity** (self-reactivity), is avoided by the processes of central tolerance (during thymic education), and peripheral tolerance which induces clonal deletion and anergy. The self-reactive clones are eliminated or rendered impotent either through

deletion within the thymus, or by being controlled by T regulatory cells in the periphery. These mechanisms can be seen as a multilayered fail-safe strategy. Should these processes break down or be circumvented, the resulting state of self-reactivity and the inflammatory damage constitutes autoimmune disease.

The form of autoimmune disease is determined by the target antigen and the form of the immune response. At its simplest, reactions against ubiquitous antigens lead to what are termed nonorgan-specific autoimmune diseases. On the other hand, reactions to unique components of individual tissues, organs or systems lead to organ-specific disease. The former are best exemplified by **systemic lupus erythematosus (SLE)**, in which the apparent target antigens are components common to all nuclei. Damage is seen in several tissues, including the skin, joints, kidneys, and nervous system. These diseases, and the others mentioned in Table 38.5, are the focus of the disciplines of clinical immunology and immunopathology. More information can be found in the books cited in the Further Reading below.

Table 38.5

The consequences of failure of the immune system

Autoimmunity	Inappropriate response to self-antigens through the breakdown of self-tolerance can lead to autoimmune conditions **Examples:** rheumatoid arthritis, SLE, type 1 diabetes mellitus
Hypersensitivity	Inappropriate or overreaction to pathogen or antigen can often result in a response which does more harm to the body than the actual cause **Examples:** hat fever in response to pollen, anaphylactic response to foodstuffs, e.g. peanuts
Immunodeficiency	Ineffective immune response to infection can cause immunodeficiency. This can often be hereditary or induced by infection or drug treatment
	Primary immunodeficiency is an intrinsic defect of one or more components of the immune response, e.g. impaired production of antibody. **Examples:** X-linked agammaglobulinemia, which presents as very low B cell numbers and serum immunoglobulin
	Severe combined immunodeficiency (SCID), occurs when thymus fails to develop and no T cells are present
	Secondary immunodeficiency can develop after infection, e.g. AIDS, where the virus infects CD4 T cells or as response to certain drugs, e.g. steroids, which can impair immune cell function

Summary

■ Integrated immune response to nonself or altered-self elements (antigens) is made up of a number of components. Some of these show unique specificity for the particular stimulating antigen(s) and comprise the specific or adaptive immune response, whilst others recognize pathogen signatures and comprise the nonspecific or innate immune response.

■ The innate response represents the first-line response and is present in all eukaryotes. The cells and soluble mediators involved are primarily those associated with the processes of inflammation and vascular activation.

■ The adaptive response is more refined and usually invoked only in the face of either failure or continued stimulation of the innate response. The cells responsible for the adaptive immune response are the T and B lymphocytes. The specificity they show for the inciting antigen is achieved via the use of specific antigen receptors, expressed on their cell surface and clonal expansion.

■ T cells recognize processed antigen via their antigen receptors, interacting with antigen presented by MHC-bearing cells. This leads to the secretion of additional cytokines and the generation of effector functions such as T cell help and T cell-mediated cytotoxicity, brought about by the T helper and T cytotoxic subsets, respectively. Historically, T cell responses have been termed the cellular immune response. A distinct $CD4^+$ subset of T cells is termed ‘T regulatory cells’, as they function to control adaptive responses and in part prevent autoreactivity by the immune response.

■ B cells recognize native antigen and secrete proteins, termed antibodies, which can bind directly to the antigen. Historically, B cells and their antibody products have been termed the humoral immune response.

■ Both T and B cells and their products are able to recruit and utilize components of the innate response in a more effective and targeted manner, with the aim of eliminating or eradicating the antigen.

■ In addition to demonstrating specificity, the adaptive immune response also demonstrates another critically important characteristic not seen with the innate response: the memory for its encounter with antigen. The benefit of this is that, on subsequent contact with the same antigen, a heightened and more efficient response will lead to a quicker removal of the causative agent, hopefully with less tissue damage than on first encounter.

Active Learning

1. What are the key features of the innate and adaptive immune responses?
2. Why has the innate immune system evolved multiple receptor systems for pathogen recognition?
3. What is the role of the different cytokine families during an immune response?
4. What is the role of the thymus in the immune response?
5. Compare the function of T and B lymphocytes.

Further reading

Abbas, AK, Lichtman, AH. *cellular and molecular immunology*, ed 7. Philadelphia, PA: Saunders Elsevier; 2012.

Chapel, H, Heaney, M, Misbah, S, *et al. Essentials of clinical immunology*, ed 5. Oxford: Blackwell; 2006.

Kumar, H, Kawai, T, Akira, S. Pathogen recognition by the innate immune system. *Int Rev Immunol*. 2011; 30:16–34.

Liu, Z, Davidson, A. Taming lupus – a new understanding of pathogenesis is leading to clinical advances. *Nature Med*. 2012; 18:871–882.

MacPherson, G, Austyn, J. *Exploring immunology*. Hoboken, NJ: Wiley-Blackwell; 2012.

Murphy, K. *Janeway's immunobiology*, ed 8. London: Garland; 2011.

Shevach, EM. Biological functions of regulatory T cells. *Adv Immunol*. 2011; 112:137–176.

CHAPTER 39

Biochemical Endocrinology

Robert K. Semple

Learning objectives

After reading this chapter you should understand:

- The principles governing endocrine signaling and its regulation.
- The anatomy, organization and regulatory role of the hypothalamus and pituitary.
- The processes controlling biosynthesis, transport and action of thyroid hormones.
- The mechanisms regulating synthesis and action of glucocorticoids.
- The regulation of the synthesis and activity of sex steroid hormones and their role in human reproduction.
- The direct and indirect actions of growth hormone, including the role of IGF-I.
- The role of prolactin in reproduction.
- The consequences of deficiencies and excesses of hormones regulated by the hypothalamo-pituitary axis.

Introduction

As a first approximation, the human body contains 10 trillion – or 10 million million – cells, which may lie more than 2 meters apart. Coordination and regulation of the growth and multiple functions of these cells in order to meet the demands of existence in constantly fluctuating environmental conditions require precise and sophisticated integrative systems. **The nervous system** is a key part of the solution, being particularly well suited to mediating reflexes and motor actions needed, for example, in 'fight or flight' situations, which may require responses in fractions of a second. **The endocrine system**, however, is the principal means of regulation of a wide range of less acute functions including growth, development, reproduction, many aspects of homeostasis, and the response to more chronic external stimuli and stress. Such endocrine responses occur over several seconds at their fastest, to days or weeks at their slowest. Failures in these highly elaborate, complex and often interconnected endocrine control systems are common, and lead to many highly prevalent diseases.

Hormones

Hormones are chemical substances that are produced by particular glands or groups of cells, and that elicit specific responses from distant cells or organs

The concept that organs and tissues may produce ‘internal secretions’ that influence the behavior of distant parts of the body probably first emerged in France in the 18th century, and by the beginning of the 20th century, adrenaline and vasopressin (antidiuretic hormone, ADH) had been isolated. However, it was only in 1905, shortly after isolating secretin, that Ernest Starling in London first coined the word ‘hormone’ to describe such internal messengers. Hormones are now understood to be chemical substances that are produced by particular glands or groups of cells, and that elicit specific responses from distant cells or organs. Classically, such hormones are blood-borne and are described as **endocrine hormones**. However, subsequently it has become clear that some hormones act locally on cells around their cell of origin – so-called **paracrine hormones** – or even on the same cell that has produced them – so-called **autocrine hormones**.

Types of hormones

Hormones have diverse chemical structures

Many different types of molecules function as hormones (Table 39.1). Among the most chemically simple of hormones are **modified amino acids**, such as epinephrine (adrenaline), and many such simple amines also function as neurotransmitters (Chapter 41.1). At the other end of the spectrum, some hormones are **polypeptides** varying in size from tripeptides (e.g. thyrotropin-releasing hormones) to complex glycoproteins (e.g. luteinizing hormone, LH). The smaller peptide hormones are synthesized as large polypeptide prohormones, that are cleaved by proteolytic enzymes to release active hormone from the endocrine gland. Sometimes one prohormone can be cleaved in different ways to produce many different hormones, according to the tissue and cellular context. Other hormones are derived by **modification of simple lipids** such as **cholesterol** or **fatty acids**.

Table 39.1

Chemical derivation of hormones

Hormones derived from amino acids	
Amino acid derivatives	Catecholamines, serotonin, thyroxine
Tripeptides	TRH
Small peptides	AVP (ADH), somatostatin
Intermediate-size peptides	Insulin, parathyroid hormone
Complex polypeptides and glycoproteins	Gonadotropins, TSH
Hormones derived from lipid precursors	
Cholesterol derivatives	Cortisol, testosterone, estradiol, vitamin D
Fatty acid derivatives	Prostaglandins, leukotrienes
Phospholipid derivative	Platelet-activating factor
Hormones derived from other chemicals	
Purines	Adenosine
Gases	Nitric oxide

AVP/ADH, arginine vasopressin or antidiuretic hormone; TRH, thyrotropin-releasing hormone; TSH, thyroid-stimulating hormone.

Many glands and tissues produce hormones

Hormones are classically known to be secreted from **ductless glands** such as the thyroid, adrenals, and pituitary which have been known to anatomists for centuries. However, many hormones are so potent that they need only circulate at tiny (e.g. picomolar) plasma concentrations to exert meaningful biological effects, and may derive from small clusters of cells, or from one particular cell

type scattered within a larger organ. Furthermore, it is now becoming clear that many tissues which were not traditionally thought to be active in endocrine terms actually secrete a wide variety of different hormones with diverse effects on other tissues. Indeed, the availability of the human genome sequence, allied to sophisticated bio-informatics and techniques such as microarray-based transcriptional profiling (Chapter 36), has led to the discovery of many new suspected hormones and hormone receptors, and understanding of the endocrine mechanisms regulating human physiology is likely to increase rapidly. Nevertheless, in this chapter the emphasis is on the very well-established hormonal systems which are the focus of current clinical endocrinology.

Advanced concept box Nonclassic endocrine organs

Many tissues apart from classic ductless endocrine glands are now known to be active endocrinologically, and the produced signals mediate important metabolic cross-talk between different tissues. Examples of this include production of leptin from **white adipose tissue**, which signals body energy stores to the brain (Chapter 22), production of fibroblast growth factor 23 (FGF23) from **bone**, (which modulates renal phosphate handling and is ectopically produced by some tumors, resulting in a clinically important renal phosphate leak and osteomalacia), and production of satiety factors such as ghrelin from the **small intestine**. The **placenta** is also a highly active endocrine tissue: as well as producing βHCG, progesterone, placental growth hormone and human placental lactogen, which have well-understood roles in pregnancy; it elaborates many other hormones at high levels whose role remains to be determined.

Principles of hormone action

In order to function effectively to relay signals from one part of the body to others, endocrine systems must exhibit certain general properties (Fig. 39.1).

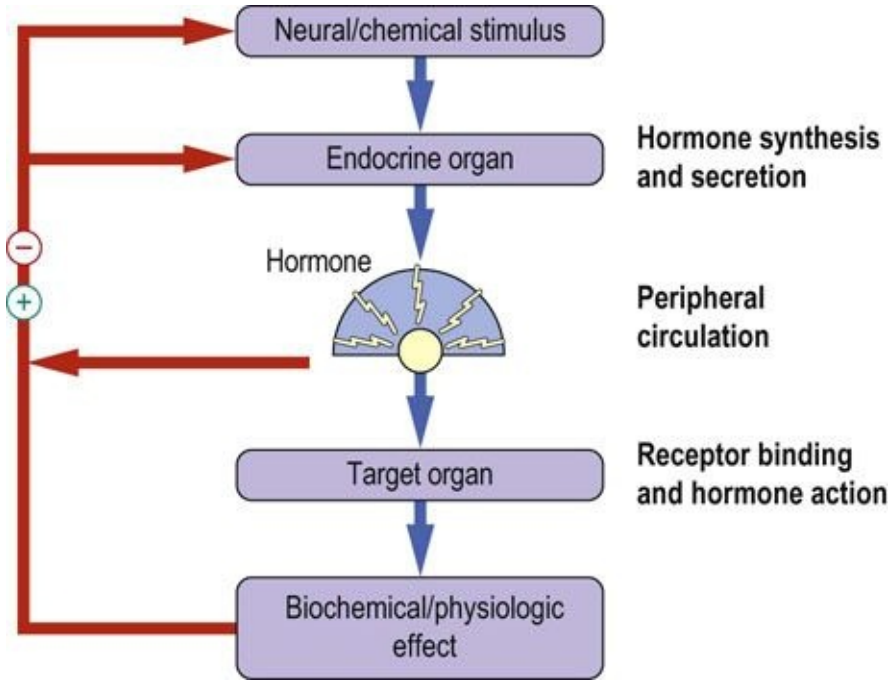


FIG. 39.1 Basic endocrine processes.
The feedback regulation of hormone action is a classic example of self-regulation. Negative feedback regulation is common. Positive feedback (the feed-forward loop) is rare. Note how the feedback loops operate at different levels of the endocrine system.

Coupling of hormone release to relevant stimuli

Some hormones function to transmit to the body important environmental cues, while others are more concerned with calibrating metabolic processes involved in maintaining homeostasis

In either situation, it is critical that there is effective translation of the magnitude of the stimulus into the amount of hormone released. How this is achieved varies

depending on the stimulus. Simple homeostatic sensing mechanisms include direct coupling of parathyroid hormone release to calcium levels via a cell surface receptor which binds calcium (Chapter 26), and direct coupling of insulin secretion to plasma glucose levels via sensing of blood glucose concentration by the pancreatic β cell (Chapter 21). Many of the endocrine systems described in this chapter, in contrast, are controlled by **modified neurons** which secrete hormones into the bloodstream. The level of secretion of such neurons is often determined by complex synaptic inputs from many different parts of the brain, as well as by the sensing of circulating metabolites and hormones. These secretory neurons thus function as important sites of integration of a variety of environmental, psychologic and physiologic cues.

Feedback regulation: a major feature of endocrine systems

A major feature of most endocrine systems is **negative feedback** (Fig. 39.1). This means that a response induced by action of a hormone feeds back to inhibit the level of hormone production. The effect of this is to damp down fluctuations in the process controlled by the hormone, thereby enhancing stability of that process and hence homeostasis. However, were negative feedback to be the only process at play, then the output of the system would remain constant. In fact, many endocrine systems, notably those controlled by the hypothalamus, display distinct **rhythmicity**. This is determined by the net product of the neural inputs to the relevant secretory neurons, and in this situation negative feedback serves instead to smooth out the hormonal profile and to prevent short-term instability. Furthermore, modulation of the susceptibility of the controlling neuron to feedback inhibition represents one potential mechanism of modulating the level of the output of the endocrine system.

Positive feedback refers to stimulation of hormone release by the response it provokes. This feed-forward loop is inherently unstable, and leads to a rapid, exponential increase in the level of signal. This is **much rarer** than negative feedback in physiology, but plays an important role in some processes which require precise timing, such as the luteinizing hormone surge which drives ovulation.

Transduction of signal at target tissues:

hormone receptors

Hormones act by binding to specific receptors, either on the cell surface or within the target cell (Chapter 40)

Classically, there is a very high degree of specificity in this binding, and it is this hormone–receptor interaction that triggers and coordinates a wide range of biological effects. Hormone receptors may be divided into different families according to their structure, and each family broadly shares signal transduction mechanisms.

Many amine-like hormones and peptides act via receptors that are intrinsic plasma membrane proteins

Many amine-like hormones and peptides which cannot cross lipid bilayers act via G-protein coupled receptors, while some larger peptides act via tyrosine kinase receptors (e.g. insulin and IGF-I) or cytokine-like receptors (e.g. leptin and growth hormone). All of these receptor types are intrinsic plasma membrane proteins, and rely on cascades of sequential phosphorylation events to alter enzyme activity and gene expression.

Lipophilic hormones enter cells before binding their receptors

In contrast lipophilic hormones including steroid hormones, some hormones derived from fatty acids, and tri-iodothyronine, which has physicochemical properties in common with steroids (Chapter 17), enter cells before binding to their receptors. These so-called **nuclear hormone receptors** are effectively ligand-activated transcription factors, and provide a very direct link between the exposure of a cell to hormonal ligand and alteration of patterns of gene expression (Chapter 35).

Many **putative receptors for unknown hormones** have been identified by virtue of sequence homology among members of each of the major hormone receptor classes. It is likely that identification of the ligand for these ‘orphan’ receptors will significantly increase the range of endocrine hormones. Such putative receptor-encoding genes are also a major focus of pharmaceutical research, due to the relative ease with which receptors may be targeted by novel drugs.

Turning off the hormonal signal

Levels of hormones are useful physiologic signals only if there is also an effective way of turning the signals off

Hormone inactivation usually occurs by further metabolism (e.g. proteolysis of peptides, or hydroxylation, conjugation and excretion of many steroids). Such degradation may occur in plasma, in organs such as the liver, or in target tissues after receptor-mediated internalization of the hormone. The rate of clearance of different hormones varies enormously, from a few minutes (insulin), through hours (steroids) to days (thyroxine). **Measurement of urinary concentrations of hormones** or their metabolites can be used in some diagnostic settings, including the diagnosis of pregnancy, and in pathologies such as secretory adrenal tumors.

Advanced concept box

Preservation of epigenetic ‘marks’ (imprinting) in endocrinology

Gene expression may be influenced not only by the DNA sequence of genes and their promoter regions, but also by covalent modifications of DNA and/or histones. **Methylation of cytosine residues** occurs particularly in promoter regions and acts to **silence genes**, for example during tissue differentiation. During the earliest stages of development most of these epigenetic ‘marks’ are wiped clean to generate totipotent embryonic stem cells. However, they are preserved at some genes, and this pattern of covalent modifications often shows which parent the gene was inherited from. Because epigenetic marks are preserved as cells divide, major ‘parent of origin’ effects can be seen subsequently on gene expression in some or all tissues.

Several endocrine systems show significant imprinting. Best known is the IGF2-IGF2 receptor system, which acts to influence placental and fetal growth, in common with nearly all known imprinted genes. Another important endocrine signaling gene which is imprinted is the GNAS gene, encoding a stimulatory G-protein subunit coupled to many classic endocrine peptide

hormone receptors. Various different mutations have been described, and depending on which parent they are inherited from, they may produce skeletal dysplasia and resistance to various hormones, including most notably parathyroid hormone. Indeed, so-called ‘pseudohypoparathyroidism’ was the first reported example of end-organ resistance to a hormone. For review of the fascinating and complex biology of GNAS, readers are referred to Weinstein LS et al. Minireview: GNAS: normal and abnormal functions. Endocrinology 2004;145:5459–5464.

Carrier proteins and ‘free’ hormones

Many small or hydrophobic hormones are transported in plasma bound to carrier proteins

Within the circulation, many small or hydrophobic hormones are transported bound to carrier proteins. For example, thyroxine and cortisol are transported on plasma proteins such as thyroid-binding globulin (TBG) and cortisol-binding globulins (CBG). This has several consequences of relevance to both physiology and clinical measurement. The transport proteins extend the biological half-life and increase the plasma concentration of the smaller hormones, which would otherwise be eliminated rapidly in the liver or kidney. This also means that there may be large differences between the total plasma levels of hormone and the unbound or ‘free’ hormone in solution, which is generally the biologically active form. Thus clinical interpretation of plasma hormone levels is simplest if assays are developed to detect free hormone only, otherwise assumptions must be made about levels of binding proteins, which in turn may be significantly influenced by the nutritional and hormonal milieu. A common example of the problem posed by variable levels of carrier proteins comes in assessing cortisol levels in women taking estrogen-containing contraceptive pills, which stimulate binding protein levels and thus make the usual laboratory normal ranges for cortisol inapplicable. It is not only small lipophilic hormones which have plasma-binding proteins, however; both growth hormone and insulin-like growth factor 1 have plasma binding proteins which influence bio-availability and tissue activity of the hormones in a variety of different ways (Table 39.2).

Table 39.2

Major hormone binding proteins

Hormone	Binding protein	Approximate % bound	Notes
Thyroxine	Thyroid-binding globulin (TBG) Albumin Transthyretin	75% 10–15% 10–15%	Clinically significant increase in early pregnancy
Testosterone	Sex hormone binding globulin (SHBG)	60–70%	Levels of SHBG are strongly hormonally regulated (e.g. suppressed by insulin and increased by thyroid hormone).
Cortisol	Cortisol-binding globulin (CBG)	75%	Variations in cortisol binding proteins may be particularly misleading as it is total cortisol which is generally assayed. Thus women taking estrogens (which increase CBG) may have high total but normal free cortisol, while in critical illness and malnutrition the converse is true.
	Albumin	15–20%	
GH	Growth hormone binding protein (GHBP)	50%	GHBP is a soluble fragment of the GH receptor. This is a common feature of many cytokine-like hormones.
IGF-1	Insulin growth factor binding protein 3 (IGFBP-3)	75%	There are 6 IGFBPs which modulate the paracrine activity of IGF-I and may have signaling roles in their own right.

Local metabolism of hormones at target tissues

Another feature of some endocrine systems, particularly those involving steroid hormones, is local metabolism of the hormones or prohormones near their cognate receptor

This may involve conversion of an active circulating hormone to a more potent form (for example, testosterone to dihydrotestosterone by 5α-reductase in androgen-dependent hair follicles and prostate gland), or creation of an active hormone from what is effectively a circulating prohormone (e.g. thyroxine to tri-iodothyronine by deiodinase, or synthesis of estradiol from adrenal androgens in adipose tissue). A similar mechanism is also employed to prevent mineralocorticoid receptors in the kidney being exposed to high concentrations of cortisol, with cortisol being inactivated by 11β-hydroxysteroid dehydrogenase type II, which is expressed highly near the mineralocorticoid receptors. Many other examples of such intracellular interconversion of steroid hormones have also emerged, and study of this phenomenon has sometimes been dubbed ‘intracrinology’ (Table 39.3).

Table 39.3

Examples of hormone metabolism in target tissues

Hormone	Enzyme	Substrate	Product effect on activity
Thyroxine (T4)	Deiodinase type 2	Tri-iodothyronine (T3)	Generates potent T3 from nearly inactive T4 in target tissues.
Testosterone	Aromatase	Estrogen	As well as playing an important role in the ovary in generating estrogen, aromatase is expressed in other tissues including bone, adipose tissue and hypothalamus. Even in men, this local conversion of androgens to estrogens is important in mediating effects of testosterone on bone density, and central negative feedback to LH/FSH production.
Testosterone	5α-reductase	Dihydrotestosterone	Produces more potent ligand for the androgen receptor, amplifying testosterone activity locally: genetic deficiency produces pseudohermaphroditism.
Cortisol	11β-hydroxysteroid dehydrogenase type II	Cortisone	Converts cortisol, which activates the mineralocorticoid receptor, to cortisone, which does not. The enzyme is highly expressed in mineralocorticoid-sensitive tissues such as the kidney. Genetic deficiency leads to exposure of mineralocorticoid receptors to high concentrations of active ligand, leading to severe hypertension and hypokalemia in the syndrome of apparent mineralocorticoid excess.

Several hormones may control one process, or one hormone may control several processes

While it may be convenient to think of the endocrine system as being compartmentalized, so that one hormone has control over one process, this is rarely the case

For example, at least four different hormones are involved in the regulation of plasma glucose concentration (Chapter 21). Conversely, single hormones such as testosterone influence a range of metabolic processes.

All of these features of endocrine systems will be illustrated later in this chapter by some or all of the endocrine axes which pass through the pituitary, to be discussed.

Biochemical assessment of hormone action

Laboratory testing of endocrine systems aims first to determine whether the system is functioning abnormally, and second to localize the functional defect

Most commonly, this means that the level of the hormone that elicits the target tissue response is determined, usually by immunoassay (Fig. 39.2), together with measurement of one or more upstream trophic hormones, where they exist. The presence of negative feedback regulation means that the system will attempt to correct perturbations in levels of the effector hormone with compensatory changes in levels of trophic hormones. Thus assessment of at least two points in such an endocrine feedback loop is essential, and permits focusing of later diagnostic imaging on the relevant gland.

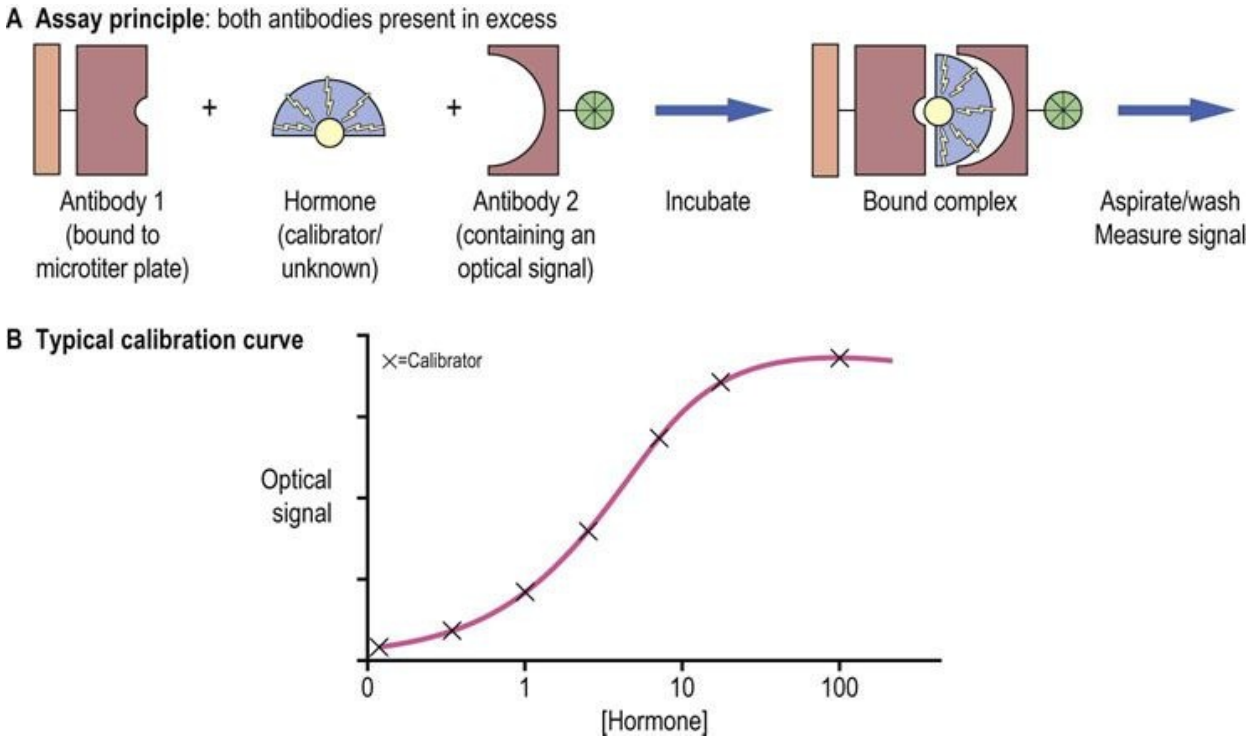


FIG. 39.2 The sandwich design for enzyme-linked immunosorbent assay (ELISA) for

the measurement of hormone concentration.
(A) The amount of signal present on the microplate after washing off the serum and reagents is a measure of the concentration of a hormone. **(B)** Calibration curve is a series of measurements of samples with a known concentration of a hormone (calibrators). Subseqently the signal obtained from the unknown samples is compared to the standard curve, and the hormone concentration is read out.

Once the hormones of most interest to the clinical problem have been chosen, care needs to be used to ensure appropriate sampling. This should take into account whether the hormone is very unstable (for example, many small peptide hormones require blood samples to be taken into tubes containing a protease inhibitor which are kept on ice), and also whether it is produced with a circadian or other rhythm. For some hormones, such as thyroxine, which have very long half-life, timing of the sample is not critical, while for others, such as cortisol, standardized timing of the sample is essential.

Single measurements are meaningless in case of hormones with pulsatile secretion

A further factor to take into account in hormone testing is whether or not secretion of the hormone is pulsatile. Nearly all of the hypothalamic and pituitary hormones show some degree of pulsatility in secretion. In most cases this simply means that repeat testing in the case of mild abnormality is required, while for some hormones, such as growth hormone, which has both a short half-life and very striking secretory spikes, single measurements are often meaningless.

Because of these problems, endocrinologists commonly employ multiple testing to build up a profile of hormone levels at several points in the day. This can be an effective way of detecting subtle perturbations of the circadian rhythm, and can also lessen the impact of oscillatory secretion of the hormone.

Some hormone are measured after a relevant stimulus has been applied

A further approach is to use provocative testing. In other words, hormone levels are measured not just in the resting state, but also after a relevant stimulus has been applied, in an attempt to assess the maximal capacity of the endocrine gland or system being tested. Frequently the stimulus is a **high dose of trophic hormone** (e.g. ACTH, TRH, GnRH, etc.); however, it may instead be a **metabolic challenge** (e.g. with an oral glucose load or severe insulin-induced

hypoglycemia to mimic stress). Examples of each of these tests are given in Table 39.4.

Table 39.4

Examples of commonly used provocative endocrine tests

Endocrine axis	Stimulus	Measurement	Rationale/use
H-P-Adrenal	Synthetic ACTH	ACTH	Tests functional integrity of adrenal glands, which rely in turn on chronic tropic actions of ACTH. This is also an indirect test of pituitary/hypothalamus.
H-P-Adrenal	Insulin-induced hypoglycemia	Cortisol	Severe hypoglycemia mimics physiologic stress and robustly tests the hypothalamus.
H-P-Thyroid	TRH	TSH	The pattern of TSH release after TRH stimulation gives useful information in diagnosing central hypothyroidism.
H-P-Growth	Insulin-induced hypoglycemia	GH	Baseline pulsatility of GH is overcome by applying a strong stimulus to its release.
H-P-Growth	Oral glucose load	GH	Failure of suppression of GH by glucose is used in the diagnosis of acromegaly.

H-P, hypothalomo-pituitary.

H-P, hypothalomo-pituitary.

Clinical test box Hormone immunoassay

Immunoassay is the most widely applied technique for measuring hormone levels. Antibodies are produced that bind to antigenic sites on the hormone. Ideally, such antibody should possess both high specificity and affinity for the hormone of interest, *e.g.* the β-subunit of glycoprotein hormones.

In the first immunoassays, the antibodies were obtained from the serum of animal species (e.g. rabbit or sheep) that had been immunized with human hormone preparations. Such antisera contained a range of different antibodies (polyclonal) capable of binding to different sites on the hormone antigen. However, antibody specificity and titer varied with the animal and duration of the immune response.

Today most immunoassays employ **monoclonal antibodies**, which are produced by fusion of spleen cells from an immunized mouse with a mouse myeloma cell line. Hybridoma cells may be cloned to produce a cell line that secretes a single antibody species for an indefinite period of time.

Many commercial producers have designed proprietary methods

for quantifying the hormone–antibody interaction. One of the most widely used formats is the two-site (sandwich) immunometric or **enzyme-linked immunosorbent assay (ELISA)** assay outlined in Figure 39.2. This assay employs two antibodies binding to different epitopes on the hormone, one of which is labeled or modified to be capable of generating an optical signal. Spectrophotometric, fluorometric, luminescence and radiochemical reporter systems are in common use.

Not all immunoassays rely on labeled antibody: other approaches include **nephelometry**, in which appearance of light-scattering immune complexes is monitored spectrophotometrically.

Major types of endocrine pathology

Autoimmunity to the thyroid, adrenal and pancreatic islets accounts for more than 50% of all organ-specific autoimmune disease

Although each endocrine system may malfunction as a consequence of organ-specific damage or disease, for example relating to the anatomic site of the gland, there are also some general types of pathology to which endocrine cells are very susceptible. The first of these is **autoimmune destruction with detectable endocrine-gland-specific antibodies**. This is a potential cause of loss of function of nearly all endocrine glands and, less commonly, may also cause gland hyperfunction. In fact, autoimmunity to the thyroid, adrenal and pancreatic islets accounts for more than 50% of all organ-specific autoimmune disease. This appears in part due to genetic predisposition, accounted for by both MHC and non-MHC genes, and also by poorly defined environmental factors (Chapter 38).

The second major group of pathologies which present as endocrine disease is neoplasia, which may be either benign or malignant

Neoplastic cells arise as autonomous clones within endocrine glands due to **somatic mutations in growth factor signaling pathways** or commonly in **G-protein coupled signaling pathways.** They may produce disease due to either excessive and dysregulated production of biologically active hormones or damage to neighboring, normal endocrine cells, with attendant loss of hormone secretion. Very small and benign adenomas which would remain undetected in nonendocrine tissues may produce florid disease by virtue of hormone hypersecretion coupled to the potency of hormones in eliciting biological responses.

As well as the pathologies arising in endocrine cells or organs, the very nature of hormones, and their frequent availability in medical preparations, means that they may also cause clinical problems when used exogenously, either as a side effect of a desired therapeutic action (e.g. corticosteroids used in inflammatory disease causing weight gain and diabetes), or due to inappropriate self

administration (e.g. thyroid hormone to induce weight loss).

The hypothalamo–pituitary regulatory system

Hormones of the posterior pituitary gland are distinct from those of the anterior pituitary

The pituitary gland is a pea-sized, oval organ encased in a bony cavity of the skull (the *sella turcica*) below the brain. It communicates with the hypothalamus via the pituitary stalk, which contains a complex array of axons and portal blood vessels (Fig. 39.3). The pituitary gland is divided into two lobes. The **posterior pituitary** (or neurohypophysis) is embryologically part of the **brain**, and consists largely of neurons which have cell bodies in the supraoptic and paraventricular nuclei of the hypothalamus. It is in these cell bodies that hormones are synthesized and packaged before transport on microtubules along axons to the pituitary, where they are released. The **anterior lobe** (adenohypophysis), accounting for approximately 80% of the gland, is embryologically derived from **oral ectoderm**, and has no direct anatomic continuity with the brain. Instead, it may be viewed as a target organ for endocrine hormones released from hypothalamic nuclei and transported from the median eminence by the portal circulation.

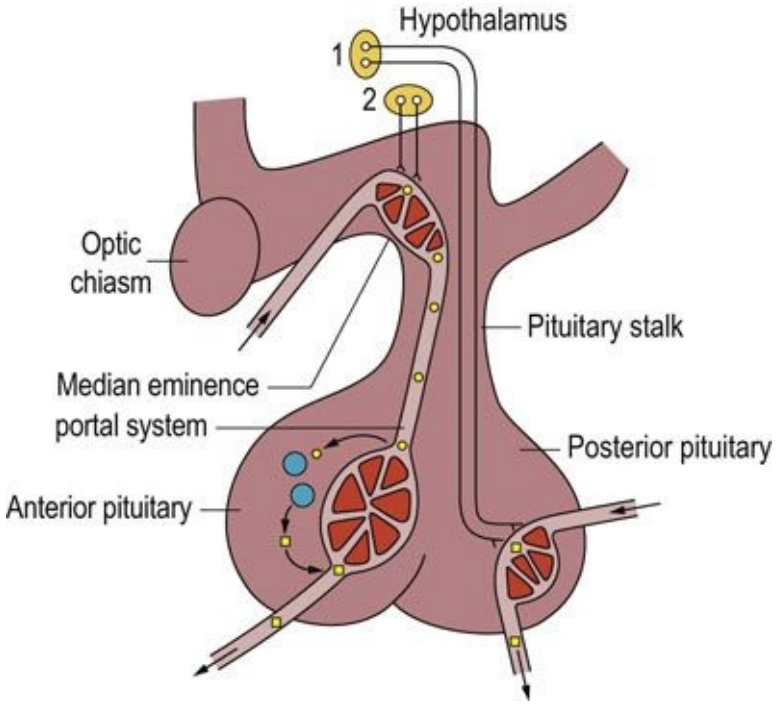


FIG. 39.3 The hypothalamo-anterior pituitary regulatory system.
Hormones of the posterior pituitary are synthesized and packaged in the supraoptic and paraventricular nuclei of the hypothalamus **(1)**, transported along axons and stored in the posterior pituitary prior to release into the circulation. The anterior pituitary releasing-or release-inhibiting hormones are synthesized in various hypothalamic nuclei **(2)** and transported to the median eminence. From there they travel to the anterior pituitary via a portal venous system.

Both posterior and anterior pituitary are controlled largely by the hypothalamus

The hypothalamus is a highly connected center of the CNS, receiving synaptic inputs from vast numbers of different parts of the brain as well as sensing peripheral signals through areas of porosity in the blood–brain barrier. The hypothalamus thus functions as an integrative center which orchestrates a huge number of endocrine and neural processes, and entrains them to relevant external stimuli. The endocrine systems which involve the hypothalamus, pituitary and downstream organs are usually termed **‘axes’** and are most usefully viewed as **functional units for the purposes of clinical diagnosis and management** (Fig. 39.4).

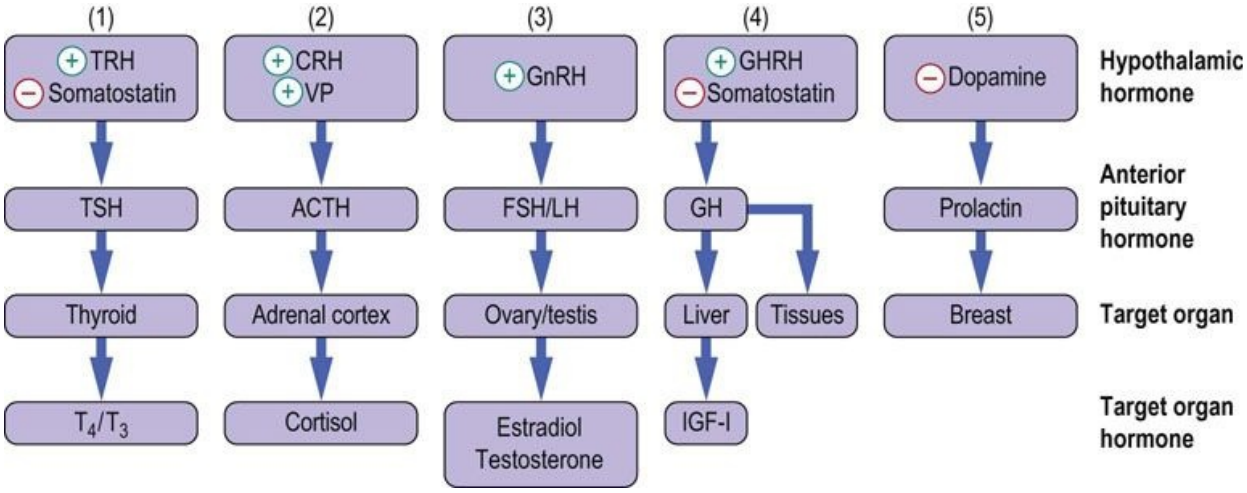


FIG. 39.4 Hypothalamo-anterior pituitary-regulatory target organ axes.
The hypothalamo-anterior pituitary regulatory system comprises 5 parallel endocrine axes, regulating the biosynthesis and release of: **(1)** thyroid hormone; **(2)** glucocorticoids; **(3)** sex steroids; **(4)** growth hormone; and **(5)** prolactin. T_4, thyroxine; T_3, tri-iodothyronine; GHRH, growth hormone-releasing hormone; GnRH, gonadotropin-releasing hormone; IGF-I, insulin growth factor-I ACTH adrenocorticotropic hormone; FSH, follicle-stimulating hormone; LH, luteinizing hormone; TRH,thyrotropin-releasing hormone, CRH, corticotropin-releasing hormone; TSH, thyroid-stimulating hormone; VP, vasopressin; (+) indicates stimulatory action and (−) inhibitory action.

Hormones of the posterior pituitary gland

The posterior pituitary gland secretes two hormones into the circulation

Oxytocin is a small peptide hormone that stimulates smooth muscle contraction in the uterus and breast; it functions in parturition (childbirth) and lactation. **Vasopressin** (VP), also known as **antidiuretic hormone** (ADH), is a 9-amino acid cyclic peptide with functions linked to the control of water clearance and homeostasis, and is described in detail in Chapter 24. In the human VP, the amino acid in position 8 in the peptide is arginine, and thus it is referred to as **arginine-vasopressin (AVP)**.

Hormonal axes of the hypothalamo–anterior pituitary regulatory system

There are five separate endocrine axes passing through

this system (Fig. 39.4)

Three of these form a complex **three-level endocrine system**, in which the hormones of the pituitary gland, thyroid-stimulating hormone (TSH), adrenocorticotropic hormone (ACTH), follicle-stimulating hormone (FSH) and luteinizing hormone (LH), may be regarded solely as trophic hormones for other target organs (i.e. thyroid, adrenals, gonads). In these axes, sophisticated control is exercised *via* a cascade in which each endocrine organ in the axis amplifies both the amount of hormone secreted and the biological half-life of its hormone product compared with the previous organ. **The fourth** endocrine axis is **a hybrid**: growth hormone (GH) is both a trophic hormone and has actions in its own right. **The fifth** and final endocrine axis mediates secretion of prolactin, which is unique in that it is **not a trophic hormone**. Distinct clinical syndromes are produced by either deficiency or excess of any of the anterior pituitary hormones except prolactin, deficiency of which does not lead to clinically significant problems in humans (Table 39.5).

Table 39.5

Clinical conditions associated with pituitary hormone disorders

Hormone	Deficiency	Excess
TSH	Hypothyroidism	Thyrotoxicosis
ACTH	Hypoadrenalism	Cushing's disease
FSH/LH	Hypogonadism	Precocious puberty
GH	Short stature	Gigantism/acromegaly
Prolactin	None	Galactorrhea/infertility

The hypothalamo–pituitary–thyroid axis

Thyrotropin-releasing hormone (TRH)

TRH is manufactured in the hypothalamus and transported via the portal circulation to the pituitary where it ultimately stimulates secretion of TSH

TRH is a modified tripeptide released in pulsatile fashion by peptidergic hypothalamic nuclei and transported to the anterior pituitary by the portal circulation. TRH stimulates TSH synthesis and secretion by binding to G-protein coupled receptors on the pituitary thyrotroph cell membrane that are linked to phospholipase C. The resulting increase in intracellular inositol trisphosphate (IP3) stimulates the release of calcium from intracellular storage sites and so leads to secretion of preformed TSH. More chronic actions of TRH include stimulation of TSH subunit biosynthesis and TSH glycosylation. The number of TRH receptors on the thyrotrophs is downregulated both by the concentration of TRH itself and by thyroid hormones.

Thyroid-stimulating hormone (TSH)

TSH, or thyrotropin, is a small glycoprotein synthesized by pituitary thyrotrophs

TSH consists of two noncovalently linked subunits and contains about 15% carbohydrate. The α-chain is identical to that found in other glycoprotein hormones (LH, FSH, βHCG) and so the specificity is conferred by the β-chain. The synthesis of each subunit is directed by separate messenger ribonucleic acids (mRNAs) encoded by separate genes on different chromosomes. The carbohydrate side chains are complex mixtures of unmodified, acetylated, and sulfated sugars. TSH is secreted in a pattern that is both pulsatile and circadian, and it has a plasma half-life of about 65 minutes.

TSH acts on the thyroid gland and influences virtually every aspect of thyroid hormone biosynthesis and

secretion

TSH, like TRH, acts via a specific G-protein coupled receptor. However, in this case the receptor is expressed on follicular cells in the thyroid gland, and it is coupled to adenylyl cyclase and thus protein kinase A (or cAMP-dependent protein kinase). cAMP-dependent protein kinases control virtually every aspect of thyroid hormone biosynthesis and secretion, including iodide transport, iodothyronine formation, thyroglobulin proteolysis, and thyroxine deiodination. TSH also stimulates growth of the thyroid gland. Indeed, activating somatic mutations in the TSH receptor are thought to account for many cases of **multinodular goiter**, a very common irregular enlargement of the thyroid gland, particularly when associated with excess secretion of thyroid hormone.

Negative feedback by thyroid hormones occurs at both hypothalamic and pituitary levels. At the pituitary level, thyroxine (T_4) and tri-iodothyronine (T_3) inhibit TSH secretion by decreasing both the biosynthesis and release of TSH through regulation of gene transcription and TSH glycosylation. T_3, the biologically active form of the hormone, is a more potent feedback inhibitor than T_4 and indeed, much of the feedback inhibition by T_4 requires its conversion intracellularly to T_3 by deiodinase type 2.

Thyroxine and tri-iodothyronine

T_4 is produced exclusively in the thyroid gland and is more abundant than T_3, which is the biologically active form

Thyroxine (T_4, also known as tetra-iodothyronine) and T_3 are structurally simple molecules, being iodinated thyronines produced by the coupling of a phenyl group detached from one tyrosine to the phenyl group of a second intact tyrosine (Fig. 39.5). The biosynthesis of T_4 and T_3 occurs on the surface of **thyroglobulin**, a large tyrosine-rich glycoprotein that accounts for about 75% of the protein content of the thyroid gland. Iodination of thyroglobulin occurs after it has been secreted into the follicular lumen, and secretion of T_4 and T_3 requires enzymatic hydrolysis of thyroglobulin, which is located in the follicular colloid. During hydrolysis and subsequent deiodination, the released iodide is conserved and reutilized (Fig. 39.6).

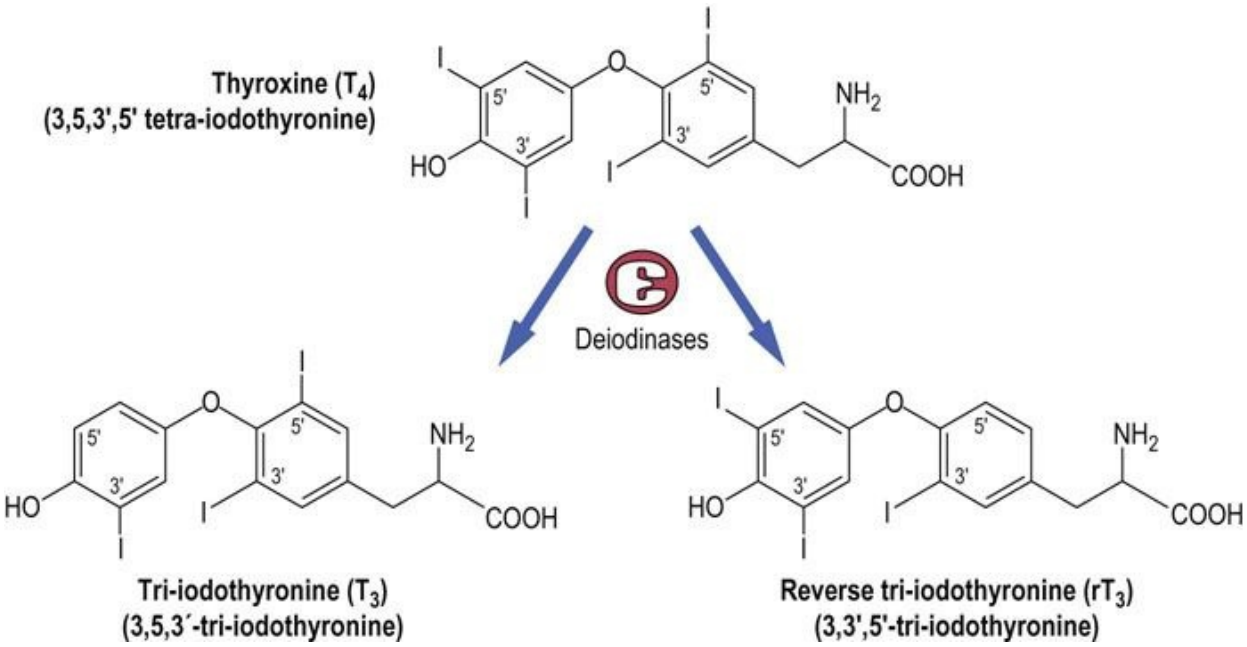


FIG. 39.5 Structures of the thyroid hormones thyroxine (T4), tri-iodothyronine (T3) and reverse T3 (rT3).

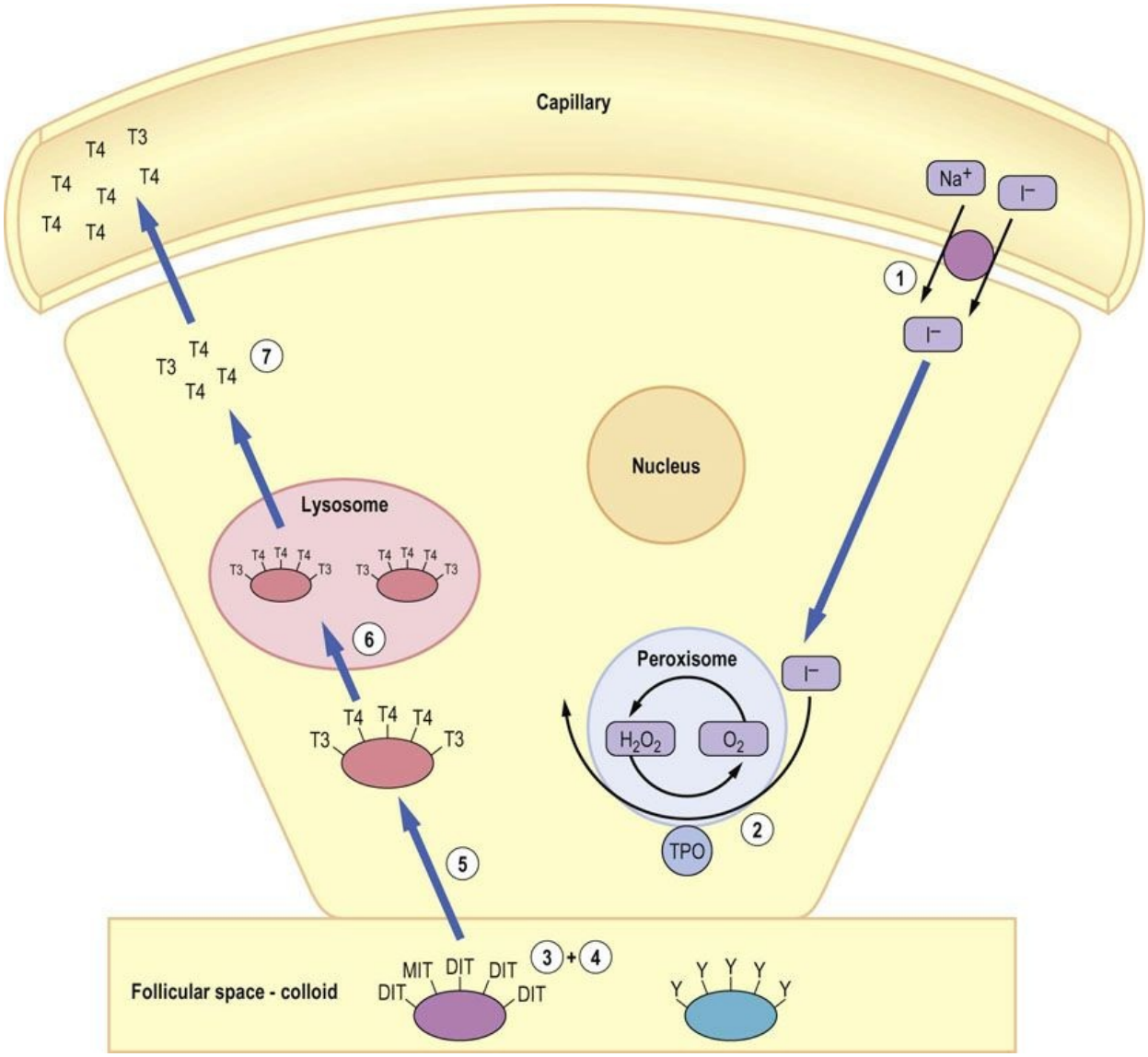


FIG. 39.6 Mechanism of biosynthesis of thyroid hormones.
(1) Iodide is concentrated in follicular epithelial cells after entry via a sodium iodide symporter, before **(2)** oxidation by thyroid peroxidase (TPO) in the peroxisome to iodine. **(3)** At the plasma membrane adjacent to the follicular lumen, conversion of tyrosyl (Y) residues on the surface of thyroglobulin to either mono-iodotyrosine (MIT) or di-iodotyrosine (DIT) occurs. Coupling of iodinated tyrosines to form either T_4 (DIT + DIT) or T_3 (MIT + DIT) then occurs. **(5)** Thyroglobulin is endocytosed and **(6)** hydrolyzed in lysosomes to release free T_3 and free T_4. **(7)** The thyroid hormones are transported to the plasma membrane and released into the bloodstream by mechanisms which are not yet clear. TSH influences virtually every step in thyroid hormone synthesis and release through the cAMP/protein kinase A pathway.

Thyroid hormone bioactivity is regulated by controlling the conversion of T_4 into T_3 by deiodination; this is mediated by **three deiodinase enzymes**.

80–95% of the thyroid hormone secreted by the thyroid gland is T_4, with T_3 as the minor component

However, **T_3,** produced by 5′ deiodination of T_4, is the **biologically active form of the hormone**, and T_4 may thus be regarded as a prohormone. 5′ deiodination may occur in the thyroid gland, in target tissues or in other peripheral tissues, and is accomplished by iodothyronine deiodinases type 1 and 2 (DI1 and DI2). The type 2 enzyme is particularly important in controlling nuclear T_3 levels, while the physiologic role of the kinetically inefficient type 1 deiodinase is less clear at present. The type 3 enzyme is the major catabolic enzyme; it catalyzes the removal of an iodide from the 3 rather than 5′ position, resulting in **reverse T_3** (rT_3), which is inactive, and similarly deiodinates and inactivates T_3. Thus, control of the deiodination of T_4 is one method of controlling thyroid hormone bioactivity. About 80% of T_4 is metabolized by deiodination, with about equal amounts of T_3 and rT_3 being produced. The remaining T_4 is conjugated with sulfate or glucuronic acid and deactivated by deamidation or decarboxylation. In severe illness there is good evidence that activation of T_4 to T_3 by DI1 and DI2 is impaired, while the expression of DI3 is increased. This results in low T_3 levels, which is an early feature of the so-called **'sick euthyroid syndrome'**.

The biologically active fraction of T_3 and T_4 in plasma (free T_3 or free T_4, *i.e.* a fraction which is not bound to protein), represents, in each case, less than 1% of the total concentration of the hormone

The daily production of T_4 is approximately 10% of the extrathyroidal pool, much of which is bound to thyroid hormone binding globulin (TBG) or albumin in plasma. Approximately 80% of this T_4 is converted to T_3 by extrathyroidal deiodination. The turnover of T_3 is much greater than that of T_4. The biologically active component of T_4 and T_3 in plasma is the free fraction – that not bound to proteins. This fraction, which is a measure of thyroid hormone status, represents less than 1% of the total T_4 and T_3. Most clinical laboratories now measure free hormone levels specifically, to avoid problems in interpretation of total hormone levels caused by fluctuations in binding proteins.

Advanced concept box Membrane

transporters for thyroid hormone

Traditionally it was thought that lipophilic hormones such as steroids and thyroid hormones simply diffused across lipid bilayers to reach their intracellular receptors. However, in the last few years it has been shown that **the plasma membrane MCT8 protein functions as a thyroid hormone transporter,** and that genetic deficiency of the protein leads to severe **psychomotor retardation**, and high T_3/T_4 ratios. Affected patients do not have peripheral features of hypothyroidism, however. Since this discovery several different thyroid hormone transporters have been identified, and whether other lipophilic hormones also have similar, specific mechanisms governing cell entry remains an interesting question.

Biochemical actions of thyroid hormones

Thyroid hormones may be considered the accelerator pedal of metabolism

T_3 increases the metabolic rate of a wide range of tissues and thereby increases the whole-body **basal metabolic rate**. Most of these actions result from binding of T_3 to its nuclear receptors α and β, which are encoded by separate genes (Chapter 40), and alteration in transcription rates of genes in target tissues. Among the many cellular changes which result from this altered program of gene expression, it has been shown that thyroid hormones **increase ATP utilization** by increasing sodium-potassium (Na^+/K^+)-dependent ATPase activity, and **increase mitochondrial oxidative metabolism**, in part by direct upregulation of key mitochondrial biogenesis factors. Adipose tissue lipolysis is also stimulated by cAMP-dependent activation of hormone-sensitive lipase, thus producing fatty acids that can be oxidized to generate the ATP used for thermogenesis. Increase in both glycogenolysis and gluconeogenesis occurs to balance the increased use of glucose as a fuel for **thermogenesis**. The rate of synthesis of many structural proteins, enzymes, and other hormones is also affected as a result of thyroid hormone action, and microarray-based

transcriptomics (Chapter 36) can now be used to define the many hundreds or thousands of genes whose expression is altered by activation of the thyroid hormone receptor.

Advanced concept box A tale of two thyroid hormone resistance syndromes

Thyroid hormone exerts its cellular effects through changes in gene expression mediated by two different nuclear receptors, thyroid hormone receptors α and β (TRα and TRβ). TRα is the dominant isoform in bone, GI tract, and in cardiac and skeletal muscle, while TRβ predominates in the liver, kidney and parts of the hypothalamus. In keeping with this, human genetic defects in both TRα and TRβ have been described. TRβ defects are much more common and cause **elevated thyroid hormone levels with detectable TSH**. The very recently described **TRα** defects, in contrast, produce delayed skeletal maturation, constipation and mild mental retardation associated with **normal TSH and high T_3/T_4 ratio** (see Bochukova et al. A mutation in the thyroid hormone receptor α gene. N Engl J Med. 2012;19:366:243–239 and Beck-Peccoz P. Syndromes of thyroid hormone resistance. Ann Endocrinol (Paris). 2005;66:264–269.

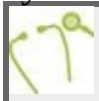
Clinical disorders of thyroid function

Thyroid disease is common, affecting almost 3% of the population; nine times as many women as men are affected

More than 95% of thyroid disease originates in the thyroid gland and much of this is **autoimmune** in origin. Antibodies may arise against several components of thyroid cells, including the peroxidase-rich microsomes. Lymphocyte infiltration and progressive destruction of the thyroid gland ensue, leading to

hypothyroidism, the commonest type of thyroid dysfunction. Some thyroid autoantibodies bind to the TSH receptor. If the autoantibodies bind to but do not stimulate the gland, the thyroid hormone production falls and the patient becomes hypothyroid, with increased plasma TSH and reduced free T_4. If the autoantibodies bind and stimulate, however, they mimic the effect of TSH, breaking the normal negative feedback loop and producing thyroid hormone oversecretion, and the patient will be **hyperthyroid (thyrotoxic)** with increased plasma free T_4 and suppressed TSH.

Hypothalamic and pituitary causes of hypothyroidism are often the result of impaired TSH secretion secondary to pressure from an adjacent **tumor**. Pituitary TSH-secreting tumors are an extremely rare cause of hyperthyroidism, and often produce abnormally high ratios of TSH α-subunit to β-subunit, and abnormal glycosylation patterns of TSH which may be useful in diagnosis.



Clinical box A 60-year-old woman with weight gain, intolerance of cold, and tiredness: hypothyroidism

A 60-year-old woman comes to the clinic and complains of weight gain, intolerance of cold, and tiredness. She also says that she has recently become less alert mentally, but attributes this to aging. On questioning she admits that she has also suffered severe constipation of late. She says that two members of her family had ‘thyroid trouble’. On examination, she is moderately obese, has dry, cool skin, a puffy face, and a slow heart rate of 50 beats per minute. The thyroid gland is not palpable. Her free thyroxine (T_4) was 5 pmol/L (0.4 ng/dL) (reference range 9–25 pmol/L (0.7-2.0 ng/dL)) and TSH was 60 mU/L (reference range 0.4–4 mU/L).

Comment.

Symptoms of hypothyroidism at an early stage can be fairly nonspecific. The best laboratory test for the diagnosis of hypothyroidism is the measurement of the plasma TSH concentration. The elevated level of TSH suggests a primary thyroid disorder. Subsequently this lady's blood was shown to be positive for anti-microsomal and anti-thyroglobulin antibodies. A diagnosis of **lymphocytic thyroiditis (Hashimoto's thyroiditis)**

was made. She was treated with thyroxine.

Clinical box A 35-year-old woman complaining of palpitations and fatigue: hyperthyroidism

A 35-year-old woman came to her physician complaining of palpitations, difficulty climbing stairs and general fatigue. She also said that she had lost 4 kg of weight recently despite a good appetite and no attempt at dieting. She also reported occasional diarrhea, and increasingly infrequent and light menstrual bleeds.

On examination, her skin was warm and moist and she had a fine tremor of outstretched hands. There was mild weakness of the thigh muscles. She had tachycardia (110/min). She also had a mild thyroid enlargement (goiter) and a bruit over the gland. Thyroid function tests show suppressed TSH level (<0.05 mU/L; reference range 0.4–4 mU/L) and increased thyroxine (T_4 = 41 pmol/L (3.3 ng/dL); reference range 9–25 pmol/L (0.7-2.0 ng/dL)) and tri-iodothyronine (T_3 = 17 pmol/L (1105 pg/dL); reference range 3.5–6.5 pmol/L (228-422 pg/dL)). Anti-thyroid receptor antibodies were detected.

Comment.

In hyperthyroidism, the TSH level is suppressed by high circulating thyroid hormones. The low TSH level and high thyroid hormone concentrations suggest hyperthyroidism. The presence of thyroid receptor antibodies confirms that the cause is **Graves' disease, an autoimmune thyroid disorder**.

Advanced concept box The GnRH

pulse generator and human genetics

For the first few months of life the hypothalamic-pituitary-gonadal axis is active before it is shut down, going into a quiescent state until it is reactivated at puberty. There are several lines of evidence suggesting that this is largely determined centrally, and a poorly defined group of hypothalamic neurons known as the GnRH ‘pulse generator’ are the main focus of interest. Human genetic studies over the past decade combined with study of genetically modified mice have, however, yielded some key insights, and it is now known that the **neuropeptides kisspeptin and neurokinin B** are critically required for normal **maturation of GnRH secretion**, with genetic defects in the genes encoding either kisspeptin, neurokinin B or either of their receptors leading to absent or severely impaired maturation. Much work is now being devoted to understanding the behavior of a group of neurons that are closely apposed to GnRH neurons, and that co-express kisspeptin, neurokinin and also dynorphin, as these neurones have connections and properties that suggest they may form an integral part of the elusive GnRH pulse generator.

The hypothalamo–pituitary–adrenal axis

Corticotropin-releasing hormone (CRH)

CRH is a 41-amino acid peptide secreted by the paraventricular nucleus (PVN)

CRH acts via a G-protein coupled receptor on pituitary corticotroph cells via the cAMP second messenger system to stimulate both synthesis and secretion of ACTH. A second hormone from the PVN, vasopressin (VP), potentiates the response of the pituitary to CRH, in part by increasing the amount and extent of expression of the CRH receptor. Negative feedback by cortisol inhibits both CRH and VP secretion, as well as reducing responsiveness of corticotrophs to stimulation.

Adrenocorticotropic hormone (ACTH)

ACTH is synthesized as a 241-amino acid precursor molecule, pro-opiomelanocortin (POMC)

POMC is cleaved at multiple sites to release several hormonally active peptides, including the **endorphins** and **melanocyte-stimulating hormones**. In addition to the pituitary, POMC may also be produced in large quantities by certain malignancies, giving rise to **ectopic ACTH syndrome**.

ACTH itself is composed of 39 amino acids with the biological activity residing in the *N*-terminal 24 residues

ACTH is secreted in stress-related bursts superimposed on a marked diurnal rhythm, with a rapid surge in production from around 03.00 h and a peak plasma concentration at about 05.00 h. It is transported unbound in plasma and has a half-life of about 10 minutes. ACTH stimulates the synthesis and release of glucocorticoid hormones by interacting with cell surface G-protein coupled receptors on the adrenal cortex that stimulate cAMP production. Acute increases in the adrenal synthesis of cortisol occur within 3 minutes, principally by stimulating the activity of cholesterol esterase. Longer-term effects of ACTH

include induction of transcription of the genes that encode steroidogenic enzymes, and it is also required to maintain the *zona fasciculata* and *zona reticularis* of the gland, meaning that long-term lack of ACTH eventually abolishes the ability of the gland to respond to an acute challenge by increasing cortisol and androgen synthesis (Chapter 17).

Negative feedback by cortisol occurs within two timeframes, acting at both the hypothalamic and pituitary levels. **Fast feedback** alters the release of hypothalamic CRH and the CRH-mediated secretion of ACTH. **Slow feedback** results from reduced synthesis of CRH plus suppression of POMC gene transcription, which results in reduced ACTH synthesis.

Biosynthesis of cortisol

Cortisol is the major glucocorticoid synthesized in man in the adrenal cortex, and is under the direct control of pituitary ACTH

A simplified scheme of steroid biosynthesis is shown in Figure 39.7 (see also Fig. 17.12). **Cholesterol** is the precursor for all steroid hormones. Cleavage of the cholesterol side chain liberates the **C-21 corticosteroids**; further side chain cleavage yields the **C-19 androgens**, and aromatization of the A ring results in the **C-18 estrogens**. The structures of the key steroid hormones are shown in Chapter 17. Several of the steroidogenic enzymes are members of the cytochrome P-450 superfamily of oxidases.

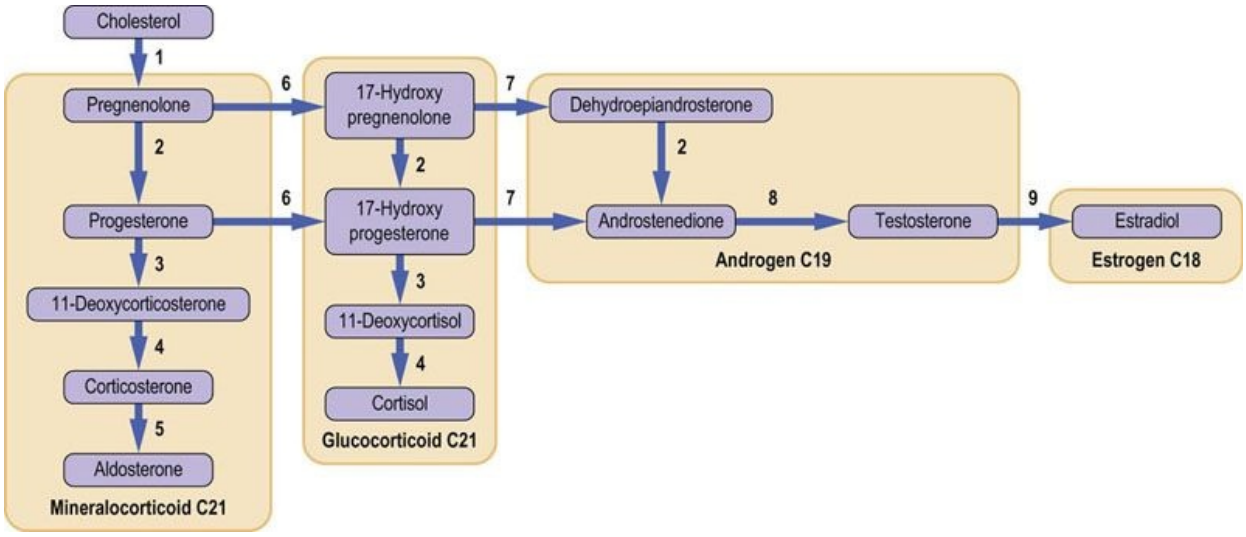


FIG. 39.7 Summary of steroid hormone biosynthesis. **(1)** cholesterol 20,22-desmolase; **(2)** 3β-hydroxysteroid dehydrogenase $\Delta\alpha^{4,5}$-oxosteroid isomerase; **(3)** 21-hydroxylase; **(4)** 11-hydroxylase; **(5)** aldosterone synthase; **(6)** 17α-hydroxylase; **(7)** 17,20-lyase/desmolase; **(8)** 17β-hydroxysteroid dehydrogenase; **(9)** aromatase. These enzymes are part of the cytochrome P-450 dependent superfamily of NADPH-dependent monooxygenases (see also Chapter 30; compare Figs 17.11 and 17.12).

The plasma concentration of **cortisol** shows a pronounced **diurnal rhythm**, being some 10 times higher at 08.00 h than at 24.00 h. This parallels the marked diurnal rhythm of secretion of ACTH. Approximately 95% of cortisol in plasma is bound to proteins, mainly the **cortisol-binding globulin**, (CBG). As cortisol concentration rises, the percentage of free cortisol also rises, indicating that CBG binding is saturable. Cortisol has a half-life in plasma of about 100 minutes. It is metabolized in the liver and other organs by a combination of reduction, side chain cleavage, and conjugation to produce a wide range of inactive metabolites that are excreted in urine (see Fig. 17.13).

Cortisol has diverse effects on metabolism, and tissue growth and repair

Cortisol has wide-ranging effects on metabolism, immune function, the cardiovascular system and skeleton, mediated by alterations in expression level of thousands of genes. As the name glucocorticoid suggests, cortisol has a major influence on **glucose homeostasis**: it is a counterregulatory hormone, working through nuclear receptors to induce the biosynthesis of gluconeogenic enzymes, while at the same time inhibiting glucose uptake and metabolism in peripheral tissues. Glycogen synthesis and deposition are increased, and lipolysis in adipose

tissue is stimulated. Protein and RNA synthesis are stimulated in the liver but inhibited in muscle; proteolysis in muscle produces the amino acid substrates for hepatic gluconeogenesis. Cortisol works in tandem with several other hormones including insulin and GH in regulating intermediary metabolism, and these metabolic effects are best understood as part of the **endocrine response to physiologic stress** (Chapter 21).

Cortisol also modulates the immune system through effects on cytokine production and enhanced leukocyte apoptosis, effects which are exploited pharmacologically in the **use of potent glucocorticoids to treat inflammatory conditions.** It is, moreover, permissive for the effect of many vasoconstrictors such as **catecholamines,** has direct effects on the heart and on the kidney, where it enhances water clearance and electrolyte reabsorption, and bone, influencing turnover through a variety of mechanisms.

Clinical disorders of cortisol secretion

Hyposecretion of cortisol may occur as a result of hypothalamic, pituitary or adrenal failure

The diagnosis of cortisol deficiency and its cause relies on the clinical presentation, carefully timed measurement of cortisol and ACTH, and the extent of the cortisol response to synthetic ACTH (Synacthen).

Addison's disease refers specifically to primary adrenal failure in which secretion of all adrenal hormones is decreased, usually because of autoimmune disease or infection (e.g. by tuberculosis or cytomegalovirus). Biochemically, it is characterized by low serum sodium, high serum potassium, acidosis, and impaired cortisol response to synthetic ACTH (Synacthen), together with an elevated plasma ACTH. Cortisol replacement, usually together with a mineralocorticoid, is an effective treatment for this life-threatening condition. **Addison's disease can be associated with a markedly elevated TSH which resolves with glucocorticoid therapy.** This is important to recognize, as erroneous diagnosis of primary hypothyroidism and treatment with thyroxine may dangerously exacerbate signs and symptoms of hypoadrenalism.

Deficiencies of CRH and ACTH are only rarely isolated and more commonly accompanied by deficiencies of other hypothalamic or pituitary hormones due, for example, to a compressive pituitary tumor. Because glucocorticoids are also involved in normal clearance of free water, their

deficiency can mask coexisting ADH deficiency, and replacement leads to dramatic diuresis.

Clinical box A 40-year-old woman who collapsed at home: primary adrenal insufficiency is an endocrine emergency

A 40-year-old woman was admitted as a medical emergency following a collapse at home 2 days after contracting influenza. For 1 year she had been complaining of chronic weakness and fatigue, and lately of abdominal pain, nausea, vomiting, anorexia and confusion. She was dehydrated and hypotensive with pigmentation of the palmar creases and buccal mucosa. Biochemical analysis revealed Na^+ 115 mmol/L (135–145 mmol/L), K^+ 5.9 mmol/L (3.5–5.0 mmol/L), urea 12 mmol/L (2.5–6.5 mmol/L) and glucose 2.9 mmol/L (4.0–6.0 mmol/L). Baseline serum cortisol was 55 nmol/L (2 ug/dL), and did not rise adequately 30 min after administration of synthetic ACTH (Synacthen).

Comment.

This is an acute presentation of **Addison's disease** most likely resulting from progressive **autoimmune destruction of the adrenal cortex**. The nonspecific symptoms of cortisol deficiency became critical after the stress of a minor illness. The hypotension and electrolyte imbalance result from mineralocorticoid deficiency. The pigmentation is due to elevated POMC-derived peptides such as the melanocyte-stimulating hormones. The diagnosis is made by the **impaired cortisol response to synthetic ACTH**. Treatment is rehydration with intravenous saline plus intravenous hydrocortisone followed by daily oral hydrocortisone and a mineralocorticoid. Other autoimmune diseases commonly coexist in the patient or their family. **Increased frequency of severe hypoglycemia in patients with type 1 diabetes** is one presentation to watch for, because of the loss of the glucose counterregulatory effects of cortisol.

Cortisol hyposecretion may also result from a genetic disorder in the steroid biosynthetic pathway: congenital adrenal hyperplasia

The most frequently affected enzyme is the **steroid 21-hydroxylase** (Fig. 39.7). In nearly complete enzyme deficiency, deficiency of cortisol and aldosterone leads to **salt wasting and hyponatremia in the first few days of life,** a dramatic presentation which is easy to recognize. Because of the loss of negative feedback by cortisol, ACTH levels rise and try to drive increased amounts of substrate through the enzymatic block at the expense of making an excess of 17-hydroxyprogesterone (17OHP), which is converted to androgens and may masculinize female external genitalia before birth. Partial defects in the enzyme may be compensated for by elevated ACTH levels sufficiently to maintain cortisol levels, though the build-up of metabolites proximal to the defective enzyme in the pathway still occurs. Consequently this presents in teenage girls as **hirsutism** (growth of male type and distribution of hair in women) and **menstrual irregularity.** The steroid profile of a patient with congenital adrenal hyperplasia is given in Fig.17.13.

Hypersecretion of cortisol results in Cushing's syndrome – among the most challenging of endocrine disorders

Exogenous glucocorticoids, commonly used to suppress the immune system in a range of inflammatory disorders, are the commonest cause of clinical **Cushing's syndrome**. However, Cushing's syndrome may also result from disorders of the hypothalamus, pituitary (around 80%) or adrenal gland (around 15%), or it may be the consequence of ectopic ACTH syndrome. Cortisol excess leads to remodeling of adipose tissue with deposition of fat in the face and trunk and loss of fat on the limbs, wasting of skeletal muscle, thinning of skin, and slow healing (Fig. 39.8). Cortisol excess also commonly produces **diabetes and hypertension**, and usually suppresses the hypothalamic gonadal axis, seen as **cessation of menstrual bleeds** in women.

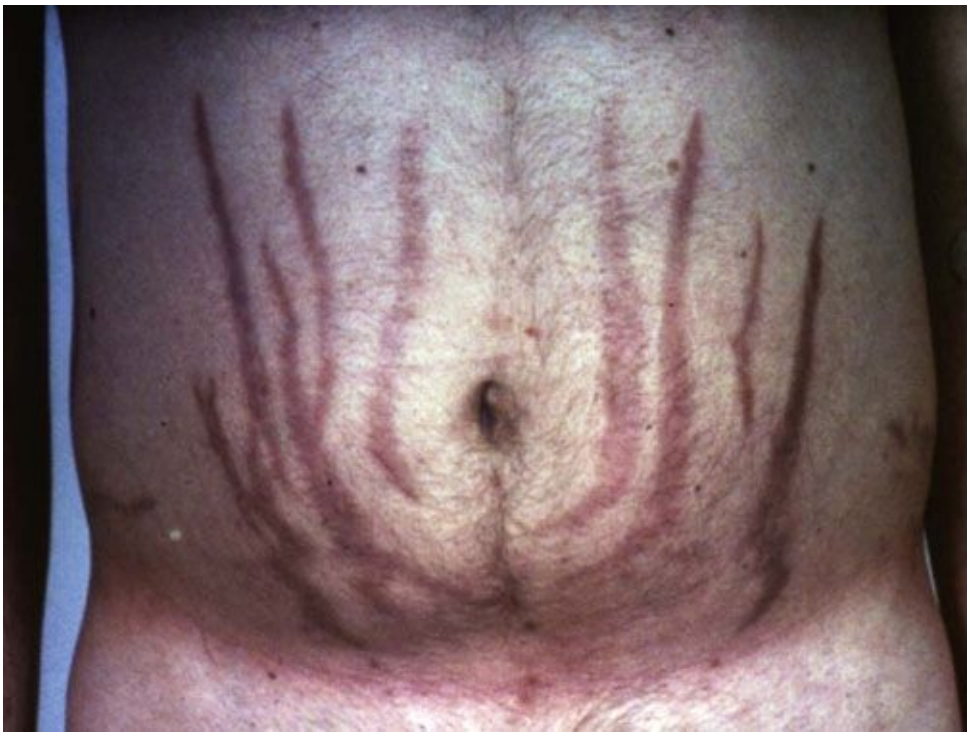


FIG. 39.8 Cushing's syndrome due to cortisol hypersecretion.
Livid ‘striae’ are shown, caused by weakening of dermal connective tissue allied to increased abdominal fat tissue.

Clinical box A woman with fatigue, and rapid weight gain cushing's syndrome may present a diagnostic dilemma

A 45-year-old woman presented with fatigue, rapid weight gain with central obesity, fullness and redness of her face (so called **‘plethora’**), and loss of regular menstrual periods. She was mildly hypertensive, and her family doctor had found her also to be diabetic for which she had received dietary advice. Urinary free cortisol was 1000 nmol/24 h (38 ug/dL) (reference range <250 nmol/24 h (<9 ug/dL/24h)); serum cortisol was 500 nmol/L at 24 h (reference range <50 nmol/L) and her 0800 h cortisol was 550 nmol/L (20.9 ug/dL) after 1 mg of dexamethasone (a potent synthetic glucocorticoid) (reference range <50 nmol/L) (<1.8 ug/dL). Plasma ACTH was 100 ng/L (reference range

<80 ng/L).

Comment.

This woman has **hypercortisolism, known as Cushing's syndrome.** The elevated urine cortisol, elevated midnight serum cortisol and failure to suppress following dexamethasone administration support the diagnosis. The ACTH result indicates that the cause is either a pituitary tumor (most likely) or ectopic ACTH secretion from an occult tumor. In this case **magnetic resonance imaging** (MRI) of the pituitary revealed a clear **tumor** of 0.5 cm diameter; diagnosis of Cushing's syndrome and localization of the problem are often not this easy! **Pituitary surgery** is the appropriate definitive treatment in this case.

The diagnosis of Cushing's syndrome may be extremely demanding for several reasons

Because of the pronounced circadian variation in serum cortisol levels, random measurement of cortisol is of little use in the diagnosis of Cushing's syndrome. Instead, 24-hour urinary free cortisol, which provides a measurement of total production of cortisol over 1 day, is a common screening test. An abnormal circadian profile of cortisol secretion, with sustained levels at night when they usually almost undetectable, may produce Cushing's syndrome even if there is little, if any, increase in total daily production of cortisol. This may be detected by sampling of cortisol at 09.00 h and midnight, showing **loss of normal circadian rhythm**. If this test is undertaken carefully without provoking an endocrine stress response by taking blood, it can be a sensitive means of detecting hypercortisolism.

Endogenous Cushing's syndrome usually arises from autonomously functioning secretory tumors

Some of the diagnostic challenges in Cushing's syndrome arise from the variable behavior of such tumors as they evolve. Sometimes only episodic cortisol hypersecretion is seen, with testing normal in between episodes, and sometimes aberrant receptor expression due to somatic mutations means that high levels of

cortisol are secreted in response to other hormones such as gut-derived peptides.

Once these pitfalls have been overcome and cortisol excess demonstrated, the anatomic site of the problem can be investigated using a range of tests which may include **ACTH measurements, selective venous catheterization of the pituitary or the adrenal glands, and dexamethasone (a synthetic glucocorticoid) suppression testing**. The principle of this suppression testing is that tumors which have become autonomous will not exhibit normal complete suppression of secretory activity by a dose of potent glucocorticoid. Definitive treatment is usually surgical and should be targeted at the primary cause of the condition. However, some medical means of reducing cortisol synthesis are commonly used in preparation for surgery.

The hypothalamo–pituitary–gonadal axis

Gonadotropin-releasing hormone (GnRH)

GnRH is essential for the secretion of intact FSH and LH

The hypothalamus has a major role in the control of gonadal function in both males and females, acting via secretion of GnRH. GnRH is a decapeptide synthesized by various **hypothalamic nuclei** and secreted by a relatively small number of neurons into the portal system for transfer to the pituitary. GnRH neurons thus serve as a fairly narrow conduit for transmission of central reproductive signals, and are vulnerable to many types of pathology. GnRH secretion is **highly pulsatile** and it induces the synthesis and secretion of both **FSH** and **LH**. GnRH secretion is held in quiescence between infancy and puberty. Increased frequency and amplitude of hypothalamic GnRH pulses is the first measurable stage of the onset of puberty. Its release is subject to negative feedback by progesterone, prolactin, and sometimes estrogen, among other hormones. GnRH acts through its cell surface receptor to increase intracellular calcium, hydrolyze phospho-inositides, and activate protein kinase C. Long-acting GnRH agonists can also cause downregulation of GnRH receptors and reduced FSH and LH secretion. Such agonists are now used to treat **prostate cancer** and to prepare infertile women for **assisted-conception programs.**

Advanced concept box Human growth and longevity

Although genetic defects in GH or IGF-I action lead to severely impaired growth, studies in many experimental organisms suggest that reduced IGF-I signaling may extend lifespan. Recent studies in Ecuador of individuals affected by Laron-type dwarfism due to GH receptor mutation suggested that this may also be true in humans, as despite their short stature, those with GH receptor defects appeared protected from diabetes, cancer, and had improved biochemical markers of ageing. (See Guevara-Aguirre J et al. Growth hormone receptor deficiency is associated with a major reduction in pro-aging signaling, cancer, and diabetes in humans.

Sci Transl Med 2011;3:70ra13.)

Follicle-stimulating hormone and luteinizing hormone

Although FSH and LH have been given their names on the basis of their function in the female, they are also secreted and functional in the male

Both FSH and LH are secreted in the male and female under the influence of GnRH, and alterations in the GnRH pulse frequency and amplitude can influence the relative amounts of FSH and LH secreted by the gonadotroph. FSH and LH are glycoproteins with an identical α-subunit (also shared with TSH) and a specific β-subunit. The gene for LH has recently undergone duplication to produce βHCG, a gonadotropin secreted by embryonic tissues.

GnRH is essential for the secretion of intact FSH and LH, but feedback from estradiol and testosterone plus gonadal peptides such as inhibin have an important role in reproductive function. Feedback by estradiol is especially interesting because it may have either a negative or a positive effect on gonadotropins, depending on the stage of the menstrual cycle. An outline of the control of the hypothalamo-pituitary-gonadal axes, both for adult men and women, is shown in Figure 39.9. The half-life of LH in plasma is approximately 50 minutes; FSH has a longer half-life of about 4 hours. Both FSH and LH concentrations vary considerably depending on age and sex.

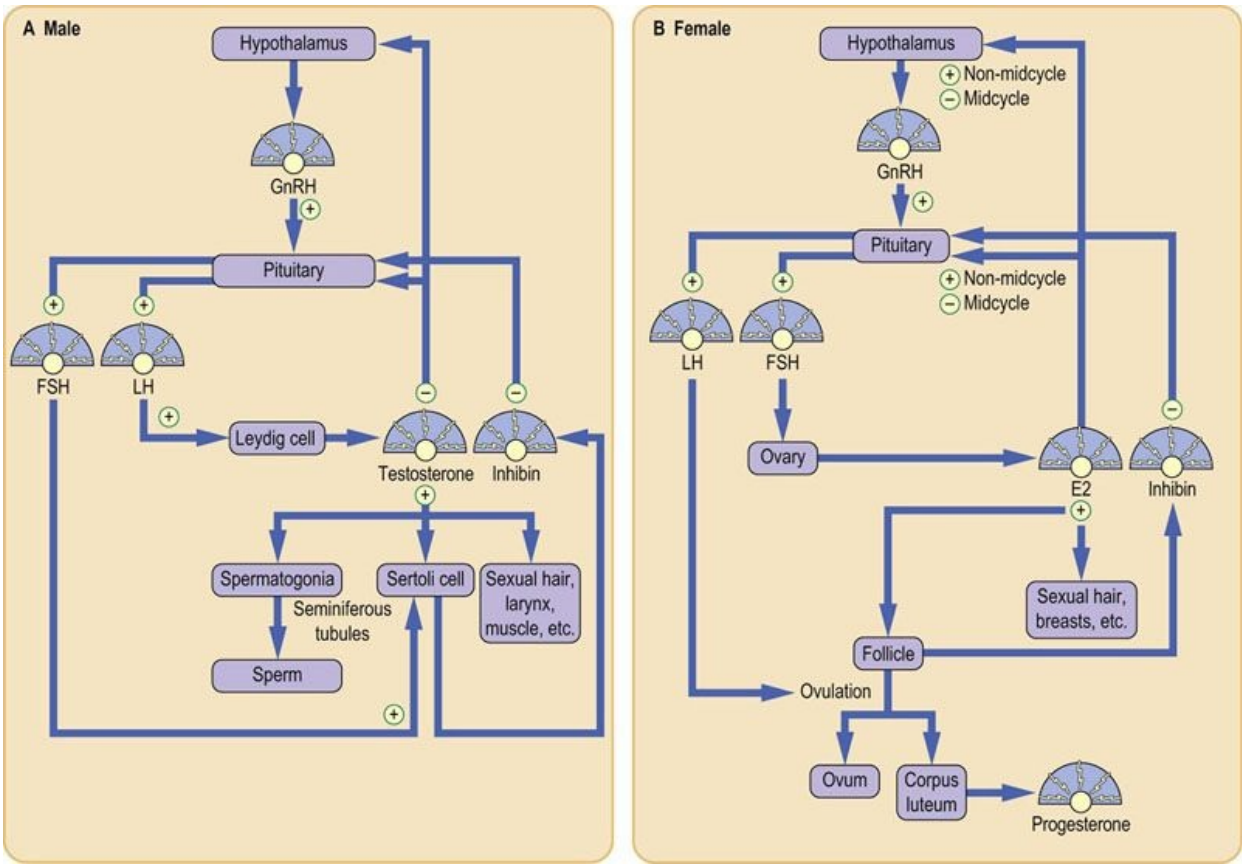


FIG. 39.9 Control of the hypothalamo-pituitary-gonadal axes.
(A) In men, testosterone is produced from cholesterol in the Leydig cell under LH stimulation. Testosterone and FSH support spermatogenesis. **(B)** In women, estradiol (E_2) is produced by the granulosa cell and the developing follicle after feedback stimulation. E_2 feedback is mainly negative but, in midcycle, there is a positive E_2 feedback resulting in the surge of LH that causes ovulation. Progesterone (P) is secreted by the resultant corpus luteum.

Actions of FSH and LH on the testes

FSH and LH influence spermatogenesis

In the male, testosterone biosynthesis occurs in the Leydig cells of the testes under the primary influence of LH (Fig. 39.10) acting through the G-protein-coupled LH receptor. High intratubular testosterone levels cooperate with FSH in promoting spermatogenesis in the spermatic tubule: FSH binds to its specific receptor on the Sertoli cell of the testes and, by a mechanism similar to LH, induces increased synthesis of several proteins, including **androgen-binding**

protein (ABP) and **inhibin**. ABP is secreted into the seminiferous tubular lumen where it binds testosterone (or its active form dihydrotestosterone). This ensures a high local androgen concentration that, together with FSH, brings about the meiotic divisions that are necessary for spermatogenesis. Inhibin has a role in the FSH negative feedback loop.

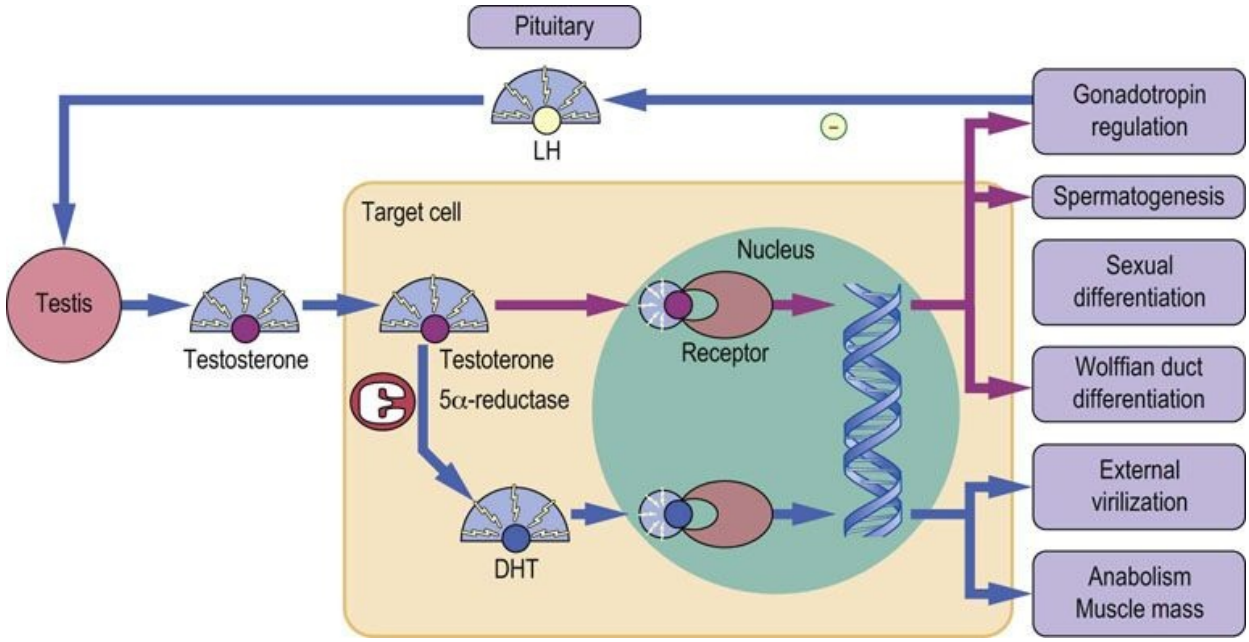


FIG. 39.10 Mechanism of action of testosterone.
Testosterone from the testis enters a target cell and binds to the androgen receptor, either directly or after conversion to 5α-dihydrotestosterone (DHT). DHT binds more tightly than testosterone, and the DHT–receptor complex binds more efficiently to chromatin. Actions mediated by testosterone are shown by purple lines, those mediated by DHT are shown by blue lines.

Biochemical actions of testosterone in the male

Testosterone not only influences gonadotropin regulation and spermatogenesis, but is also a natural anabolic steroid

Nearly 95% of serum testosterone comes from the testes, the remainder originating from the adrenal gland. More than 97% of this circulating testosterone is bound to protein, with equal amounts bound to **albumin** and to the specific **sex hormone-binding globulin** (SHBG), which is similar in

structure to ABP. Testosterone is metabolized in target tissues to **dihydrotestosterone** (DHT) via 5α-reduction, more than doubling its affinity for the nuclear androgen receptor. In addition to effects on gonadotropin regulation and spermatogenesis, androgens induce the development of the reproductive tract from the Wolffian ducts during male sexual differentiation. They also have anabolic effects, stimulating protein synthesis and an increase in muscle mass (Fig. 39.10).

Clinical disorders of testosterone secretion in males

Testosterone deficiency may originate from a wide range of disorders of the hypothalamus, pituitary or testes

Endocrine failure of the testes may occur due to trauma or inflammation of the testes themselves (e.g. due to mumps infection), or due to failure at the level of the hypothalamus or pituitary. Indeed, the **gonadotrophs are among the most sensitive of the anterior pituitary cell types to pituitary damage** (e.g. due to growth of an adenoma within the bony constraints of the *sella turcica*) and consequently gonadal failure is often the earliest manifestation of pituitary disease. Central hypogonadism may also be a response to **severe weight loss or physiological stress**, when it is best seen as an emergency energy-saving mechanism, and is additionally seen in **Cushing's syndrome** or chronic use of **opioid drugs**.

There is a series of genetic disorders which manifest as gonadal failure

Klinefelter's syndrome, caused by acquisition of an additional X chromosome in males (karyotype 47, XXY) presents with gynecomastia (abnormal increase in size of the male breast), eunuchoidism and varying degrees of hypogonadism, and is due to abnormal development of the testes. Serum FSH is elevated, and LH is usually elevated, but testosterone may be subnormal. **Kallmann's syndrome**, in contrast, is a genetic hypothalamic disorder in children, featuring both deficient GnRH production and commonly absent smell (anosmia) due to defective migration of GnRH-producing (and other) neurones to hypothalamus during intrauterine development. Affected individuals present with delayed

puberty and subnormal FSH, LH, and testosterone. Androgen deficiency may lead to muscle and fat redistribution and loss of bone mineral density, and these may be corrected, to some extent, by androgen administration (Table 39.6).

Table 39.6
Causes of hypogonadism

Site of defect	Biochemistry	Causes
Hypothalamus	Low GnRH* Low LH, FSH Low E2/testosterone	Tumors Irradiation Genetic disorders, *e.g.* Kallmann's syndrome Mutations in kisspeptin or its receptor Mutations in neurokinin B or its receptor
Pituitary	High GnRH* Low LH, FSH Low E2/testosterone	Tumors Irradiation Trauma Inflammatory conditions Genetic disorders GnRHR mutations Mutations in pituitary transcription factors
Gonads	High GnRH* High LH, FSH Low E2/testosterone	Irradiation Physiologic (menopause) Autoimmunity Infection (e.g. mumps orchitis in men) Trauma Genetic/developmental abnormalities Klinefelter's syndrome (47, XXY: men) Turner's syndrome (45, XO: women)

E2, estradiol

*GnRH is not easily measured clinically.

Androgen excess in males is only seen in precocious puberty

Precocious puberty is a rare condition that may result from early maturation of the normal hypothalamo-pituitary-gonadal axis, gain-of-function mutations in the LH receptor or a receptor for another protein known as kisspeptin (metastin), or as a result of a tumor that is secreting either androgen or HCG.

Actions of FSH and LH on the ovary

In the female, FSH promotes estradiol synthesis leading to follicular maturation, while LH leads to follicle rupture and oocyte release

In the mature female, the GnRH pulse generator engages in a dynamic interplay with ovarian cells in maintaining the hormonally driven **menstrual cycle** (Figs 39.9 and 39.11). Primordial follicles actually begin hormone-independent growth and maturation many weeks before the start of the menstrual cycle, but only at the start of the cycle, when they have acquired the capacity to respond to FSH, are a few follicles preserved from atresia. Rising FSH concentrations stimulate estradiol synthesis in granulosa cells through induction of aromatase and other enzymes (Fig. 39.7). As estradiol is secreted so FSH falls, and this combination plays an important role in the selection of a dominant follicle for further development, though details of the precise mechanism remain to be fully established. Follicular maturation continues under the influence of rising estradiol concentrations. Increasing estradiol because of positive feedback within the dominant follicle, causes the negative central feedback of estrogen to flip into positive feedback, initiating a surge of LH. This LH binds to receptors on the dominant follicle and, in tandem with steroid hormones and other factors such as prostaglandins, induces completion of the first stage of oocyte meiosis, rupture of the follicle and release of the oocyte some 36 hours later. At this time there is a sharp fall in plasma estradiol, followed by a fall in LH.

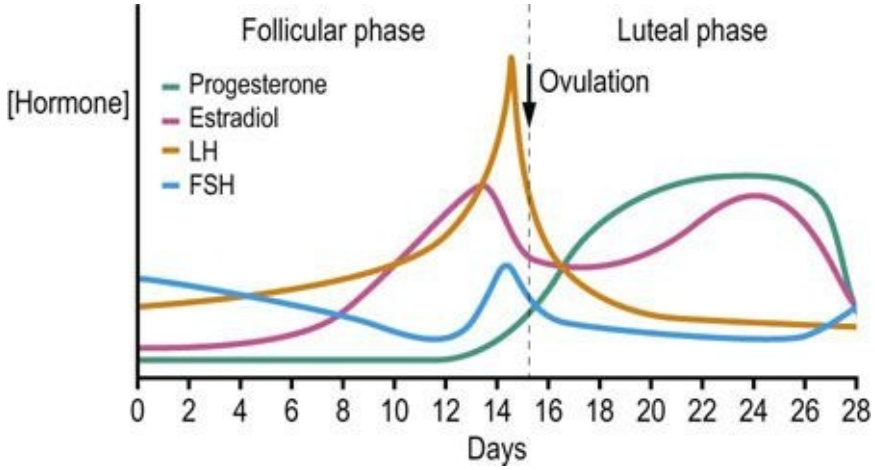


FIG. 39.11 Changes in hormone secretion during the normal menstrual cycle. LH, luteinizing hormone; FSH, follicle-stimulating hormone.

The ruptured follicle transforms into the *corpus luteum*, which secretes progesterone and lesser amounts of estradiol both to sustain the oocyte and to prepare the estrogen-primed uterine endometrium for implantation of a fertilized

ovum and the establishment of early pregnancy. In the absence of fertilization, corpus luteum function declines, and progesterone and estradiol secretion falls. This brings about vascular changes in the endometrium leading to tissue involution and menstruation. The decline in steroid secretion stimulates FSH secretion, setting the stage for initiation of the next cycle.

Clinical box Hormone replacement therapy (HRT)

Women normally reach menopause at 45–55 years of age. The absence of ovarian follicles leads to an absence of estrogen and inhibin secretion, resulting in a marked rise in pituitary FSH. There are vasomotor symptoms of estrogen deficiency (flushing) associated with menopause. Long-term postmenopausal estrogen deficiency is known to increase the rate of bone loss, leading to **osteoporosis**, and also alters lipoprotein metabolism in a way that increases the **risk of cardiovascular disease.** HRT with an estrogen preparation was standard practice in most developed countries both to relieve the acute symptoms and to provide prophylaxis against the long-term risks. However, recent studies have suggested that HRT may increase a woman's risk for serious conditions including **breast cancer, cardiovascular disease, stroke, and pulmonary embolism.** Although HRT did have a modest effect on preventing osteoporosis, alternative bone-targeting drugs are also effective. The current recommendations are that **HRT should only be used as short-term therapy** for vasomotor symptoms that do not respond to other treatment.

Actions of steroid hormones in the female

In a woman with a normal menstrual cycle, it is the progesterone level that is of diagnostic significance

The concentrations of FSH, LH, estradiol, and progesterone vary considerably during the **menstrual cycle.** In a normally cycling woman, **progesterone** is of

particular diagnostic utility: a serum concentration of more than around 20 nmol/L in the luteal phase (precise values vary among laboratories) is consistent with ovulation (Fig. 39.11). Aside from their roles in the menstrual cycle and reproduction, both estradiol and progesterone have other effects, acting via specific nuclear receptors in target cells. Estradiol, working in tandem with other hormones such as insulin-like growth factor-I (IGF-I), is responsible for linear growth, breast development and maturation of the urogenital tract and the female body shape. In adult life, both estradiol and progesterone support breast function, and estradiol has an important role in preserving bone mineral density. Progesterone is responsible for the rise in basal body temperature during the luteal phase of the menstrual cycle and decreases in progesterone secretion may contribute to premenstrual changes in mood.

Disorders of steroid secretion in the female

Just as in the male, endocrine disorders producing subnormal sex steroid secretion in the female may result from a wide range of disorders in the hypothalamus, pituitary or ovary

Just as for the testes, endocrine failure of the ovaries may occur due to damage or dysfunction of the ovaries themselves, or due to defects at the level of the hypothalamus or pituitary. Indeed, several pathologies, including pituitary adenomas and Kallmann's syndrome, are common to men and women. However, in women the most common genetic disorder affecting the ovary is ***Turner's syndrome*** (karyotype 45X), which has characteristic somatic features including short stature, elevated plasma FSH and prepubertal estradiol, and vestigial or ‘streak’ ovaries. Various other endocrine causes of infertility are shown in Table 39.6, but are largely beyond the scope of this text.

Syndromes of excess ovarian steroid secretion lead to precocious puberty in children and infertility and/or hirsutism in the adult

Precocious puberty may arise from early maturation of the normal hypothalamic gonadal axis, an estradiol-or androgen-secreting cyst or tumor of the ovary or the adrenal gland, congenital adrenal hyperplasia, or severe insulin

resistance. In the mature female, hypersecretion of androgen in the **polycystic ovary syndrome** (PCOS) may result in infertility and/or hirsutism. Although the etiology of PCOS remains uncertain, it shows a strong genetic predisposition and current theories favor a key role for aberrant intraovarian androgen metabolism, in some cases strongly influenced by systemic insulin resistance. The nature of the clinically important cross-talk between insulin action and ovarian androgen production is not yet fully worked out, however.

The growth hormone axis

Growth hormone-releasing hormone (GHRH) and somatostatin

GHRH is a 44-amino acid peptide synthesized as part of a 108-amino acid prohormone in the arcuate and ventro-medial nuclei of the hypothalamus, and in the median eminence

GHRH binds to its receptor on the pituitary somatotroph cell and triggers both the adenylyl cyclase and intracellular calcium–calmodulin systems (Chapter 40) to stimulate GH transcription and secretion from the anterior pituitary. Negative feedback from GH and IGF-I results in both a decrease in GHRH synthesis and secretion and an increase in somatostatin synthesis and secretion.

Somatostatin is found as two isoforms, with 14 and 28 amino acids respectively, both of which are produced from the same 116-amino acid gene product. Somatostatin and its receptors are found throughout the brain and also in other organs, notably the gut. Binding of somatostatin to its receptor is coupled to adenylyl cyclase by an inhibitory guanine nucleotide-binding protein, resulting in a decrease in intracellular cAMP. In the context of growth, somatostatin inhibits the secretion of GH. GHRH and somatostatin are released in separate bursts to provide a very fine level of control of GH release. Somatostatin also inhibits the production or action of many other hormones including TSH, insulin, and glucagon and gastrin. **Long-acting analogs of somatostatin are thus effective in the management of GH excess** as well as tumors secreting a wide range of other hormones, and as an adjunct to pancreatic surgery to inhibit exocrine secretions from the pancreatic residue during healing.

Growth hormone

The greatest secretion of GH occurs in children and young adults, chiefly during sleep

Although GH can exist in variant forms, the major species is a 22 kDa protein.

Nearly two-thirds of GH in the circulation is associated with a 29 kDa **binding protein** that is identical to the extracellular domain of the GH receptor. This binding protein prolongs the half-life of GH in plasma to about 20 minutes.

The normal human pituitary contains approximately 10 mg of GH; less than 5% of this is released each day. GH is released in bursts with a periodicity of 3–4 hours, and greatest secretory activity occurs during sleep. At the peak of a secretory burst, the plasma GH concentration may be 100-fold greater than baseline; this means that **no reference interval can be derived for this hormone,** and so meaningful measurement requires either a **provocative test** or **multiple samples** over the course of a day. Secretory bursts occur most frequently in children and young adults.

GH has a wide range of actions in regulating growth and intermediary metabolism

It is not surprising, in view of the wide range of actions of GH, that several factors other than GHRH and somatostatin can influence GH secretion. These include other hormones (estradiol) and metabolic fuels (glucose). Several pharmacologic agents are also known to influence GH secretion. These factors act at the higher centers and the hypothalamus to alter the pattern of GHRH and somatostatin secretion.

Binding of GH to its receptor precipitates a complex series of intracellular events that lead to the transcription of many enzymes, hormones and growth factors, including IGF-I

The diversity of action of GH makes it difficult to understand or integrate all of its functions; thus it is convenient to think in two distinct phases (Fig. 39.12). The direct actions of GH are on **lipid, carbohydrate, and protein metabolism**. During hypoglycemia, GH stimulates lipolysis and induces peripheral resistance to insulin. These effects stimulate the use of fatty acids in peripheral tissues, sparing glucose for use in the brain. Previously, GH therapy was used only for treatment of GH deficiency in children in order to stimulate growth; however, treatment is now often extended into adulthood when it can sometimes improve quality of life. This change in treatment strategy illustrates the complex effects of GH on metabolism. During growth, GH stimulates the uptake of amino acids and their incorporation into protein, especially in muscle. The **indirect actions of GH are mediated by IGF-I.** These actions promote the proliferation of

chondrocytes and the synthesis of cartilage matrix in skeletal tissues. Although GH is most often associated with its role in stimulating linear growth, it clearly has other roles in influencing the relative amount and distribution of fat and muscle tissue. Its anabolic effects are the rationale for its current **illicit use in sport**.

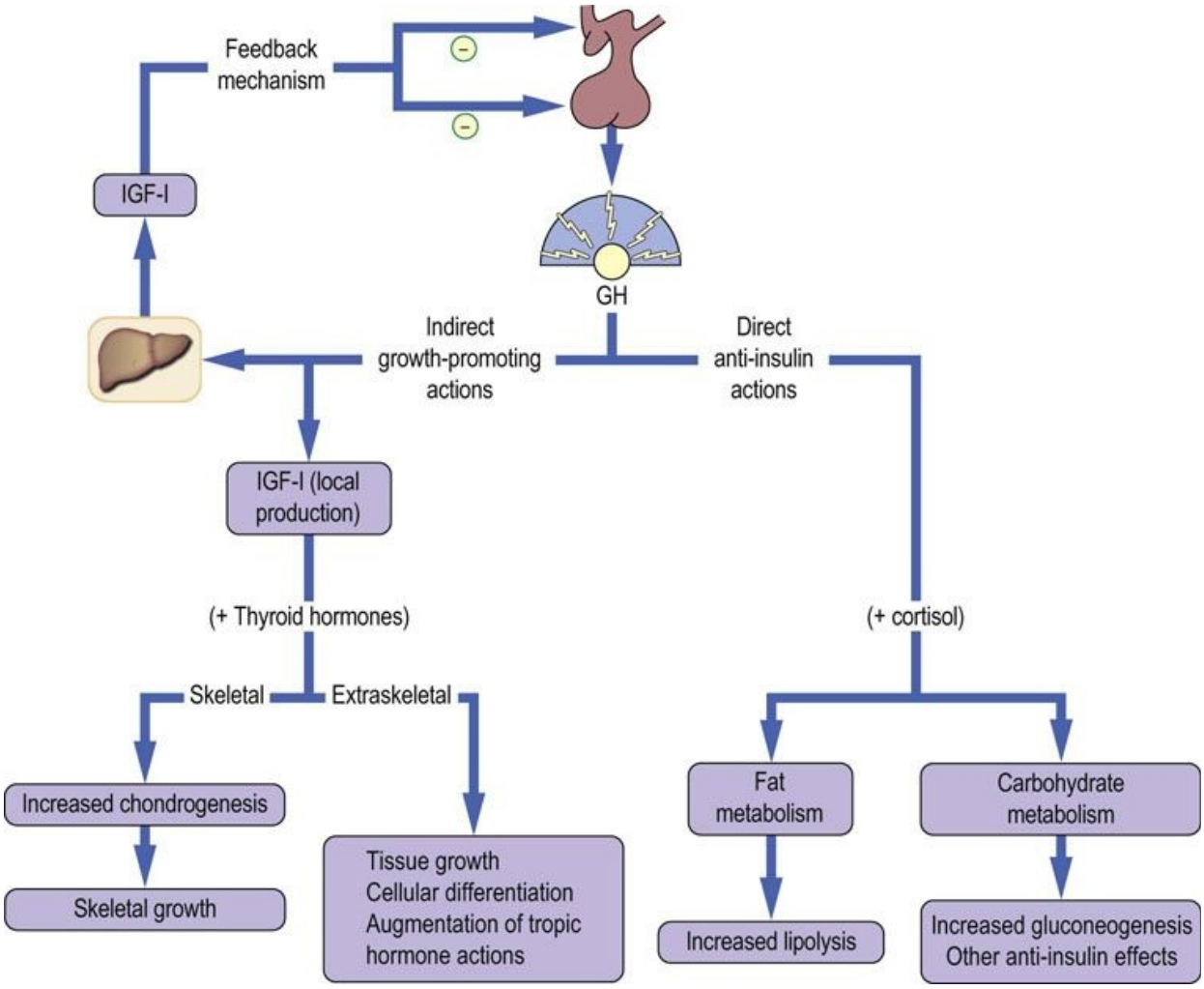


FIG. 39.12 Biochemical functions of the growth hormone (GH). These can be divided conveniently into direct actions on lipid and carbohydrate metabolism, and indirect actions on protein synthesis and cell proliferation.

Insulin-like growth factor I

IGF-I is the most GH-dependent of a series of growth factors

IGF-I is a 70-amino acid single-chain basic peptide, which has considerable homology with proinsulin. In response to tropic hormones like GH, IGF-I is produced in many tissues, whose growth is then stimulated; that is, IGF-I acts as a **paracrine hormone**. The liver is the major source of circulating IGF-I, whose function is primarily feedback inhibition of GH secretion. In plasma and other extracellular fluids, IGF-I is complexed to a series of **IGF-binding proteins** (IGFBPs) of which IGFBP-3 is the most abundant. IGF-I works through the **type 1 IGF receptor**, which is structurally similar to the insulin receptor and linked to intracellular tyrosine kinase activity. The relative affinities of insulin and IGF-I for their respective receptors mean that, in normal physiology, there is little cross-stimulation, although in pathologic and pharmacologic situations it is possible for insulin to have some action through the IGF-I receptor, and vice versa. This has implications for states of **severe insulin resistance**, where local activation of the IGF1 receptor has been suggested to underlie some clinical features such as skin overgrowth **(acanthosis nigricans)**. Conversely, IGF-I has a limited therapeutic role in some rare conditions caused by loss of insulin receptor function.

The plasma reference interval for IGF-I in adults aged 20–60 years is fairly constant. It is much lower in young children but rises dramatically during the period of growth and progression through puberty. IGF-I concentrations fall after the sixth decade of life. Thus, IGF-I appears to be a good marker of integrated GH activity.

Clinical test box Biochemical confirmation of pregnancy

Biochemical confirmation of pregnancy is now achieved by means of a simple and sensitive test that can give reliable results **within 2 weeks of fertilization** (i.e. before the next menstrual period). Pregnancy tests may be performed by doctors, nurses or even the patients themselves since they are readily purchased from pharmacies. Pregnancy tests are immunoassays that measure the production of **human chorionic gonadotropin (HCG)**. A specific two-site immunoassay is employed (see Fig. 39.2). One antibody recognizes the α-subunit of HCG, while the other only binds to the

32-amino acid segment of the β-subunit that is unique to HCG (Fig. 39.10). Since nonpregnant women do not normally produce HCG, the detection of even low levels from the first stages of placental development is sufficient to confirm pregnancy.

Clinical disorders of GH secretion

GH deficiency in children is one of several causes of short stature, while GH excess in children leads to gigantism and, in adults, to acromegaly

The absence of a plasma reference range for GH means that GH deficiency may only be diagnosed by studying the dynamics of GH secretion, either during sleep or following a stimulation test. **Basal IGF-I or IGFBP-3 measurements** may serve as a preliminary screening test. Treatment is by regular injection of recombinant human GH. Adults with a definite cause of GH deficiency (hypopituitarism) are also candidates for GH replacement therapy. A rare genetic cause of short stature is **Laron dwarfism** in which GH levels are elevated but IGF-I levels are subnormal – a GH receptor defect is responsible.

In most cases, GH excess is caused by a pituitary tumor

GH excess is most commonly due to a **GH-secreting pituitary tumor,** although GHRH-secreting tumors in the hypothalamus and ectopic GHRH production from pancreatic tumors have been described. In children, GH excess manifests itself as **gigantism**. In adults, the epiphyses of the long bones have closed and so further linear growth is not possible. Therefore, the adult form, **acromegaly**, is characterized by a thickening of tissues and overgrowth of the bones in the hands, feet and face (Fig. 39.13). Enlarged lips and tongue with lower jaw overgrowth give rise to a classic facial appearance. Excessive sweating and joint pains are also common symptoms. Unfortunately, the progression of acromegaly is so slow that an average of 9 years elapses between the onset of symptoms and diagnosis. Classically, the diagnosis of GH excess has relied on showing inadequate GH suppression during a standard 75 g oral glucose tolerance test and on elevated IGF-I levels. However, pubescent patients may have paradoxic

responses to glucose, and IGF-I levels are normally elevated during this developmental stage. Therefore evidence of a pituitary tumor obtained using **magnetic resonance imaging** (MRI) is critical in making the proper diagnosis. **Surgery** is the preferred treatment, although **long-acting somatostatin** is also effective. More recently, pegvisomant, a GH antagonist rationally designed based on structural features of the hormone–receptor interaction, has also proved useful in refractory cases.

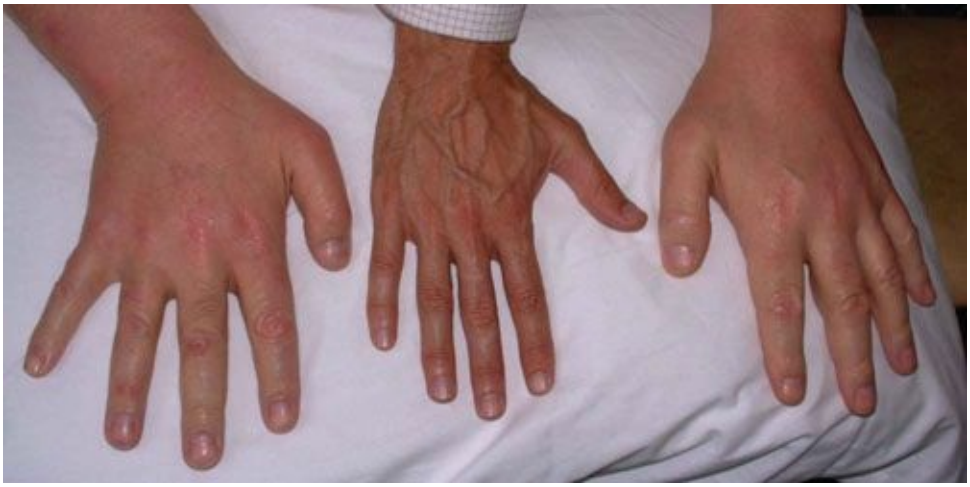


FIG. 39.13 Acromegaly.
Enlarged, somewhat square or ‘spadelike’ hands are shown compared to a normal hand in the middle.

The prolactin axis

Dopamine

Dopamine is an inhibitor of prolactin secretion

Prolactin is unique among the pituitary hormones in that it is under predominantly **inhibitory control** from the hypothalamus (Fig. 39.4). Furthermore, the controlling agent is the simple molecule **dopamine** (Chapter 41.1), which is produced by tuberoinfundibular dopamine neurons. Dopamine works by stimulating the pituitary lactotroph D2 receptor to inhibit adenylyl cyclase and consequently inhibits both prolactin synthesis and secretion. Several neuropeptides, including TRH, have prolactin-releasing properties, but there is little evidence for a physiologic role.

Prolactin

Prolactin may assist breast growth and milk formation, in association with other pregnancy-related hormones

Prolactin is a 23 kDa protein which is homologous to GH. The primary role for prolactin in humans occurs during **pregnancy** when prolactin binds to its receptor in mammary tissue and stimulates the synthesis of several milk proteins, including lactalbumin. In other animals, prolactin also has gonadotropic, immunologic, and hematologic effects: it is required for maintenance of pregnancy in rodents, is mitogenic for some immune cells, and it ameliorates some forms of anemia. Prolactin exerts its effects on female reproductive function at multiple levels. Its actions include blocking the action of FSH on follicular estrogen secretion and enhancing progesterone levels by inhibiting steroid-metabolizing enzymes.

Clinical test box Macroprolactin

Although antibody-based immunoassays are extremely powerful tools in endocrine diagnosis, they are subject to interference by endogenous antibodies in several ways. **One way in which endogenous antibodies can produce confusing results is by**

binding and aggregating the target hormone in high molecular weight (HMW) complexes. This may be seen for several peptide hormones, but is particularly problematic for prolactin, where so-called ‘macroprolactin’ may account for up to 10% of cases of apparently elevated serum prolactin. This is a problem, as macroprolactin is generally not biologically active, so its detection could lead to unnecessary treatment. There are several ways around this. Most simply, as different immunoassays often have different capabilities to detect macroprolactin, use of more than one assay may provide a strong clue to its presence. Second, a common technique is to precipitate HMW complexes with polyethylene glycol (PEG), with immunoassay both before and after to establish how much signal is lost in the HMW fraction. More accurate quantification may require laborious techniques such as gel filtration.

Clinical disorders of prolactin secretion

There are no known prolactin deficiency syndromes but hyperprolactinemia is very common

Hyperprolactinemia may result from a prolactin-secreting pituitary tumor **(prolactinoma)**, a deficient supply of dopamine from the hypothalamus, or the use of any of a wide range of antidopaminergic drugs. In women, the presenting features of hyperprolactinemia include **menstrual irregularity and galactorrhea** (discharge of milk from the breast).

A grossly elevated serum prolactin is usually diagnostic of a prolactinoma

For subjects with modest hyperprolactinemia, who are not taking antidopaminergic drugs, the diagnosis is more difficult; pituitary imaging and/or dynamic tests of prolactin secretion will assist the diagnosis of a microprolactinoma. Treatment options include **long-acting dopamine agonist drugs or surgery.** Prolactinomas need to be observed particularly carefully

during pregnancy, when there is physiologic hyperplasia of pituitary lactotrophs in anticipation of lactation.

In men, hyperprolactinemia can cause **impotence and prostatic hyperplasia**. However, because other conditions more commonly produce these symptoms, the diagnosis of hyperprolactinemia is often missed until prolactinomas become very large and present as hypopituitarism with visual field defects. The latter symptoms arise when the tumor expands out of the pituitary fossa and impinges on the optic chiasm. Such tumors often shrink and fibrose with dopamine agonist therapy. For this reason it is critically important to identify which patients with large pituitary tumors have very high prolactin levels, as they may be able to avoid surgery.

Endocrine systems not considered in this chapter

The human body contains several other endocrine systems not considered in this chapter, though they are governed by the same general principles. Some of these systems are considered in other chapters as part of the physiologic function that they control. Thus the reader is referred to Chapter 21 for **carbohydrate homeostasis**, Chapter 26 for **calcium homeostasis** and Chapter 24 for water and **electrolyte homeostasis** and the **control of blood pressure**. The intracellular signaling systems through which hormones exert their effects are described in Chapter 40.

Summary

■ **The endocrine system** produces hormones, a structurally diverse group of chemical messengers that regulate and coordinate whole-body metabolism, growth, reproduction, and responses to external stimuli.
■ **The hypothalamic-anterior pituitary axis** is a critical link between the brain and peripheral endocrine glands, and integrates diverse internal and external stimuli into only a few hormonal signals: it controls the synthesis and action of thyroid hormone, glucocorticoids, sex steroids, growth hormone, and prolactin.
■ **The hypothalamic-posterior pituitary axis** controls the production and release of oxytocin and AVP/ADH. The action of these hormones on cells is controlled by receptors located either on cell membranes or intracellularly.
■ **Feedback mechanisms** are important in controlling endocrine systems, and both overactivity and underactivity of these axes produce distinct clinical syndromes.
■ **Laboratory diagnosis of endocrine disorders** relies heavily on measurement of hormones in blood; however, careful attention must often be paid to the timing of sampling, and to the nature of the sample required, before meaningful interpretation of the results is possible.

Active Learning

1. Trace the two-way flow of information between hypothalamus and ovarian follicles early in the menstrual cycle.
2. How do GH, cortisol, and insulin interact to regulate lipid and carbohydrate metabolism?
3. Because dietary iodide can vary greatly, thyroid hormones need to be synthesized and stored when iodide is available, but thyroid hormones are hydrophobic and cannot be contained within membrane vesicles. How does the thyroid gland solve this problem?

Further reading

Anderson, MS. Autoimmune endocrine disease. *Curr Opin Immunol*. 2002; 14:760–764.

Cooper, DS. Hyperthyroidism. *Lancet*. 2003; 363:459–468.

Cooper, DS, Biondi, B. Subclinical thyroid disease. *Lancet*. 2012; 379:1142–1154.

Dattani, M, Preece, M. Growth hormone deficiency and related disorders: insights into causation, diagnosis and treatment. *Lancet*. 2004; 363:1977–1987.

Dayan, CM. Interpretation of thyroid function tests. *Lancet*. 2001; 357:619–624.

Franklyn, JA, Boelaer, K. Thyrotoxicosis. *Lancet*. 2012; 379:1155–1166.

Henderson, J. Ernest Starling and 'hormones': an historical commentary. *J Endocrinol*. 2005; 184:5–10.

Løvås, K, Husebye, ES. Addison's disease. *Lancet*. 2005; 365:2058–2061.

Melmed, S. Acromegaly. *N Engl J Med*. 2006; 355:2558–2573.

Newell Price, J, Bertagna, X, Grossman, A, *et al*. Cushing's syndrome. *Lancet*. 2006; 367:1605–1617.

Roberts, CGP, Ladenson, PW. Hypothyroidism. *Lancet*. 2004; 362:793–803.

Semple, RK, Topaloglu, AK. The recent genetics of hypogonadotrophic hypogonadism – novel insights and new questions. *Clin Endocrinol (Oxf)*. 2010; 72:427–435.

Plant, TM. Hypothalamic control of the pituitary-gonadal axis in higher primates: key advances over the last two decades. *J Neuroendocrinol*. 2008; 20:719–726.

Schlechte, JA. Prolactinoma. *N Eng J Med*. 2003; 349:2035–2041.

Schneider, HJ, Aimaretti, G, Kreitschmann-Andermahr, I, *et al*. Hypopituitarism. *Lancet*. 2007; 369:1461–1470.

Zimmermann, MB, Jooste, PL, Pandav, CS. Iodine-deficiency disorders. *Lancet*. 2008; 372:1152–1168.

Websites and downloads

The Endocrine Societ. www.endo-society.org

Endotext.otg. The endocrine source. www.endotext.org.

Thyroid disease manager. www.thyroidmanager.org.

Pituitary Network Associatio. www.pituitary.org

Cushing's Support and Research Associatio. www.CSRF.net

National Endocrine and Metabolic Diseases Information Servic. www.endocrine.niddk.nih.gov/

CHAPTER 40

Membrane Receptors and Signal Transduction

Ian P. Salt, Margaret M. Harnett and Helen S. Goodridge

Learning objectives

After reading this chapter you should be able to:

- Distinguish between steroid and polypeptide hormones, and outline their mechanisms of action.
- Describe G-protein coupled receptors.
- Outline the activation of downstream intracellular signaling cascades by heterotrimeric G-proteins.
- Discuss the generation of second messengers such as cyclic AMP, inositol trisphosphate (IP_3), diacylglycerol (DAG) and Ca^{2+}, and explain how they activate key protein kinases.
- Explain how phospholipases generate a diverse array of lipid second messengers.
- Discuss how the generation of a variety of second messengers can amplify signals and lead to the generation of specific biological responses.

Introduction

Cellular signals are processed by cassettes comprising specific receptors-effector elements and regulatory proteins

Cells sense, respond to, and integrate a multiplicity of signals from their environment. These signals can be hormones produced elsewhere from their site of action **(endocrine signaling);** signals generated locally to the target cell **(paracrine signaling);** signals from cells in physical contact with the target cell **(juxtacrine signaling);** or signals generated by the target cell itself **(autocrine signaling).** Hydrophobic signals and small molecules can cross the plasma membrane and exert their effects via receptors within the cell, whereas the majority of signals are hydrophilic, are not able to cross the lipid plasma membrane and require specific cell surface membrane receptors. In either case, signals are sensed and processed by **cellular signal transduction cassettes** that comprise specific receptors, effector signaling elements, and regulatory proteins. These signaling cassettes serve to detect, amplify, and integrate the diverse external signals to generate the appropriate cellular response (Fig. 40.1). Ultimately, signals can rapidly alter cellular processes such as exocytosis and metabolism, or alter transcription factor activity, leading to changes in target gene expression.

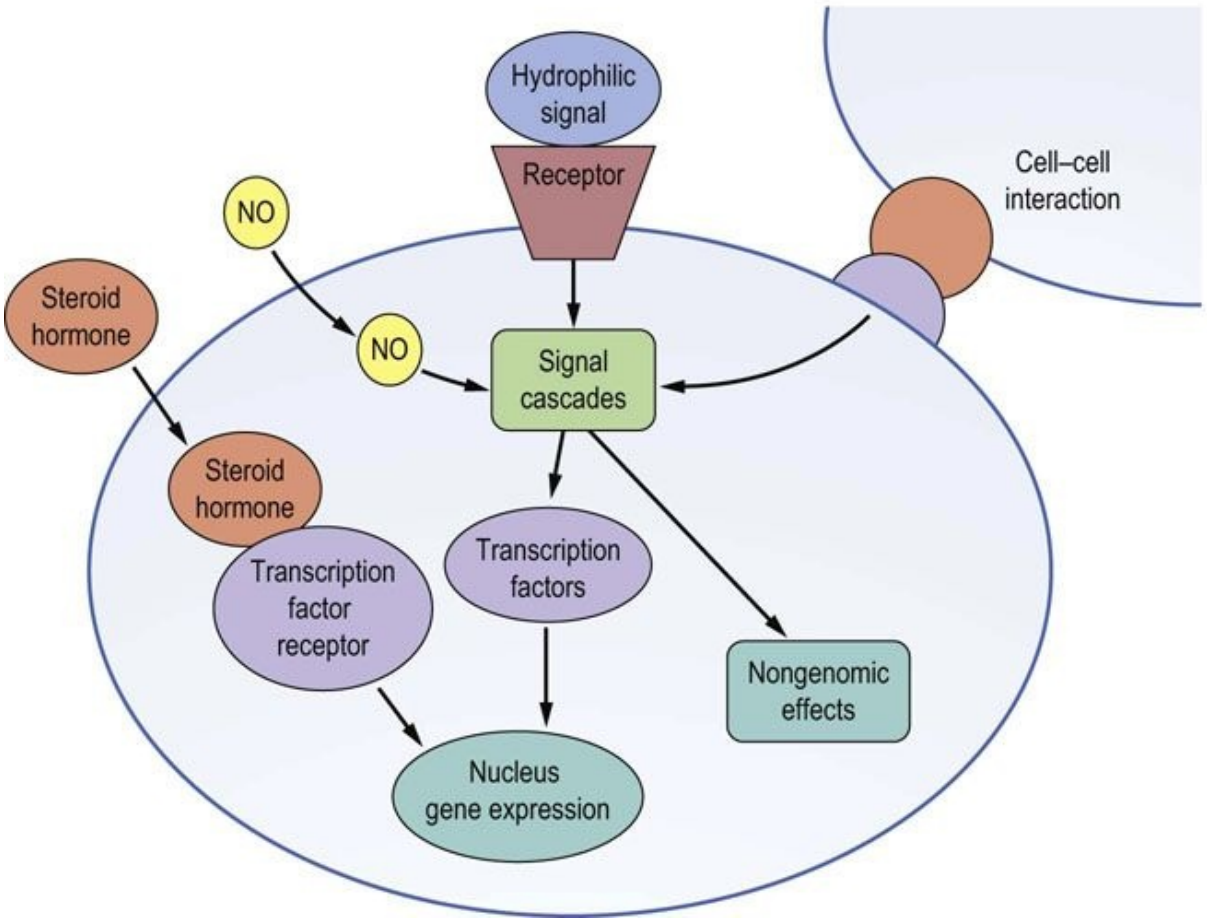


FIG. 40. 1 Mechanisms of cell signaling.
NO, nitric oxide.

In this chapter, we first discuss how cell surface receptors sense and transduce their specific signal by transmembrane coupling to effector enzyme systems, generating low-molecular-mass molecules termed second messengers. We then discuss the diversity of these second messengers and how they influence the activity of a range of key protein kinases with distinct substrates that ultimately determine the type of obtained biological response.

Types of hormone and monoamine receptors

Hormone receptors for steroid hormones are different from those for polypeptide hormones and monoamines

Hormones are biochemical messengers that act to orchestrate the responses of different cells within a multicellular organism (Chapter 39). They are generally synthesized by specific tissues and secreted directly into the blood, which transports them to their target responsive organs. Hormone signaling can broadly be subdivided into two major classes:

- steroid hormone signaling
- polypeptide and monoamine hormone signaling.

Hormones achieve their biological effects by interacting with specific receptors to induce intracellular signaling cascades (Table 40.1).

Table 40.1

Classification of membrane receptors

Receptor class	Transmembrane-spanning domains	Intrinsic catalytic activity	Accessory coupling/ regulatory molecules	Examples of receptor classes/ ligands
G-protein coupled receptors (serpentine receptors)	Multipass (seven transmembrane α-helices)	None	G-proteins	Glucagon α-Adrenergic, β-adrenergic (epinephrine) Muscarinic (acetylcholine) rhodopsin (vision) Chemokines (IL-8)
Ion channel receptors (ligand-gated channels)	Multipass; generally form multimeric complexes	Ion channel activity	None	Neurotransmitters Ions Nucleotides Inositol trisphosphate (IP_3)
Intrinsic tyrosine kinase receptors	Single-pass transmembrane domain, but may be multimeric (e.g. insulin receptor)	Tyrosine kinase	None	Insulin Peptide growth factors (e.g. PDGF, FGF, NGF, EGF)
Tyrosine kinase-associated receptors	Single-pass transmembrane domain, but generally form multimeric receptors	None	Some require ITAM/ ITIM-containing proteins	Antigen receptors (ITAM-Src-related kinases) FcγR (ITIM–Src-related kinases) Leptin, IL-6 (Janus kinases)
Intrinsic tyrosine phosphatase receptors	Single-pass transmembrane domain	Tyrosine phosphatase	None	CD45-phosphatase receptor
Intrinsic serine/threonine kinase receptors	Single-pass transmembrane domain	Serine/threonine kinase	None	Tumor growth factor-β (TGF-β)
Intrinsic guanylate cyclase receptors	Single-pass transmembrane domain	Guanylyl cyclase (generates cGMP)	None	Atrial natriuretic protein (ANP)
Death domain receptors	Single-pass transmembrane domain	None	Death domain accessory proteins (TRADD, FADD, RIP, TRAFs)	Tumor necrosis factor-α (TNF-α) Fas

cGMP, cyclic guanosine monophosphate; FADD, Fas-associated death domain; FcγR, Fc-γ receptor (receptor for immunoglobulin G); IL, interleukin; ITAM/ITIM, immunoreceptor tyrosine activation/inhibition motif; RIP, receptor-interacting protein; Src, Src-tyrosine kinase; TRADD, TNF-receptor-associated death domain; TRAFs, TNF-receptor-associated factors. PDGF, platelet-derived growth factor; FGF, fibroblast growth factor; EGF, epidermal growth factor; NGF, nerve growth factor FGF, fibroblast growth factor; EGF, epidermal growth factor; NGF, nerve growth factor

Steroid hormones traverse cell membranes

Because of the cholesterol-based nature of their structure, steroid hormones, such as glucocorticoids, mineralocorticoids, sex hormones, and vitamin D, can traverse the plasma membrane of cells to initiate their responses via cytoplasmic located receptors called steroid hormone receptors (Fig. 40.1). These receptors belong to a superfamily of cytoplasmic receptors called the intracellular receptor superfamily, which also transduce signals from other small hydrophobic signaling molecules such as the tyrosine-derived thyroid hormones (e.g. thyroxine) and the vitamin A-derived retinoids (e.g. retinoic acid).

Intracellular receptors for steroid and thyroid hormones, and retinoids, are transcription factors

The intracellular receptors for these steroid and thyroid hormones, and retinoids, are **transcription factors;** they bind to regulatory regions of the DNA of genes that are responsive to the particular steroid/thyroid hormone. Such 'ligand binding' (ligation) induces a conformational change in the transcription factor that allows it to activate or repress gene induction. Although all the target cells have specific receptors for the individual hormones, they express distinct combinations of cell type-specific regulatory proteins that cooperate with the intracellular hormone receptor to dictate the precise repertoire of genes that are induced. Hence the hormones induce distinct sets of responses in different target cells (Chapter 35).

Polypeptide hormones act through membrane receptors

In contrast to the steroid hormones, polypeptide hormones cannot cross cell membranes and must initiate their effects on their target cells via specific cell **surface receptors** (Fig. 40.1). As they do not themselves enter the target cell, they are sometimes termed **'first messengers'.** Binding to the specific cell surface receptor causes a conformational change in that receptor that can engage **signaling cascades** in several different ways. Receptor binding can:

- Regulate the production of low-molecular-weight signaling molecules, such as cyclic adenosine monophosphate (cAMP) or calcium, which are called **'second messengers'**
- Alter the intrinsic catalytic activity of the receptor.
- Alter the recruitment of regulatory molecules to the receptor (Table 40.1).

Other molecules that signal through membrane receptors

In addition to polypeptide hormones, a wide range of signaling molecules use transmembrane signal transduction cassettes to elicit their biological effects. Such signals include polypeptide growth factors, polypeptide signals that mediate inflammation and immunity (cytokines and chemokines) and small hydrophilic molecules (such as acetylcholine, purines, nucleotides or inositol trisphosphate) (Table 40.1).

Some low-molecular-mass signaling molecules traverse the cell membrane

Although most extracellular signals mediate their effects via receptor–ligand interaction of either cell surface or cytoplasmic receptors, some low-molecular-

mass signaling molecules are able to traverse the plasma membrane and directly modulate the activity of the catalytic domains of transmembrane receptors or cytoplasmic signal transducing enzymes (Fig. 40.1). For example, nitric oxide (NO), which has a variety of functions including the relaxation of smooth muscle cells in blood vessels, can stimulate guanylyl cyclase, leading to the generation of the second messenger, cGMP. Patients with **angina pectoris** are treated with glyceryl trinitrate (GTN), which is converted to NO, resulting in relaxation of blood vessels delivering oxygen and nutrients to the heart. The consequent improvement in oxygen delivery to the heart muscle eases the pain that was caused by inadequate blood flow to the heart.

Receptor coupling to intracellular signal transduction

Membrane receptors couple to signaling pathways utilizing diverse mechanisms

Some membrane receptors, for example the β-adrenergic receptors or the antigen receptors on lymphocytes, have no intrinsic catalytic activity and serve simply as specific recognition units. These receptors use a variety of mechanisms, including adaptor molecules or catalytically active regulatory molecules such as G-proteins (guanosine triphosphatases, GTPases, which hydrolyze GTP) to couple them to their effector signaling elements, which are generally enzymes (often called signaling enzymes or signal transducers) or ion channels (Fig. 40.2). In contrast, other receptors – such as the intrinsic tyrosine kinase receptors for insulin and many growth factors, or the intrinsic serine kinase receptors for molecules like transforming growth factor-β (TGF-β) – have extracellular ligand-binding domains and cytoplasmic catalytic domains. After receptor ligation, these receptors can directly initiate their signaling cascades by phosphorylating and modulating the activities of target signal-transducing molecules (downstream signaling enzymes). These in turn propagate the growth factor signal by modulating the activity of further specific signal transducers or transcription factors, leading to gene induction (Chapter 35). Furthermore, sensory systems such as vision (Chapter 41.2), taste and smell use similar mechanisms of cell surface membrane receptor-coupled signal transduction (Table 40.1).

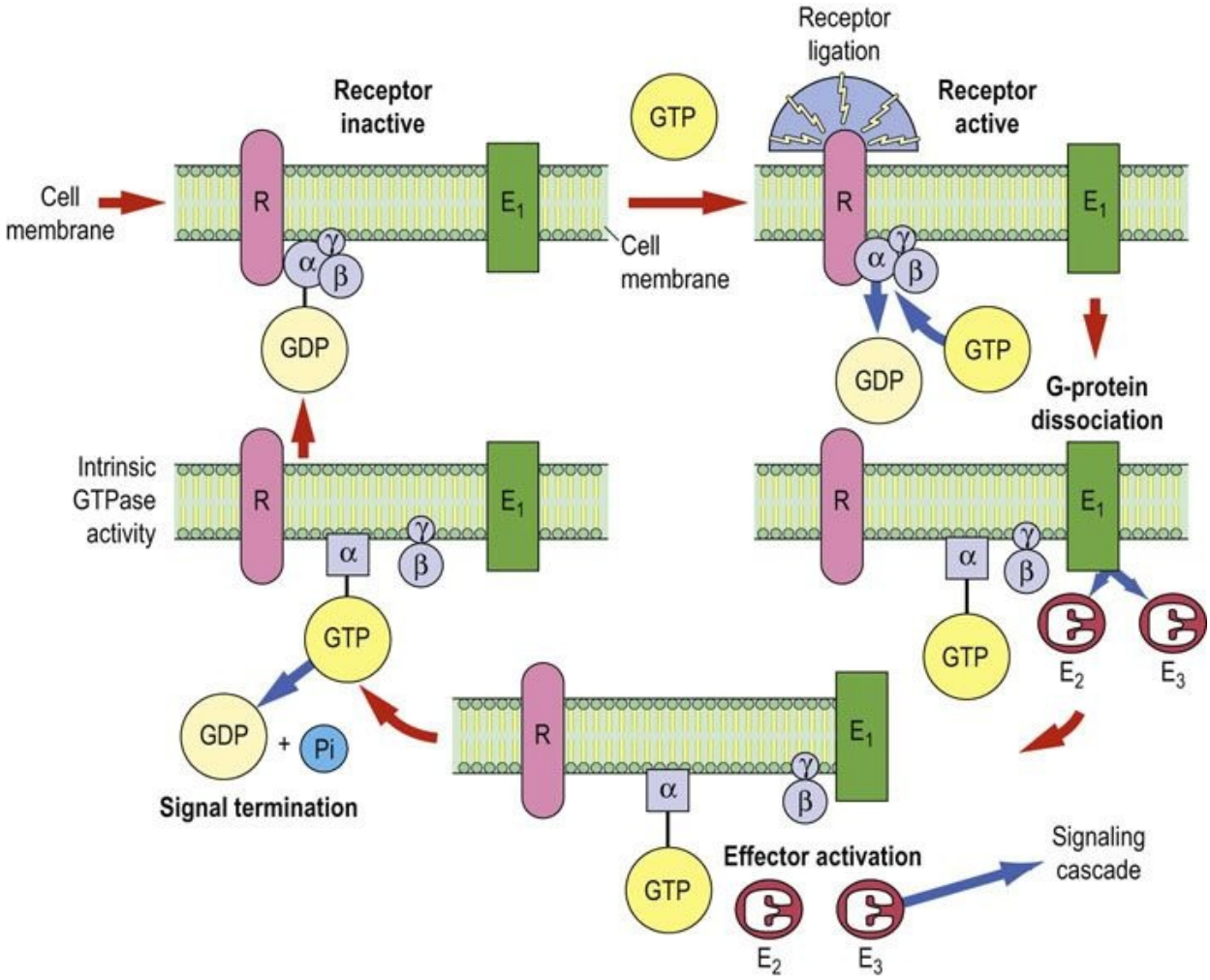


FIG. 40.2 Mechanism of G-protein signaling.
In the inactive state, G-proteins exist as heterotrimers with GDP bound tightly to the α-subunit. None of the subunits are integral membrane proteins; however, the G-protein is anchored to the plasma membrane by lipid modification of the γ-subunits (prenylation) and some of the α-subunits (myristoylation in the $G_{i\alpha}$ family). Ligation of the receptor (R) drives exchange of GDP for GTP and induces a conformational change in Gα, which results in a decrease in its affinity both for the receptor and for the βγ-subunits, leading to dissociation of the receptor–G-protein complex. The activated Gα (GTP-bound) or the released βγ-subunits, or both, can then interact with one or more effectors, to generate intracellular second messengers that activate downstream signaling cascades. Signaling is terminated by the intrinsic GTPase activity of the α-subunit, which hydrolyzes GTP to GDP to allow reassociation of the inactive, heterotrimeric G-protein, Gαβγ.

Some receptors possess intrinsic protein kinase activity

Ligand binding to many growth factor receptors stimulates protein kinase activity of an intracellular domain of the receptor complex. The activated

receptor subsequently phosphorylates substrate proteins, in which the γ-phosphate from ATP is transferred to the side-chain hydroxyl groups on serine, threonine or tyrosine residues. All receptor protein kinases are either specific serine/threonine kinases or tyrosine kinases, but never both. Furthermore, protein kinases phosphorylate substrates at specific serine, threonine or tyrosine residues, depending on the sequence surrounding the site of phosphorylation. Upon ligand binding, receptor protein kinases autophosphorylate (self-phosphorylate). The introduction of the bulky charged phosphate moiety during autophosphorylation or onto other substrate proteins markedly alters the conformation of the protein, leading to changes in activity or acting as docking sites for other (adapter) proteins. Adapter proteins contain specific domains that recognize and bind to the phosphorylated proteins. As receptor protein kinases often phosphorylate at several sites, this can lead to several different adapter proteins being recruited to the activated receptor complex. Subsequently these adapter proteins can engage several different signaling pathways.

The example of insulin signaling

An example of this is insulin signaling. Insulin binding to the insulin receptor (IR) causes activation and autophosphorylation of the intracellular tyrosine kinase domains. Adapter proteins including the insulin receptor substrate (IRS) proteins and Shc (Src-homology and collagen-like) proteins bind the phosphotyrosine residues on the IR. Next, IRSs are themselves tyrosine phosphorylated by the IR on tyrosine residues, generating docking sites for the lipid kinase phosphatidylinositol 3′-kinase (PI3K), which generates phosphatidylinositol 3,4,5-trisphosphate (PIP_3) at the plasma membrane. Newly formed PIP_3 recruits the serine/threonine protein kinase Akt to the plasma membrane, and Akt is subsequently phosphorylated and activated by other protein kinases (Fig. 40.3). Activation of Akt is the key signaling pathway by which insulin exerts the majority of its metabolic effects, including stimulation of glucose transport and suppression of gluconeogenesis (Chapter 21). Binding of Shc to the activated IR leads to recruitment of growth factor receptor-bound protein (Grb2) to Shc, which subsequently activates a guanine nucleotide exchange factor (SOS), which stimulates the small G-protein Ras, which initiates the a protein kinase cascade in which several protein kinases phosphorylate each other in turn. Signaling through this pathway is associated with the mitogenic, growth-promoting actions of insulin. Therefore ligand binding can initiate multiple signaling pathways with different cellular effects.

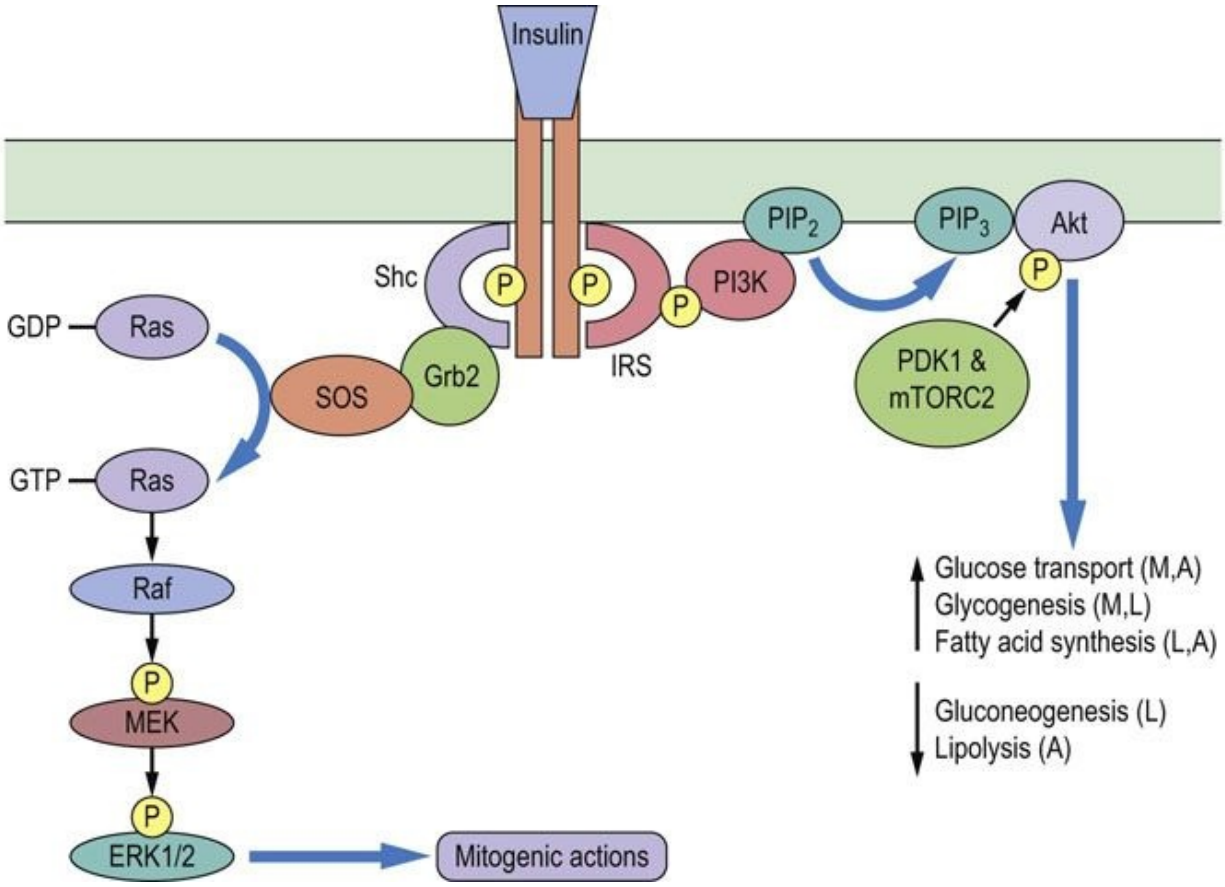


FIG. 40.3 Insulin signaling pathways.
Insulin binding to the dimeric insulin receptor tyrosine kinase stimulates autophosphorylation. The insulin receptor substrate (IRS) and Src-homology and collagen-like (Shc) adapter proteins bind to the phosphotyrosines on the insulin receptor. IRSs are subsequently phosphorylated by the insulin receptor, generating docking sites for phosphatidylinositol 3'-kinase (PI3K), which generates the phosphorylated lipid signaling molecule phosphatidylinositol 3,4,5-trisphosphate (PIP_3) from phosphatidylinositol 4,5-bisphosphate (PIP_2). PIP_3 recruits the serine/threonine protein kinase Akt to the plasma membrane, where Akt is phosphorylated and activated by phosphoinositide-dependent kinase 1 (PDK1) and the mammalian target of rapamycin complex 2 (mTORC2). Akt is essential for the metabolic effects of insulin in muscle (M), liver (L) and adipose tissue (A). Insulin receptor-bound Shc recruits growth factor receptor-bound protein 2 (Grb2), which is bound to the guanine nucleotide exchange factor Son of Sevenless (SOS). SOS catalyzes GDP–GTP exchange on the small G-protein Ras. The GTP-bound active Ras initiates a protein kinase cascade in which the protein kinase Raf phosphorylates and activates another protein kinase MEK, which subsequently phosphorylates and activates the protein kinases ERK1 and ERK2, which mediate many of the mitogenic actions of insulin.

Some membrane receptors are coupled to G-proteins

G-protein coupled receptors comprise a superfamily of structurally related receptors for hormones, neurotransmitters, inflammatory mediators, proteinases, taste and odorant molecules, and light photons. A classic example of this class of receptors is the **β-adrenergic receptor** (for which the ligand is epinephrine) as its structure–function properties have been extensively studied with respect to its activation of signal transduction cascades. G-protein coupled receptors are integral membrane proteins characterized by the seven transmembrane-spanning helices within their structure. They generally comprise an extracellular *N*-terminus, seven transmembrane-spanning α-helices (20–28 hydrophobic amino acids each), three extracellular and intracellular loops, and an intracellular *C*-terminal tail. Ligands, such as epinephrine, typically bind to the G-protein coupled receptor by sitting in a pocket formed by the transmembrane helices (Fig. 13.5). G-protein coupled receptors have no intrinsic catalytic domains; upon activation they recruit G-proteins via their third cytoplasmic loop to couple to their signal transduction elements. G-protein coupled receptors are often the target of drugs; indeed it has been estimated that approximately 30% of all currently available therapeutics act on G-protein coupled receptors. Furthermore, using the information gained by sequencing the human genome, it is apparent that there are many additional members of the G-protein coupled receptor family for which the signal has not yet been identified.

G-proteins regulate a diverse range of biological processes

G-proteins constitute a group of regulatory molecules that are involved in the regulation of a diverse range of biological processes, including signal transduction, protein synthesis, intracellular trafficking (targeted delivery to the plasma membrane or intracellular organelles) and exocytosis, as well as cell movement, growth, proliferation and differentiation. The G-protein superfamily predominantly comprises two major subfamilies: the small, monomeric *Ras*-like G-proteins (Chapter 42) and the heterotrimeric G-proteins. Heterotrimeric G-proteins regulate the transduction of transmembrane signals from cell surface receptors to a variety of intracellular effectors, such as adenylyl cyclase, phospholipase C (PLC), cGMP-phosphodiesterase (PDE) and ion channel effector systems (Fig. 40.2). Heterotrimeric G-proteins consist of 3 subunits: α (39–46 kDa), β (37 kDa), and γ (8 kDa). In general, effector specificity is conferred by the α-subunit, which contains the GTP-binding site and an intrinsic GTPase activity. However, it is now widely accepted that βγ-complexes can also directly regulate effectors such as phospholipase A_2 (PLA_2), PLC-β isoforms,

adenylyl cyclase and ion channels in mammalian systems, as well as cellular responses such as mating factor receptor pathways in yeast. Four major subfamilies of α-subunit genes have been identified on the basis of their cDNA homology and function: $G_s\alpha$, $G_i\alpha$, $G_{q/11}\alpha$, and $G_{12/13}\alpha$ (Table 40.2). Many of these Gα subunits have been shown to exhibit a rather ubiquitous pattern of expression in mammalian systems, at least at the mRNA level, but it is also clear that certain α-subunits have a tissue-restricted profile of expression. Moreover, there is evidence of differential expression of α-subunits during cellular development.

Table 40.2

Properties of mammalian G-protein α-subunits

G-protein subfamily	α subunit	Tissue distribution	Toxin substrate	Examples of effectors
$G_s\alpha$	$G_s\alpha$	Ubiquitous	Cholera toxin	Activates adenylyl cyclase(G_s, $G_{olf}\alpha$) K^+ channels ($G_s\alpha$) Src tyrosine kinases ($G_s\alpha$)
	$G_{olf}\alpha$	Olfactory neurons, central nervous system	Cholera toxin	
$G_i\alpha$	$G_i\alpha$	Ubiquitous	Pertussis toxin	Inhibits adenylyl cyclase (G_i,G_o,$G_z\alpha$) Activates K^+ channels (G_i,G_o,$G_z\alpha$) cGMP phosphodiesterase ($G_j\alpha$),
	$G_o\alpha$	Neuronal/neuroendocrine tissues	Pertussis toxin	
	$G_z\alpha$	Neurons, platelets	None	
	$G_t\alpha$	Retina	Pertussis toxin	
	$G_{gust}\alpha$	Taste buds	Pertussis toxin	
$G_{q/11}\alpha$	$G_q\alpha$	Ubiquitous	None	Activates PLC indirectly through activation of calcium channels Activates K^+ channels ($G_q\alpha$)
	$G_{11}\alpha$	Ubiquitous	None	
	$G_{14}\alpha$	Lung, kidney, liver, spleen, testis	None	
	$G_{16}\alpha$ ($G_{15}\alpha$ in mouse)	Hematopoietic cells	None	
$G_{12/13}\alpha$	$G_{12}\alpha$	Ubiquitous	None	Activates PLCε indirectly through activation of calcium channels Activates PLD Activates Rho GEF
	$G_{13}\alpha$	Ubiquitous	None	

cGMP, cyclic guanosine monophosphate; PLC, phospholipase C; PLD, phospholipase D; Rho GEF, Rho GTPase guanine nucleotide exchange factor.

G-proteins act as molecular switches

Heterotrimeric G-proteins regulate transmembrane signals by acting as molecular switches, linking cell surface G-protein coupled receptors to one or more downstream signaling molecules (Fig. 40.2). Ligation of the receptor initiates an interaction with the inactive, GDP-bound heterotrimeric G-protein. This interaction drives exchange of GDP for GTP, inducing a conformational change in Gα, which results in a decrease in its affinity for both the receptor and the βγ-subunits, leading to dissociation of the receptor–G-protein complex. The

activated Gα (GTP-bound) or released βγ-subunits, or both, can then interact with one or more effectors to generate intracellular second messengers, which activate downstream signaling cascades. Signaling is terminated by the intrinsic GTPase activity of the α-subunit, which hydrolyzes GTP to GDP to allow reassociation of the inactive heterotrimeric G-protein (Gαβγ).

Second messengers

Advanced concept box Bacterial toxins which target G-proteins cause several diseases

A variety of bacterial toxins exert their toxic effects by covalently modifying G-proteins and hence irreversibly modulating their function. For example, **cholera toxin** from *Vibrio cholerae* contains an enzyme (subunit A) that catalyzes the transfer of ADP-ribose from intracellular NAD^+ to the α-subunit of G_s; this modification prevents the hydrolysis of G_s-bound GTP, resulting in a constitutively (permanently) active form of the G-protein. The resulting prolonged increase in cAMP concentrations within the intestinal epithelial cells leads to PKA-mediated phosphorylation of Cl^- channels, causing a large efflux of electrolytes and water into the gut, which is responsible for the severe diarrhea that is characteristic of cholera. Enterotoxin action is initiated by specific binding of the B (binding) subunits of the cholera toxin (AB_5) to the oligosaccharide moiety of the monosialoganglioside, GM_1, on epithelial cells. A similar molecular mechanism has been attributed to the action of the heat-**labile enterotoxin**, labile toxin secreted by several strains of *Escherichia coli* responsible for 'traveler's diarrhea'.

In contrast, **pertussis toxin** (another AB_5 toxin) from *Bordetella pertussis*, the causative agent of whooping cough, catalyzes the ADP-ribosylation of $G_{i\alpha}$, which prevents $G_{i\alpha}$ from interacting with activated receptors. Hence, the G-protein is inactivated and cannot act to inhibit adenylyl cyclase, activate PLA_2 or PLC, open K^+ channels, or open and close Ca^{2+} channels, causing a generalized uncoupling of hormone receptors from their signaling cascades.

Cyclic AMP (cAMP) is a key molecule in signal transduction

cAMP is a small molecule that has a key role in the regulation of intracellular signal transduction cAMP is derived from ATP by the catalytic action of the signaling enzyme **adenylyl cyclase** (Fig. 40.4). This cyclization reaction involves the intramolecular attack of the 3′-OH group of the ribose unit on the *α*-phosphoryl group of ATP to form a phosphodiester bond; this is driven by the subsequent hydrolysis of the released pyrophosphate. The activity of cAMP is terminated by the hydrolysis of cAMP to 5′-AMP by specific cAMP phosphodiesterases.

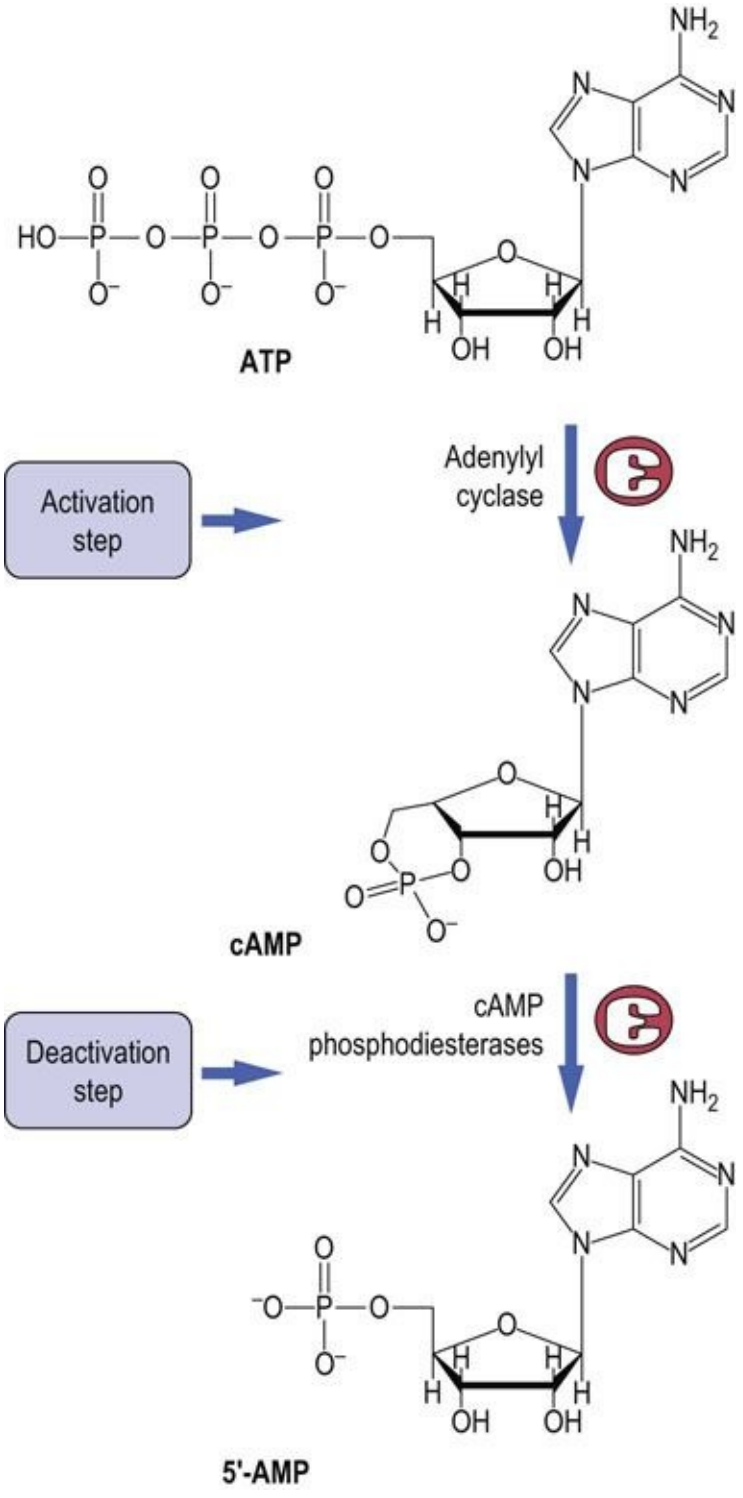


FIG. 40.4 Metabolism of cyclic AMP.
Adenylyl cyclase catalyzes a cyclization reaction to produce the active cAMP, which is then deactivated by cAMP phosphodiesterases. cAMP, cyclic adenosine monophosphate.

Glucagon and β-adrenergic receptors are coupled to cAMP

Glucagon and β-adrenergic receptors are coupled to the generation of cAMP. The β-adrenergic hormone epinephrine induces the breakdown of glycogen to glucose-1-phosphate in muscle and, to a lesser extent, in the liver. The breakdown of glycogen in the liver is predominantly stimulated by the polypeptide hormone glucagon, which is secreted by the pancreas when plasma glucose is low (Chapters 13 and 21). One of the earliest signaling events after binding of these hormones to their receptors is the generation of cAMP. The importance of cAMP in regulating glycogen breakdown was demonstrated by a series of experiments showing not only that hormones that activate adenylyl cyclase activity in liver and muscle cells, also stimulate glycogen breakdown, but also demonstrated that cell-permeant analogues of cAMP, such as dibutyryl cAMP, can mimic the effects of these hormones in inducing glycogen breakdown.

Adenylyl cyclase is regulated by G-protein α-subunits

The β-adrenergic and glucagon receptors are coupled to adenylyl cyclase activation by the action of a specific form of the α-subunit of the G-protein, termed $G_s\alpha$. Although hydrolysis of GTP by the intrinsic GTPase of the $G_s\alpha$-subunit acts to switch off adenylyl cyclase activation, the hormone–receptor complex must also be deactivated to return the cell to its resting, unstimulated state. In the case of β-adrenergic receptors, this receptor desensitization, which occurs after prolonged exposure to the hormone, involves phosphorylation of the C-terminal tail of the hormone-occupied β-adrenergic receptor by a kinase known as β-adrenergic receptor kinase. Other G-protein coupled receptors, such as α_2-adrenergic receptors in smooth muscle, act to inhibit adenylyl cyclase and cAMP generation. In this case, receptors are coupled to a specific inhibitory form of the α-subunit of the G-protein, termed $G_i\alpha$ (Table 40.2), which inhibits adenylyl cyclase activity, reducing cAMP concentrations.

Signals can activate different receptor subtypes, with different consequences

Receptor subtypes are expressed in a tissue-specific manner for some signals such as epinephrine and angiotensin II. These different receptor subtypes may couple in different ways; for example, epinephrine stimulates cAMP synthesis through $G_s\alpha$-coupled β-adrenergic receptors in skeletal muscle. whereas in

smooth muscle, epinephrine inhibits cAMP synthesis through $G_i\alpha$-coupled α_2-adrenergic receptors. Therefore the same signal can have differing effects on intracellular signaling cascades dependent on the tissue being examined.

Protein kinase A

cAMP transduces its effects on glycogen–glucose-1-phosphate interconversion by regulating a key signaling enzyme, protein kinase A (PKA), which phosphorylates target proteins on serine and threonine residues.

Protein kinase A binds cAMP and phosphorylates other enzymes

PKA is a multimeric enzyme comprising two regulatory (R) subunits and two catalytic (C) subunits: the R_2C_2 tetrameric form of PKA is inactive, but binding of four molecules of cAMP to the R-subunits leads to the release of catalytically active C-subunits, which can then phosphorylate and modulate the activity of two key enzymes, phosphorylase kinase and glycogen synthase (Fig. 40.5), which are involved in regulation of glycogen metabolism (Chapter 13).

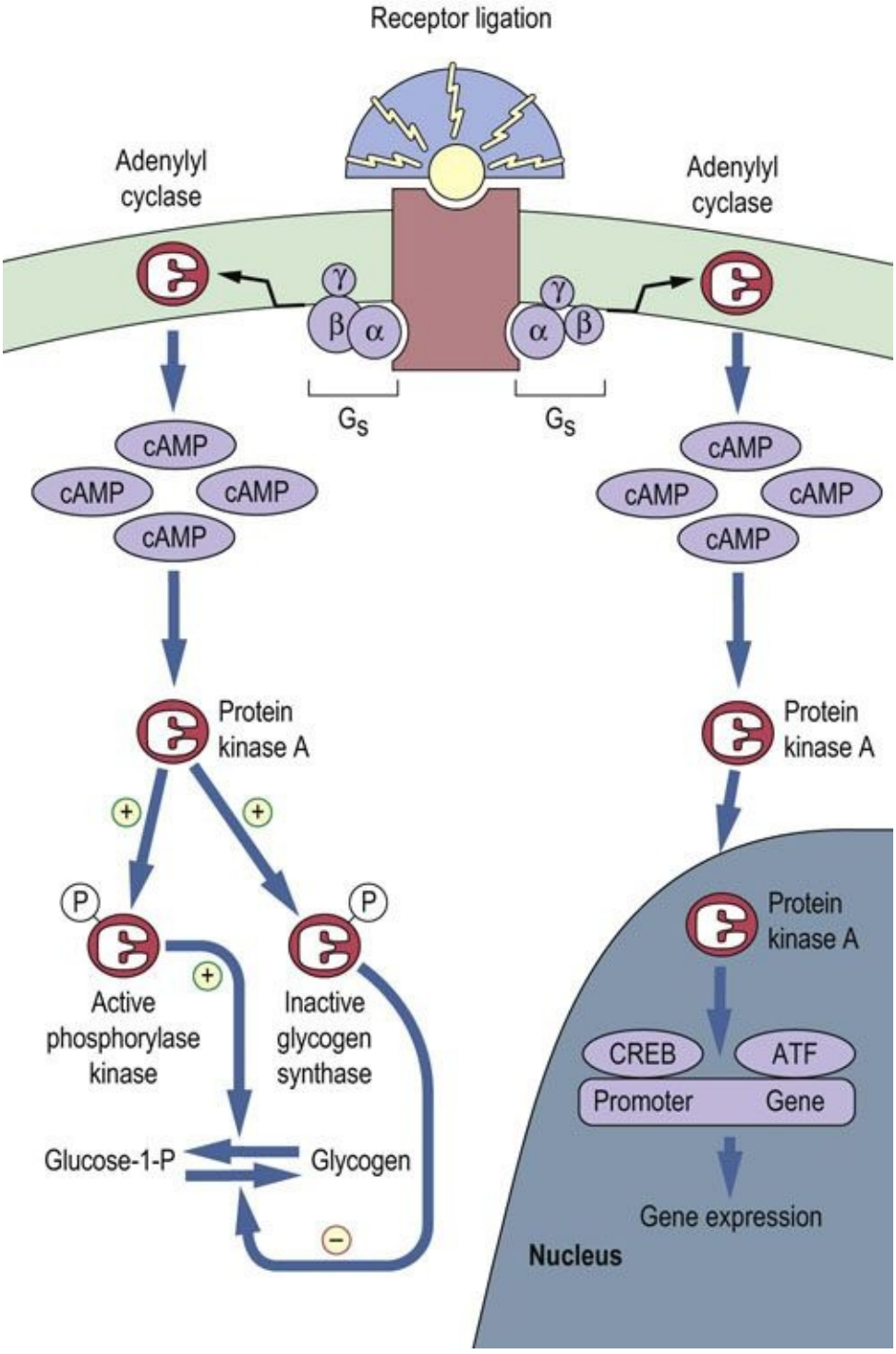


FIG. 40.5 Protein kinase A (PKA) acts as a signaling enzyme for the second messenger, cAMP.
Binding of a stimulatory G-protein (G_s) to the hormone–receptor complex activates adenylyl cyclase, which catalyzes the production of cAMP. PKA is activated by binding four molecules of cAMP. Translocation of PKA into the nucleus modulates activity of

transcription factors, for example CREB and ATF (see text), leading to induction or repression of gene expression. (Compare Fig. 13.6.)

Many other cellular responses can be mediated by the cAMP–PKA signaling cassette

PKA-mediated phosphorylation can regulate the activity of a number of ion channels, such as K^+, Cl^- and Ca^{2+} channels, and that of phosphatases involved in the regulation of cell signaling. In addition, translocation of activated PKA into the nucleus allows modulation of the activity of transcription factors such as the cAMP response element-binding protein (CREB) and the activation transcription factor (ATF) families, leading to either the induction or the repression of expression of specific genes (Fig. 40.5 and Chapter 34).

cAMP can stimulate cellular signaling independent of PKA

It has become clear that not all actions of cAMP are mediated by PKA. cAMP can also bind to Epacs (exchange proteins directly activated by cAMP), which are guanine-nucleotide-exchange factors for the Rap small GTPase. Epac activation has been implicated in the anti-inflammatory action of cAMP and in neuronal growth and development.

Signal cascades amplify signals initiated by receptor binding

The concentrations of hormones and other signals are often in the nanomolar (10^{-9} mol/L) or picomolar (10^{-12} mol/L) range. As a consequence, it is important that the signal is amplified. Multilayered signal transduction cascades cause substantial amplification of the original signal at each stage of the cascade, ensuring that binding of only a few hormone molecules leads to an appropriate biological response. For example, stimulation of glycogen breakdown by glucagon or epinephrine involves amplification of the signal at the level of G-proteins, adenylyl cyclase, protein kinase A and phosphorylase, such that many glucose-1-phosphate molecules are released (Fig. 40.6).

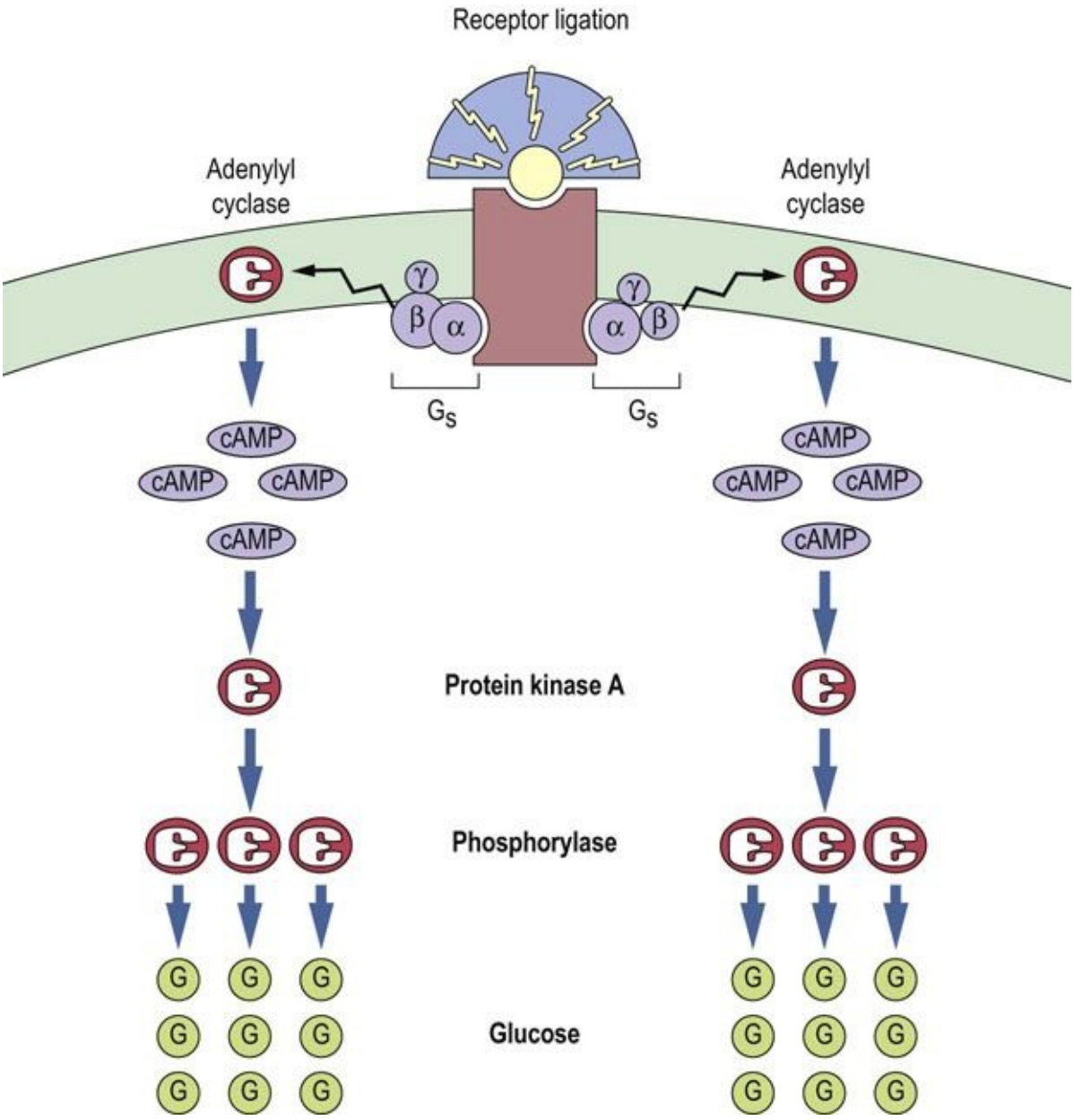


FIG. 40.6 Signal cascade induces amplification of hormone signal.
Each activated hormone–receptor complex can stimulate multiple G_s molecules. Each adenylyl cyclase can catalyze the generation of many cAMP molecules, and each protein kinase A can activate many phosphorylase molecules, leading to the breakdown of glycogen into many glucose-1-phosphate molecules as a result of glycogen degradation. (Compare Fig. 13.4.)

Clinical box Heart failure – potential for targeting β-adrenergic

receptor (βAR) signaling by gene therapy?

Heart disease is a major cause of death in developed countries and despite advances with electrophysiologic therapies and pharmacologic agents including β-adrenergic receptor (βAR) blockers (commonly known as β-blockers), angiotensin-converting enzyme (ACE) inhibitors and diuretics, the intracellular signal transduction abnormalities that underlie the development and progression of heart failure are not addressed.

Normally, binding of catecholamines (epinephrine and norepinephrine) to βARs activates Gαs to stimulate adenylyl cyclase and generate cAMP, resulting in the activation of PKA and, ultimately, via calcium-mediated excitation–contraction coupling, regulation of heart rate (chronotropy), contractility (inotropy) and relaxation (lusitropy). By contrast, in heart failure, myocytes exhibit characteristic changes in βAR signaling and consequent calcium mobilization, resulting in reduction of contractile function. Specifically, βAR expression is reduced and the remaining receptors appear to be desensitized due to increased levels of the negative regulator of βAR signaling, βAR kinase (βARK). Moreover, specific polymorphisms (mutations) in βARs have been identified in humans that are associated with poor prognosis following heart failure. As a consequence, much interest has focused on the therapeutic potential of modulating signaling via this pathway by cardiac gene therapy.

Gene therapy, where genetic material is introduced directly into the cell using viral vectors is the only approach that would effectively target signaling proteins within cardiomyocytes, in contrast to drugs that also have actions in tissues other than cardiac muscle. For example, studies in animal models to date have suggested that there is potential therapeutic value in terms of contractility and myocyte survival in increasing expression of β_2AR, but not β_1AR. Moreover, expression of the carboxy-terminal sequence of βARK (βARKct), which inhibits βARK-mediated desensitization of βARs, has been shown to prevent heart failure in rodents. Likewise, overexpression of downstream βAR effectors

such as $AC_{VI,}$ the major isoform of adenylyl cyclase in myocardium, has produced therapeutic effects. Finally, as βAR coupling to PKA signaling regulates the receptors and ion channels responsible for the calcium signals required for excitation–contraction coupling, downstream inotropic regulators of contractile function, such as the sarcoplasmic reticulum Ca^{2+}-ATPase (SERCA2a) and the EF hand-type calcium sensing protein (S100A1), which are also downregulated in heart failure, are also currently under study as potential therapeutic targets. Indeed, a clinical trial using adeno-associated virus to increase SERCA2a is currently underway in patients with advanced heart failure.

Phosphodiesterases terminate the cAMP signal

Phosphodiesterases (PDEs) terminate the cAMP signal by converting cAMP to its 5′-AMP metabolite (Fig. 40.4); thus they have the potential to play key roles in the regulation of various physiologic responses in many different cells and tissues. There are many different isoforms of PDEs that exhibit a tissue-specific pattern of expression and different selectivity for cAMP or cGMP. PDEs have been demonstrated to regulate platelet activation, vascular relaxation, cardiac muscle contraction and inflammation.

Inhibitors of phosphodiesterases are used as therapeutic agents

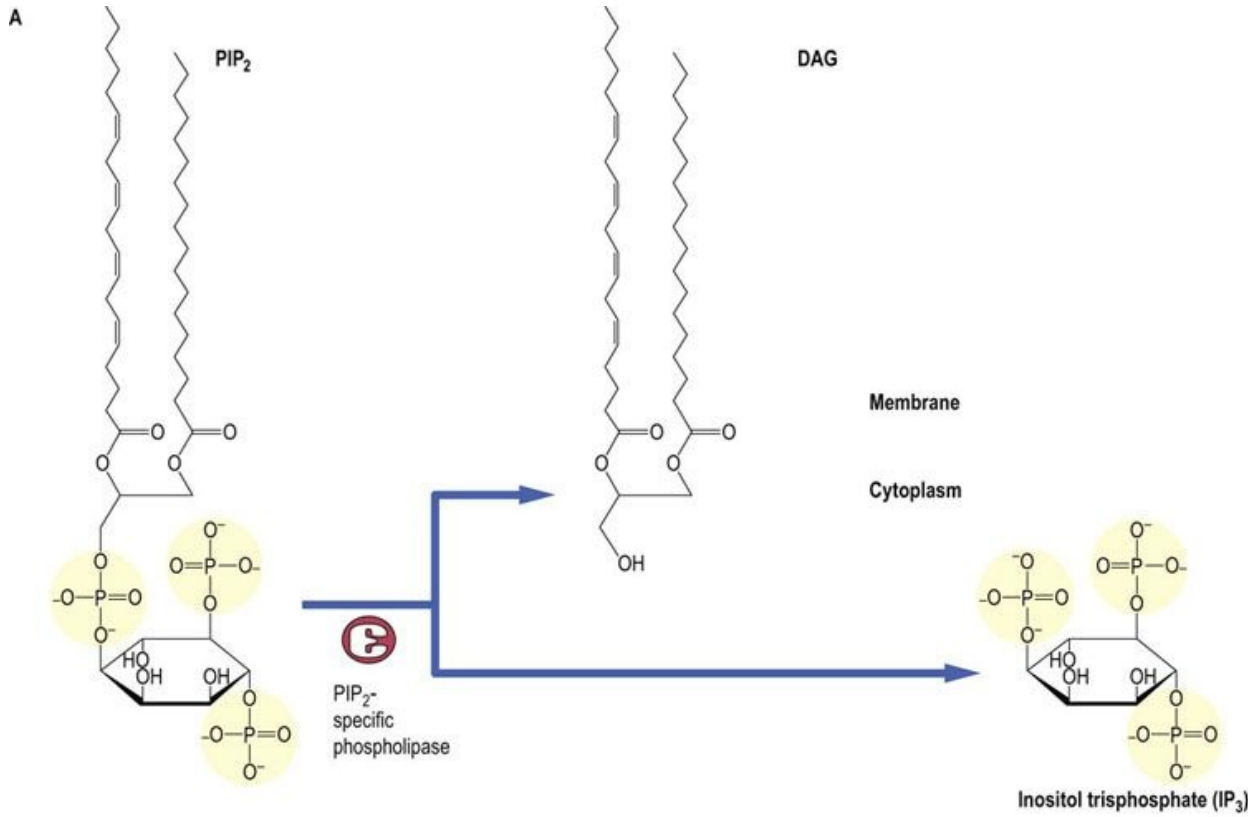
Selective inhibitors of PDEs have been used as therapeutic agents for **asthma** (methylxanthines), **erectile dysfunction** (sildenafil) and **heart failure** (milrinone). Milrinone is selective for the PDE_3 isoforms, which increases the force of contraction of the heart, presumably due to increased cAMP concentrations and PKA activity, leading to phosphorylation of cardiac calcium channels and a subsequent increase in intracellular calcium concentration.

Phospholipase-derived second messengers

Phospholipase C hydrolyzes the membrane phospholipid

phosphatidylinositol 4,5-bisphosphate to generate two second messengers

G-protein coupled receptors that are coupled to the $G_q\alpha$ subtype of the G-protein α-subunit stimulate the activity of **phospholipase C (PLC).** In addition, other types of membrane receptor, such as the vascular endothelial growth factor (VEGF) receptor, that has intrinsic tyrosine kinase activity, also are able to stimulate PLC. PLC catalyzes the hydrolysis of a minor phospholipid species, phosphatidylinositol 4,5-bisphosphate (PIP_2). PLC-mediated hydrolysis of PIP_2, which typically represents about 0.4% of total phospholipids in membranes, generates two second messengers: **inositol -1,4.5-trisphosphate** (IP_3) and **diacylglycerol** (DAG) (Fig. 40.7). IP_3 is a water-soluble product, which is released into the cytosol and has been shown to mobilize intracellular stores of calcium. DAG is a lipid second messenger, which is anchored in the plasma membrane by virtue of its hydrophobic fatty acid side chains and activates a key family of signaling enzymes known as protein kinase C (PKC).



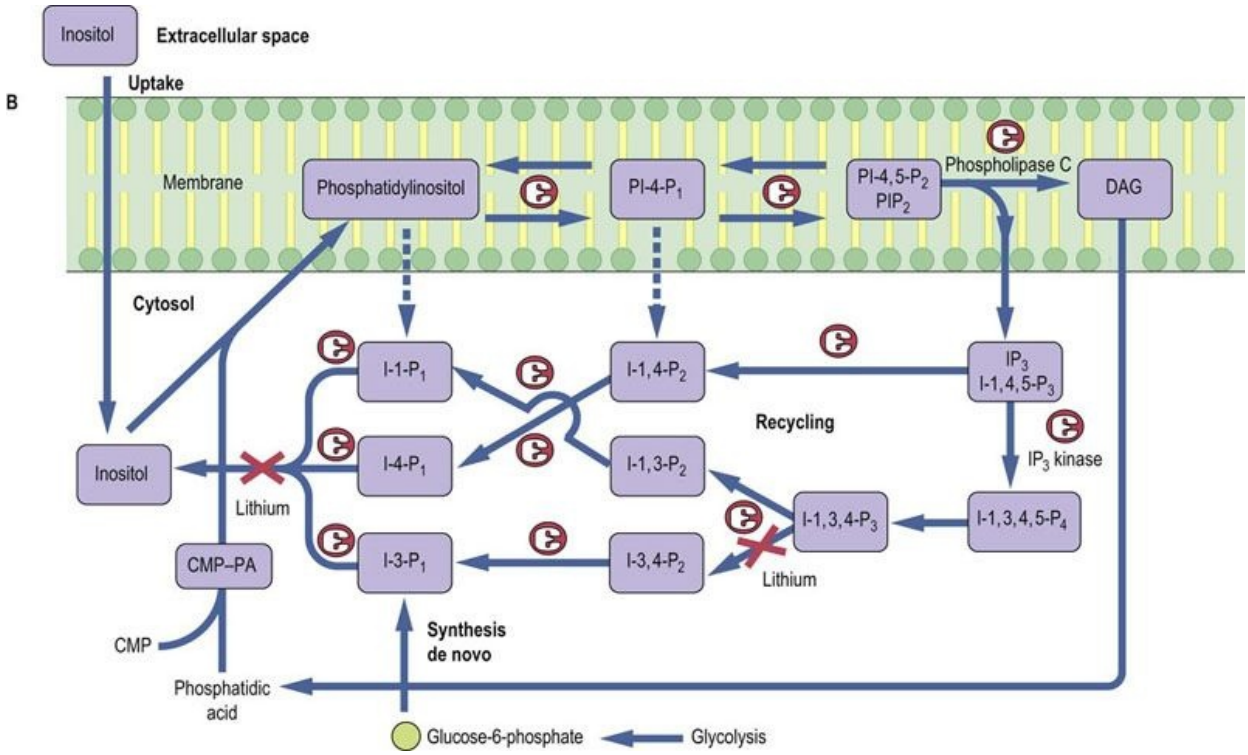


FIG. 40.7 Synthesis and metabolism of phosphatidylinositol 4,5-bisphosphate. Phosphatidylinositol 4,5-bisphosphate (PIP_2) is hydrolyzed by a PIP_2-specific PLC, to generate two second messengers: Inositol-1,4,5-P_3 (IP_3) and DAG **(A)**. IP_3 is released into the cytosol and has been shown to mobilize intracellular stores of calcium. DAG is a lipid second messenger which is anchored in the plasma membrane and activates PKCs. PIP_2 is generated from phosphatidylinositol by phosphatidylinositol 4-kinase and phosphatidylinositol 5-kinase **(B)**. IP_3 is degraded by (i) the sequential action of phosphatases converting Ins1,4,5-P_3 to inositol, and (ii) an IP_3 kinase, which generates Inositol-1,3,4,5-P_4, which is in turn sequentially degraded to inositol by inositol phosphate-specific phosphatases, some of which can be inhibited by lithium. DAG, diacylglycerol; PA, phosphatidic acid; CMP-PA, cytosine monophosphate-phosphatidic acid.

Inositol 1,4,5-trisphosphate (IP_3) stimulates intracellular calcium mobilization

Once synthesized from PIP_2, IP_3 binds to receptors found on the endoplasmic reticulum of all cells. IP_3 receptors are a family of related glycoproteins (molecular mass 250 kDa) comprising six transmembrane-spanning domains. The active receptor is expressed as a multimer of four IP_3 receptor molecules that acts as a **ligand-gated Ca^{2+}** channel. The tetrameric structure of the IP_3

receptor gives rise to cooperativity in Ca^{2+} channel activity. It has been estimated that stimulation with IP_3 causes transport of 20–30 calcium ions, revealing the amplification inherent in this signaling cascade. Consistent with the transient nature of the release of intracellular calcium that is observed after hormone receptor ligation, cellular concentrations of IP_3 are rapidly returned to resting values (10 nmol/L) by more than one route of degradation (Fig. 40.7).

Calcium and calmodulin. Signal transduction by calcium ions

Calcium (Ca^{2+}) is a ubiquitous messenger with an important role in the transduction of signals leading to diverse cellular responses such as cell motility changes, egg fertilization, neurotransmission, secretion, differentiation and proliferation. Cells expend a considerable amount of energy maintaining a Ca^{2+} concentration gradient such that the intracellular Ca^{2+} concentration in resting, unstimulated cells is of the order of 10^{-7} mol/L, whereas the extracellular Ca^{2+} concentration is approximately 10,000-fold greater, typically 10^{-3} mol/L. This steep gradient allows for rapid, abrupt, transient changes in Ca^{2+} concentration in response to signals. Ligand binding by a wide range of receptors leads to a PLC-mediated rapid (within seconds) and transient increase in intracellular Ca^{2+} concentration to the micromolar range (compare Fig. 8.4). The rapid changes in Ca^{2+} concentrations are very tightly regulated and utilize a variety of mechanisms involving cell compartmentalization. For example, intracellular Ca^{2+} concentrations can be lowered by sequestration of Ca^{2+} into the endoplasmic reticulum by Ca^{2+}-ATPases, or into the mitochondria using the energy-driven electrochemical gradient. Alternatively, free Ca^{2+} can be chelated by Ca^{2+}-binding proteins such as calsequestrin.

Many downstream signaling events mediated by Ca^{2+} are modulated by a Ca^{2+}-sensing and binding protein, calmodulin

Calmodulin (CaM) is a 17 kDa protein found in all animal and plant cells (comprising up to 1% of cellular protein). It belongs to a family of proteins characterized by one or more copies of a Ca^{2+}-binding structural motif called an EF hand motif (Fig. 40.8). CaM is composed of two similar globular domains joined by a long α-helix, each globular lobe having two EF hand motifs/Ca^{2+}-binding sites placed 1.1 nm (11 Å) apart. Binding of three to four calcium ions

occurs when the intracellular Ca^{2+} concentration is increased to about 500 nmol/L, inducing a major conformational change that allows CaM to bind to and modify target proteins. Binding of several calcium ions allows cooperativity in the activation of CaM, such that small changes in Ca^{2+} concentration cause large changes in the concentration of an active Ca^{2+}/CaM complex, providing amplification of the original hormone signal.

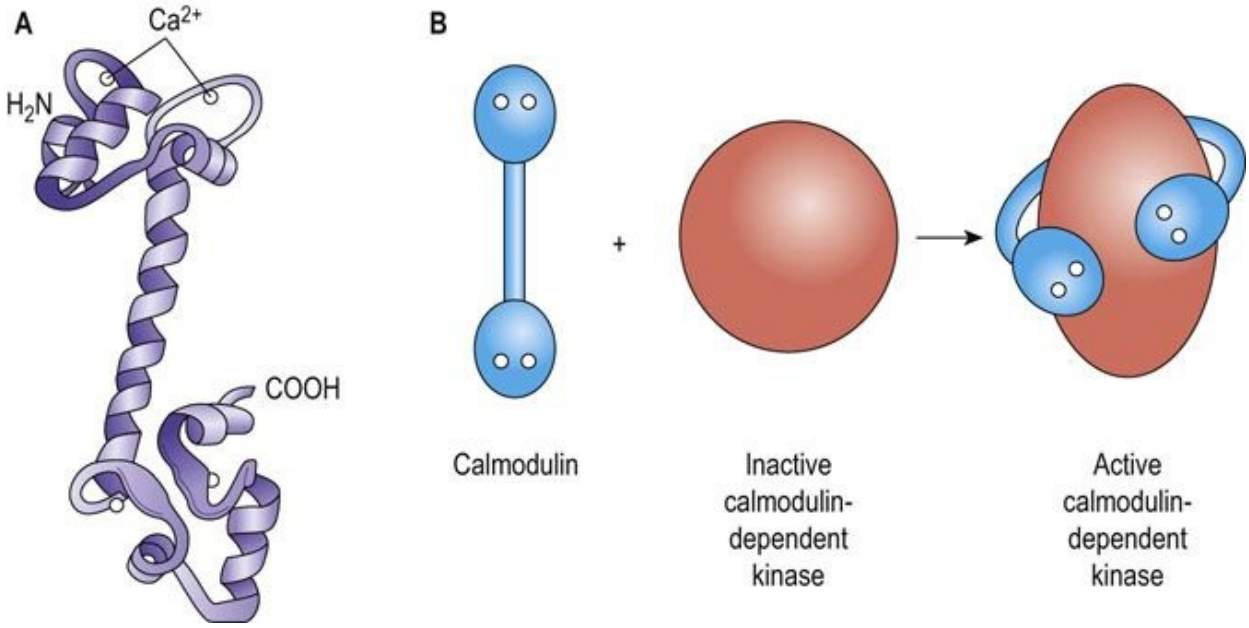


FIG. 40.8 Calmodulin.
(A) Structure of calmodulin. **(B)** Binding of calcium induces a conformational change, allowing calmodulin to bind to and modify the activity of target signaling enzymes.

Calmodulin has a wide range of target effectors

CaM has a wide range of target effectors, including nitric oxide synthase (NOS), which stimulates NO synthesis in response to Ca^{2+}-mobilizing signals, and Ca^{2+}/CaM-dependent protein kinases, which phosphorylate serine-threonine residues on proteins to regulate a variety of processes. For example, the broad-specificity kinase, Ca^{2+}/CaM-kinase II, is involved in the regulation of fuel metabolism, ion permeability, neurotransmitter biology and muscle contraction. Interestingly, CaM serves as a permanent regulatory subunit of phosphorylase kinase, and may also regulate non-kinase effectors such as certain adenylyl cyclase isoforms and also cAMP-specific PDEs, permitting ‘cross-talk’ between

cAMP-and Ca^{2+}-dependent signaling pathways.

Diacylglycerol (DAG) activates protein kinase C

DAG fulfills its second messenger role by activating protein kinase C (PKC), which phosphorylates a wide range of target signal transduction proteins on serine or threonine residues. PKC was originally identified as a calcium-and lipid (phosphatidylserine)-dependent kinase important in the regulation of cell proliferation. However, in recent years, it has become clear that PKC is, in reality, a generic name for a superfamily of related kinases that have different activation requirements (Table 40.3). PKC family members also exhibit tissue-specific expression. Nevertheless, all these enzymes share some conserved features; most notably they comprise two major domains: an *N*-terminal regulatory domain and a *C*-terminal catalytic kinase domain. The regulatory domain contains a pseudosubstrate sequence that resembles the consensus phosphorylation site in PKC substrates. In the absence of activating cofactors (Ca^{2+}, phospholipid, DAG), this pseudosubstrate sequence interacts with the substrate-binding pocket in the catalytic domain and represses PKC activity; binding of cofactors reduces the affinity of this interaction, induces a conformational change in the PKC, and allows stimulation of PKC activity. Consistent with the fact that the activator/cofactor, DAG, is anchored in the membranes, PKC activation is generally associated with translocation from the cytosol to the plasma mebrane or nuclear membranes.

Table 40.3

The protein kinase C superfamily

	Conventional PKCs			Novel PKCs			Atypical PKCs		
	α	β	γ	δ	ε	η	θ	λ (ι in mouse)	ζ
Ca^{2+}-sensitive	Yes	Yes	Yes	No	No	No	No	No	No
DAG-sensitive	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	No

Other phospholipases hydrolyze phosphatidylcholine or phosphatidylethanolamine, generating a range of lipid second messengers

Additional receptor-coupled lipid signaling pathways have been identified involving hydrolysis of phosphatidylcholine or phosphatidylethanolamine,

which can give rise to DAG and other biologically active lipids (Fig. 40.9) in response to a wide range of growth factors and mitogens. Phosphatidylcholine comprises about 40% of total cellular phospholipids. It can be hydrolyzed by distinct phospholipases, generating a diversity of lipid second messengers, including fatty acids such as arachidonic acid (generated by PLA_2) as well as different species of DAG (generated by PLC) and phosphatidic acid (generated by PLD). Hormone-stimulated phosphatidylethanolamine-PLD activities have also been reported. Some hormones or growth factors can stimulate only one or other of these phospholipases, but other ligands can stimulate all these pathways after binding to their specific receptors.

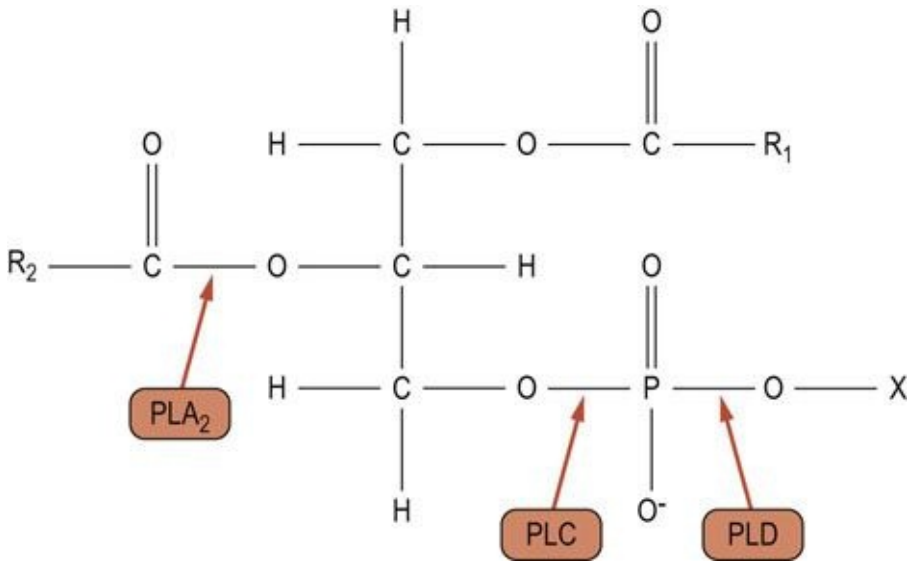


FIG. 40.9 Sites of action of phospholipases.
Hydrolysis of phosphatidylcholine or phosphatidylethanolamine by phospholipase A_2 (PLA_2) results in the production of lysophosphatidylcholine or lysophosphatidylethanolamine and a fatty acid. Hydrolysis by phospholipase C (PLC) results in the synthesis of DAG and phosphocholine or phosphoethanolamine. Hydrolysis of phosphatidylcholine or phosphatidylethanolamine by phospholipase D (PLD) results in the production of phosphatidic acid and choline or ethanolamine. R1, R2: fatty acyl chains; X: choline/ethanolamine.

There is growing evidence that all these distinct lipid second messengers have different targets. For example, it has recently been suggested that the saturated/monounsaturated fatty acid-containing DAGs derived from phosphatidylcholine-specific phospholipase D (PLD) activation are unable to

activate PKC isoforms, and that it is only the stearoyl-arachidonyl-phosphatidic acid species that can modulate activity of the GTPase Ras (Chapter 43). Generation of these diverse but related lipid second messengers therefore provides a mechanism for initiating or terminating hormone-specific responses via particular signal transducers, including differential activation of PKC isoforms.

Arachidonic acid is a second messenger regulating phospholipases and protein kinases

Arachidonic acid is a C20 polyunsaturated fatty acid containing four double bonds. Increased arachidonic acid synthesis has been demonstrated to regulate several signaling enzymes, including PLC and PKC-α, -β and -γ isoforms. Furthermore, arachidonic acid is a key inflammatory intermediate. However, the arachidonic acid involved in these disparate functions appears to be generated by two distinct PLA_2 routes. Arachidonic acid generated for signaling purposes appears to be derived from the action of a phosphatidylcholine-specific cytosolic phospholipase A_2 ($cPLA_2$), which has a molecular mass of 85 kDa and is regulated by phosphorylation of key serine residues. In contrast, inflammatory arachidonic acid is generated by the action of a family of low-molecular-mass secretory PLA_2 ($sPLA_2$) proteins (14–18 kDa), which appear to be ubiquitous and are found in high concentrations in snake venom and pancreatic juices. In addition, arachidonic acid can be generated by DAG lipase.

Arachidonic acid is the precursor of eicosanoids, which encompass prostaglandins, prostacyclins, thromboxanes, and leukotrienes

As a key inflammatory mediator, arachidonic acid is the major precursor of the group of molecules termed eicosanoids, which encompass **prostaglandins, prostacyclins, thromboxanes,** and **leukotrienes.** Eicosanoids act like hormones and signal via G-protein coupled receptors. They have a wide variety of biological activities, including modulating smooth muscle contraction (vascular tone), platelet aggregation, gastric acid secretion, and salt and water balance, as well as mediating pain and inflammatory responses. Prostaglandins, prostacyclins, and thromboxanes are synthesized in membranes from arachidonic acid by the successive actions of several enzymes, starting with cyclooxygenase (Fig. 40.10 and see also Fig 7.2).

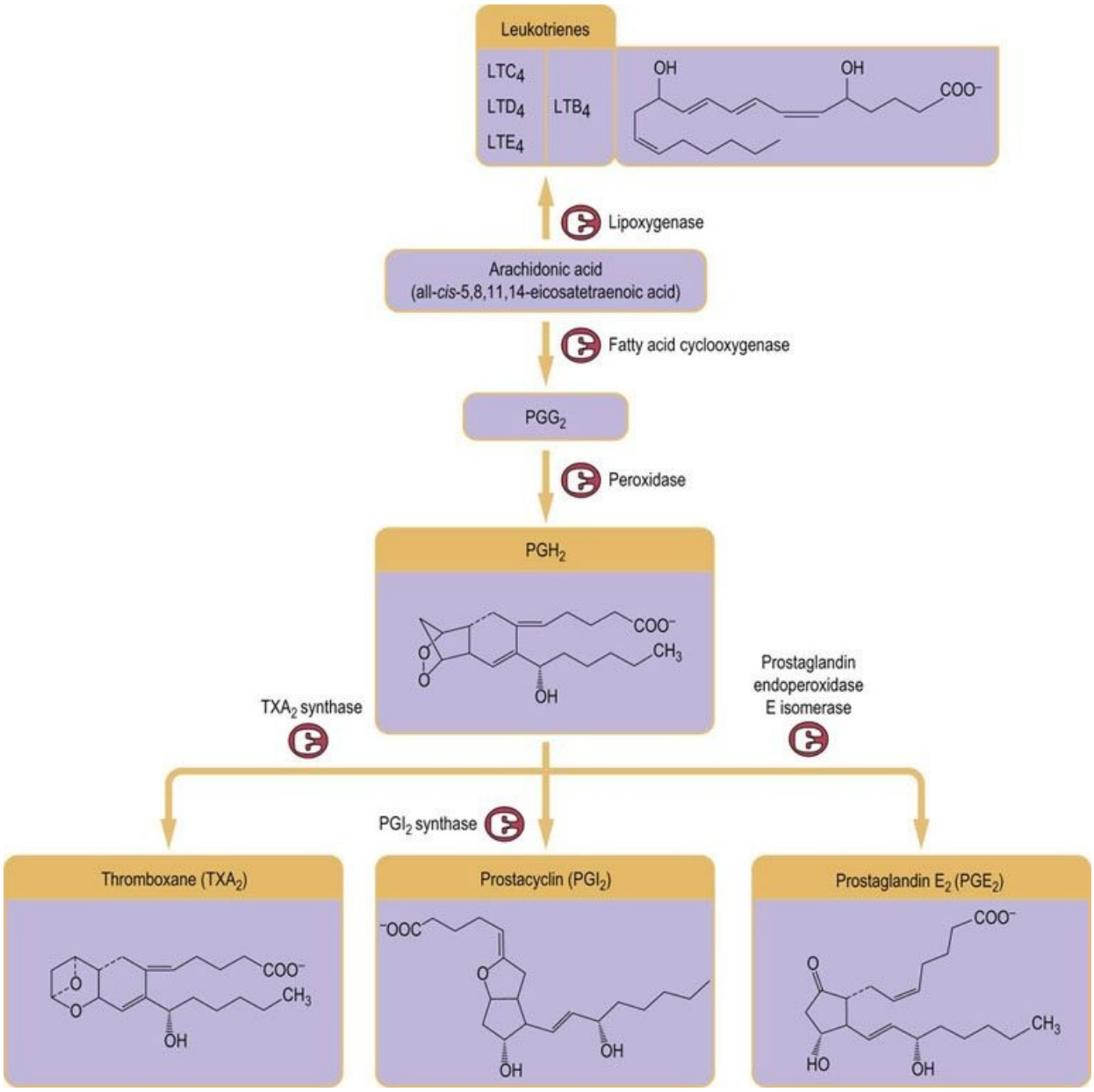


FIG. 40.10 Synthesis of eicosanoids.
Eicosanoids are primarily derived from arachidonic acid. Leukotrienes (LT) are synthesized via a lipoxygenase-dependent pathway, whereas prostaglandins (PG), prostacyclins and thromboxanes (TX) arise from cyclooxygenase-dependent routes.

Clinical box A child with premature breast development: mcCune–albright syndrome

A 3-year-old girl was brought to hospital because her mother had been concerned about apparent breast development over the last 6 months, and a spot of blood on her pants last week. On examination, she had Tanner Stage 3 breast development. On her trunk she had three areas of brown skin pigmentation with ragged edges.

Comment.

This child is suffering from McCune–Albright syndrome. She is likely to develop polyostotic fibrous dysplasia, with areas of thinning and sclerosis in her long bones which may fracture. Other endocrinopathies include thyrotoxicosis, growth hormone hypersecretion, Cushing's syndrome (cortisol excess) and hyperparathyroidism. The cause is an activating missense mutation in the gene encoding the $G_{s\alpha}$-subunit of the G-protein that stimulates cyclic AMP formation. The problem presents following a somatic cell mutation with clinical features dependent on a mosaic distribution of aberrant cells. The incidence of the syndrome is 1 in 25,000.

Clinical box Fibroblast growth factor receptor 3 and achondroplasia

Fibroblast growth factor (FGF) receptor 3 (FGFR3) is an intrinsic tyrosine kinase receptor with an important role in the regulation of bone growth. In chondrocytes (the cells that synthesize cartilage at the epiphyses of long bones), FGF binding to two FGFR3 monomers causes dimerization of the receptor, allowing the intracellular tyrosine kinase domains to transphosphorylate each other. This autophosphorylation of FGFR3 leads to activation of signaling pathways including the transcription factor, signal transducer and activators of transcription-1 (STAT1) and the small G-protein Ras, which subsequently activates the Raf-MEK-ERK protein kinase cascade. Activation of these pathways inhibits

chondrocyte differentiation and proliferation. Thus, FGFR3 stimulation suppresses long bone growth as the reduced number of chondrocytes decreases cartilage deposition that would normally serve as a template for osteoblasts to form bone by ossification.

Achondroplasia, characterized by short stature and macrocephaly, has an incidence of 1 in 15,000–40,000 newborns. The majority of people with achondroplasia have mutations in FGFR3 that increase its tyrosine kinase activity in the absence of FGF. As a consequence, chondrocyte proliferation and cartilage deposition are impaired, leading to reduced length of long bones.

Summary

■ Cells specifically respond to a multiplicity of signals from their environment via signal transduction cassettes, which comprise specific cell surface membrane receptors, effector signaling systems (e.g. adenylyl cyclase, phospholipases, or ion channels) and regulatory proteins (e.g. G-proteins or protein kinases).
■ These signal transduction cassettes serve to detect, amplify and integrate diverse external signals to generate the appropriate cellular response.
■ The families of cell surface receptors sense and transduce their specific hormone signal by transmembrane coupling to different effector systems to generate low-molecular-weight molecules, termed second messengers, such as cAMP, IP_3, DAG, and Ca^{2+}, which mediate their signaling functions by activating key protein kinases.
■ The specificity of a particular hormone response can be further heightened by the variety of available phospholipase-signaling activities (PLC, PLD and PLA_2). Taking into account their range of potential lipid substrates (e.g. PIP_2, phosphatidylcholine, and phosphatidylethanolamine) and products (e.g. DAG, phosphatidic acid and arachidonic acid), the phospholipases can generate a diverse array of lipid second messengers.

Active learning

1. Compare and contrast a receptor for a polypeptide hormone and for a steroid.
2. Describe the mechanism of G-protein signaling.
3. Give an example of a receptor possessing protein kinase activity and the components of its signaling cascade.
4. Comment on the diversity of phospholipase enzymes.
5. Describe the termination mechanism of cAMP activation.

Further reading

Fredholm, BB, Hökfelt, T, Milligan, G. G-protein-coupled receptors: an update. *Acta Physiol (Oxf)*. 2007; 190:3–7.

Halls, ML, Cooper, DM. Regulation by Ca^{2+}-signaling pathways of adenylyl cyclases. *Cold Spring Harb Perspect Biol*. 2011; 3:a004143.

Houslay, MD. Underpinning compartmentalised cAMP signaling through targeted cAMP breakdown. *Trends Biochem Sci*. 2010; 35:91–100.

Lefkowitz, RJ. Seven transmembrane receptors: something old, something new. *Acta Physiol (Oxf)*. 2007; 190:9–19.

Leto, D, Saltiel, AR. Regulation of glucose transport by insulin: traffic control of GLUT4. *Nat Rev Mol Cell Biol*. 2012; 13:383–396.

Michel, T, Vanhoutte, PM. Cellular signaling and NO production. *Pflugers Arch*. 2010; 459:807–816.

Osborne, JK, Zaganjor, E, Cobb, MH. Signal control through Raf: in sickness and in health. *Cell Res*. 2012; 22:14–22.

Parekh, AB. Decoding cytosolic Ca^{2+} oscillations. *Trends Biochem Sci*. 2011; 36:78–87.

Smith, WL. Nutritionally essential fatty acids and biologically indispensable cyclooxygenases. *Trends Biochem Sci*. 2008; 33:27–37.

Websites and downloads

Science magazine's signal transduction knowledge environment. www.stke.org.

Kimball's biology pages. http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/C/CellSignaling.html.

The Biology Project (University of Arizona). www.biology.arizona.edu/cell_bio/problem_sets/signaling/Index.htm.

BioCarta's cell signaling pathways. www.biocarta.com/genes/CellSignaling.asp.

Cell signaling pathway maps. www.cellsignal.com/reference/pathway/index.html.

CHAPTER 41.1

Neurotransmitters

Simon J.R. Heales

Learning objectives

After reading this chapter you should be able to:

■ Outline the criteria that need to be met before a molecule can be classified as a neurotransmitter.

■ Identify the major neurotransmitter types and be aware that some molecules have neurotransmitter properties but cannot in the strictest sense be classified as neurotransmitters.

■ Explain the generation of action potentials, appreciate how neurotransmitters can be excitatory or inhibitory, and summarize the process whereby a neurotransmitter is released from the presynaptic cell.

■ Describe the different neurotransmitter receptors and their general mode of action.

■ Describe the major biochemical pathways for neurotransmitter synthesis and degradation.

■ Identify some clinical disorders that can arise as a result of disruption of neurotransmitter metabolism.

Introduction

Neurotransmitters are molecules that act as chemical signals between nerve cells

Nerve cells communicate with each other and with target tissues by secreting chemical messengers, called neurotransmitters. This chapter describes the various classes of neurotransmitters and how they interact with their target cells. It will discuss their effects on the body, how alterations in their signaling may cause disease, and how pharmacologic manipulation of their concentrations may be used therapeutically.

Definition of a neurotransmitter

Traditionally, for a molecule to be labeled as a neurotransmitter, a number of criteria have to be met:

- Synthesis of the molecule occurs within the neuron, *i.e.* all biosynthetic enzymes, substrates, cofactors, etc., must be present for de novo synthesis.
- Storage of the molecule occurs within the nerve ending prior to release, *e.g.* in synaptic vesicles.
- Release of the molecule from the presynaptic ending occurs in response to an appropriate stimulus such as an action potential.
- There is binding and recognition of the putative neurotransmitter molecule on the postsynaptic target cell.
- Mechanisms exist for the inactivation and termination of the biological activity of the neurotransmitter.

Rigorous adherence to the above criteria means that some molecules that are involved in the cross-talk between neurons are not in the strict sense classified as neurotransmitters. Thus, nitric oxide (NO), adenosine, neurosteroids, polyamines, etc., are often termed **neuromodulators** rather than neurotransmitters.

Classification of neurotransmitters

A classification of neurotransmitters based on chemical composition is shown in Table 41.1.1. Many are derived from simple compounds, such as amino acids (Table 41.1.2), but peptides are also now known to be extremely important. The principal transmitters in the peripheral nervous system are norepinephrine and acetylcholine (ACh) (Fig. 41.1.1).

Table 41.1.1

Classification of neurotransmitters

Group	Examples
Amines	Acetylcholine (ACh), norepinephrine, epinephrine, dopamine, 5-HT
Amino acids	Glutamate, GABA
Purines	ATP, adenosine
Gases	Nitric oxide
Peptides	Endorphins, tachykinins, many others

5-HT, 5-hydroxytryptamine; GABA, *y*-amino butyric acid.

Neurotransmitters can be classified in several ways. The scheme shown relies on chemical similarities. All except the peptides are synthesized at the nerve ending and packaged into vesicles there; peptides are synthesized in the cell body and transported down the axon.

Table 41.1.2

Neurotransmitters of low molecular weight

Compound	Precursor	Site of production
Amino acids		
Glutamate		Central nervous system (CNS)
Aspartate		CNS
Glycine		Spinal cord
Amino acid derivatives		
GABA	Glutamate	CNS
Histamine	Histidine	Hypothalamus
Norepinephrine	Tyrosine	Sympathetic nerves, CNS
Epinephrine	Tyrosine	Adrenal medulla, a few CNS nerves
Dopamine	Tyrosine	CNS
5-HT	Tryptophan	CNS, enterochromaffin gut cells, enteric nerves
Purine derivatives		
ATP		Sensory, enteric, sympathetic nerves
Adenosine	ATP	CNS, peripheral nerves
Gas		
Nitric oxide	Arginine	Genitourinary tract, CNS
Miscellaneous		
Acetylcholine	Choline	Parasympathetic nerves, CNS

Many neurotransmitters are simple compounds, often derived from common amino acids.

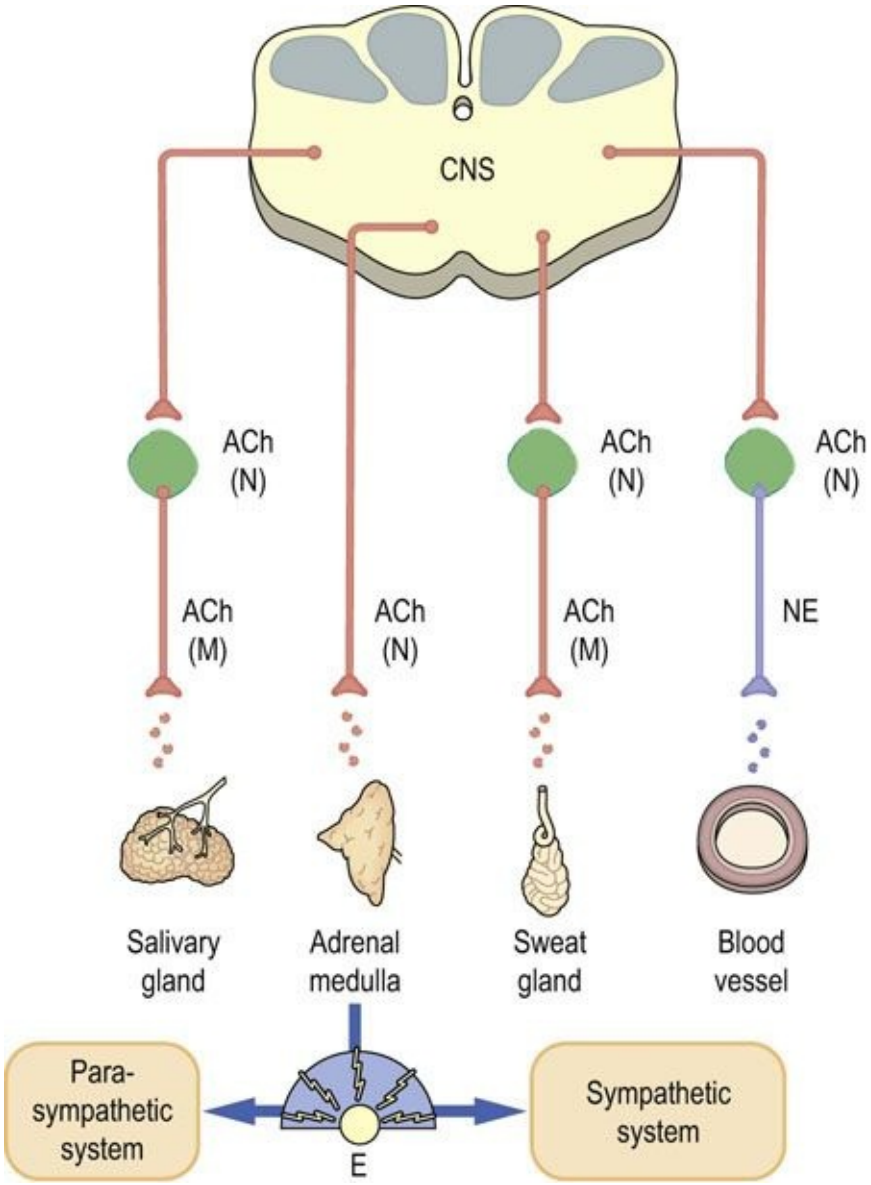


FIG. 41.1.1 Transmitters in the autonomic nervous system.
Catecholamines and acetylcholine (ACh) are transmitters in the **sympathetic** and parasympathetic nervous systems. Preganglionic nerves all release ACh, which binds to **nicotinic (N) receptors.** Most postganglionic sympathetic nerves release norepinephrine (NE), whereas postganglionic parasympathetic nerves release ACh, which acts at **muscarinic (M) receptors.** Adrenal glands release epinephrine. Motor neurons release ACh, which acts at distinct nicotinic receptors. E, epinephrine (see also Fig. 41.1.3).

Several transmitters may be found in one nerve

An early dogma of nerve function held that one nerve contained one transmitter.

However, this is now known to be an oversimplification, and combinations of transmitters are the rule. The pattern of cellular transmitters may characterize a particular functional role, but details of this also remain unclear. A major low-molecular-weight transmitter such as an amine is often present, along with several peptides, an amino acid, and a purine. Sometimes, there may even be more than one possible transmitter in a particular vesicle, as is believed to be the case for adenosine triphosphate (ATP) and norepinephrine in sympathetic nerves. In some cases, the intensity of stimulation may control which transmitter is released, peptides often requiring greater levels of stimulus. Furthermore, different transmitters may have a different timescale of action. Sympathetic nerves are good examples of nerves for which this is the case: it is believed that ATP causes their rapid excitation, whereas norepinephrine and the neuromodulator neuropeptide Y (NPY) cause a slower phase of action. In some tissues, NPY on its own may be able to produce a very slow excitation.

Neurotransmission

Action potentials are caused by changes in ion flows across cell membranes

The signal carried by a nerve cell reflects an abrupt change in the voltage potential difference across the cell membrane. The normal **resting potential** difference is a few millivolts, with the **inside of the cell being negative**, and is caused by an imbalance of ions across the plasma membrane: the concentration of K^+ ion is much greater inside cells than outside, whereas the opposite is true for Na^+ ion. This difference is maintained by the action of the **Na^+/K^+-ATPase** (Chapter 24). Only those ions to which the membrane is permeable can affect the potential, as they can come to an electrochemical steady state under the combined influence of concentration and voltage differences. Because the membrane in all resting cells is comparatively permeable to K^+ as a result of the presence of voltage-independent (leakage) K^+ channels, this ion largely controls the resting potential.

A change in voltage which tends to drive the resting potential towards zero from the normal negative voltage is known as a depolarization, whereas a process that increases the negative potential is called hyperpolarization

So far, this picture is common to all cells. However, nerve cells contain voltage-dependent sodium channels that open very rapidly when a depolarizing change in voltage is applied. When they open, they allow the inward passage of huge numbers of Na^+ ions from the extracellular fluid (Fig. 41.1.2), which swamps the resting voltage and drives the membrane potential to positive values. This reversal of voltage is the **action potential.** Almost immediately afterwards, the sodium channels close and so-called delayed potassium channels open. These restore the normal resting balance of ions across the membrane and, after a short refractory period, the cell can conduct another action potential. Meanwhile, the action potential has spread by electrical conductance to the next segment of nerve membrane, and the entire cycle starts again.

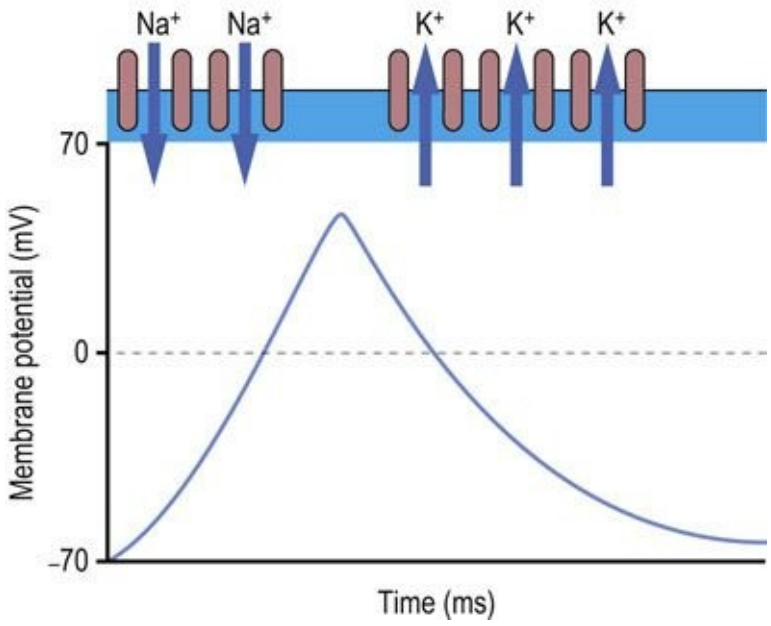


FIG. 41.1.2 Generation of action potential.
Action potential is formed as follows. At the start of an action potential, the membrane is at its resting potential of about −70 mV. This is maintained by voltage-independent K^+ channels. When an impulse is initiated by a signal from a neurotransmitter, voltage-dependent Na^+ channels open. These allow inflow of Na^+ ions, which alter the membrane potential to positive values. The Na^+ channels then close and K^+ channels, called delayed rectifier channels, open to restore the initial balance of ions and the negative membrane potential.

Neurotransmitters alter the activity of various ion channels to cause changes in the membrane potential

Excitatory neurotransmitters cause a **depolarizing** change in voltage, in which case an action potential is more likely to occur. In contrast, **inhibitory** transmitters **hyperpolarize** the membrane and an action potential is then less likely to occur.

Neurotransmitters act at synapses

Neurotransmitters are released into the space between cells at a specialized area known as a synapse (Fig. 41.1.3). In the simplest case, they diffuse from the presynaptic membrane across the synaptic space or cleft, and bind to receptors at the postsynaptic membrane. However, many neurons, particularly those containing amines, have several varicosities along the axon, containing transmitter. These varicosities may not be close to any neighboring cell, so

transmitter released from them has the possibility of affecting many neurons. Nerves innervating smooth muscle are commonly of this kind.

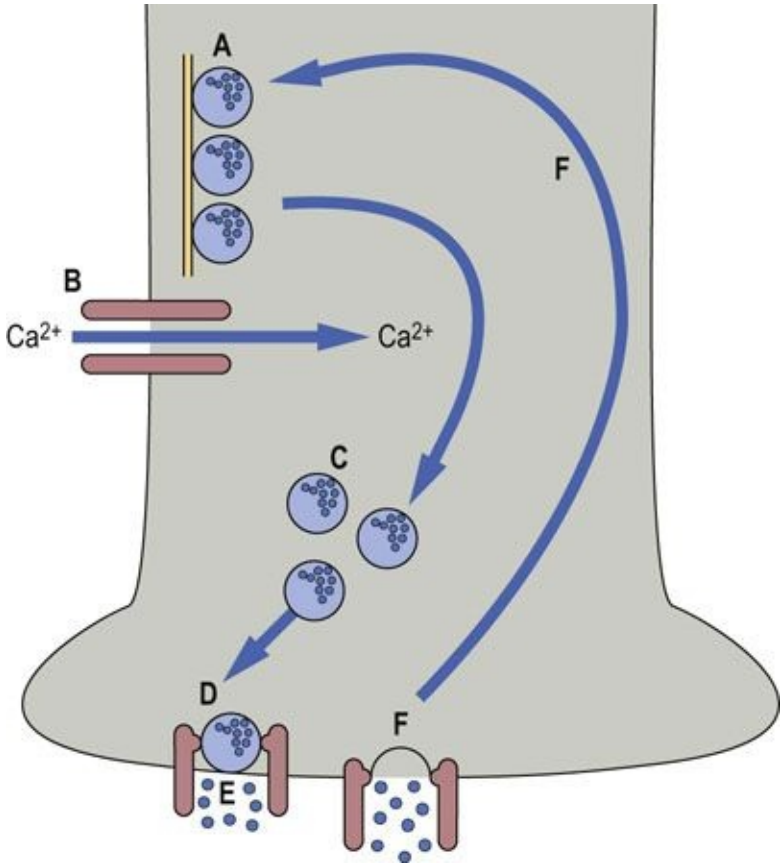


FIG. 41.1.3 Release of neurotransmitters.
Neurotransmitters are released from vesicles at the synaptic membrane. **(A)** In the resting state, vesicles are attached to microtubules. **(B)** When an action potential is received, calcium channels open. **(C)** Vesicles move to the plasma membrane, and **(D)** bind to a complex of docking proteins. **(E)** Neurotransmitter is released, and **(F)** vesicles are recycled.

When the action potential arrives at the end of the axon, the change in voltage opens calcium channels. Calcium entry is essential for mobilization of vesicles containing transmitter, and for their eventual fusion with the synaptic membrane and release through it.

Because transmitters are released from vesicles, impulses arrive at the postsynaptic cell in individual packets, or quanta. At the neuromuscular junction between nerves and skeletal muscle cells, a large number of vesicles are discharged at a time, and a single impulse may therefore be enough to stimulate

contraction of the muscle cell. The number of vesicles released at synapses between neurons, however, is much smaller; consequently, the recipient cell will be stimulated only if the total algebraic sum of the various positive and negative stimuli exceeds its threshold. As each cell in the brain receives input from a huge number of neurons, this implies that there is a far greater capability for the fine control of responses in the central nervous system (CNS) than there is at the neuromuscular junction.

Receptors

Neurotransmitters act by binding to specific receptors, and opening or closing ion channels

There are several mechanisms by which receptors for excitatory neurotransmitters can cause the propagation of an action potential in a postsynaptic neuron. Directly or indirectly, they cause changes in ion flow across the membrane, until the potential reaches the critical point, or threshold, for initiation of an action potential. Receptors that directly control the opening of an ion channel are called **ionotropic,** whereas **metabotropic** receptors cause changes in second messenger systems, which in turn alter the function of channels that are separate from the receptor.

Ionotropic receptors (ion channels)

Ionotropic receptors contain an ion channel within their structure (Fig. 41.1.4; see also Chapter 8). Examples include the **nicotinic ACh receptor** and some glutamate and **γ-amino butyric acid (GABA) receptors.** These are transmembrane proteins, with several subunits, usually five, surrounding a pore through the membrane. Each subunit has four transmembrane regions. When the ligand binds, there is a change in the three-dimensional structure of the complex, which allows the flow of ions through it. The effect on membrane potential depends on the particular ions that are allowed to pass: the nicotinic ACh receptor is comparatively nonspecific towards sodium and potassium and causes depolarization, whereas the $GABA_A$ receptor is a chloride channel and causes hyperpolarization.

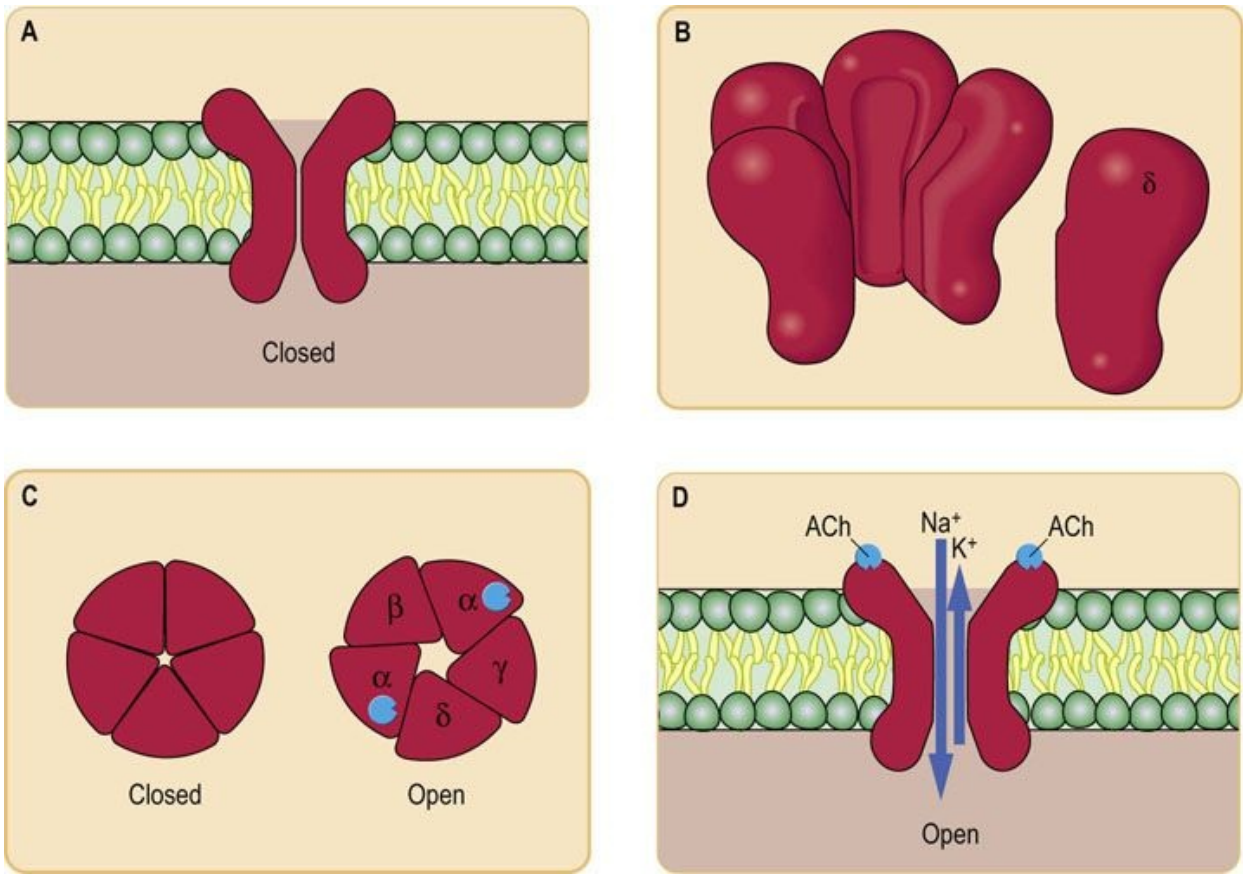


FIG. 41.1.4 Mechanism of action of ionotropic receptors.
Ionotropic receptors directly open ion channels (in fact, they are themselves ion channels). The best studied example is the nicotinic ACh receptor. This is a transmembrane protein **(A)** consisting of five nonidentical subunits **(B)**, each one passing right through the membrane. The subunits surround a pore **(C)** that selectively allows certain ions through when it is opened by a ligand **(D)**.

Metabotropic receptors

All known metabotropic receptors are coupled to G-proteins

Metabotropic receptors are coupled to second messenger pathways and act more slowly than ionotropic receptors. All known metabotropic receptors are coupled to **G-proteins** (Chapter 40) and, like hormone receptors, have seven transmembrane regions. Typically, they then couple either to adenylate cyclase, altering the production of cyclic adenosine monophosphate (cAMP), or to the phosphatidyl inositol pathway, which alters calcium fluxes. Ion channels that are

separate from the receptor are then usually modified by phosphorylation. For instance, the β-adrenergic receptor, which responds to norepinephrine and epinephrine (**Fig.** 13.5), causes an increase in cAMP, which stimulates a kinase to phosphorylate and activate a calcium channel. Some of the muscarinic class of ACh receptors have similar effects on K^+ channels.

Regulation of neurotransmitters

The action of transmitters must be halted by their removal from the synaptic cleft

When transmitters have served their function, they must be removed from the synaptic space. Simple diffusion is probably the major mechanism of removal of neuropeptides. Enzymes such as **acetylcholinesterase**, which cleaves ACh, may destroy any remaining transmitter. Surplus transmitters may also be taken back up into the presynaptic neuron for reuse, and this is a major route of removal for catecholamines and amino acids. Interference with uptake causes an increase in the concentration of transmitter in the synaptic space; this often has useful therapeutic consequences.

Concentrations of neurotransmitters may be manipulated

The effects of neurotransmitters can be altered by changing their effective concentrations or the number of receptors. Concentrations can be altered by:

- changing the rate of synthesis
- altering the rate of release at the synapse
- blocking reuptake
- blocking degradation.

Changes in the number of receptors may be involved in long-term adaptations to the administration of drugs.

Classes of neurotransmitters

Amino acids

It has been particularly difficult to prove that amino acids are true neurotransmitters; they are present in high concentrations because of their other metabolic roles, and therefore simple measurement of their concentrations did not provide conclusive evidence. Pharmacologic studies of responses to different analogues and the cloning of specific receptors finally provided the proof.

Glutamate

Glutamate is the most important excitatory transmitter in the CNS

Glutamate acts on both ionotropic and metabotropic receptors. Clinically, the receptor characterized in vitro by **N-methyl-D-aspartate (NMDA)** binding is particularly important (Fig. 41.1.5).

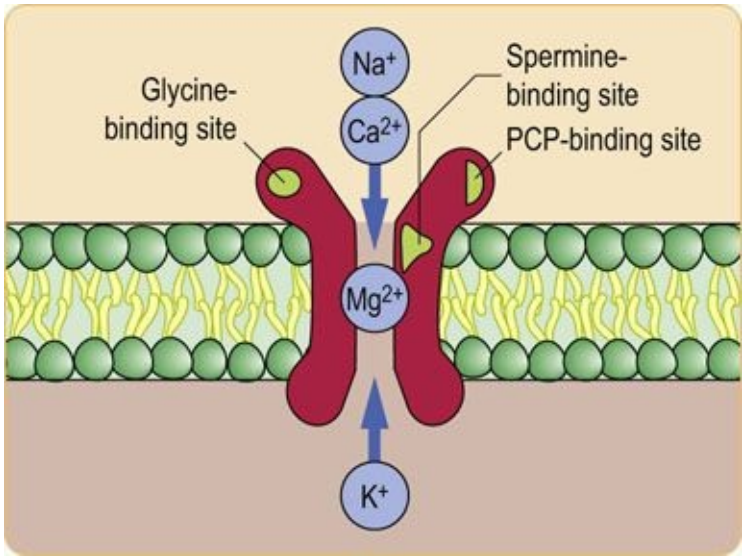


FIG. 41.1.5 The NMDA glutamate receptor.
The glutamate receptor that binds *N*-methyl-D-aspartate (NMDA) is complex. This receptor is clinically important because it may cause damage to neurons after stroke (excitotoxicity). It contains several modulatory binding sites, so it may be possible to

develop drugs that could alter its function. Glycine is an obligatory cofactor, as are polyamines such as spermine. Magnesium physiologically blocks the channel at the resting potential, so the channel can open only when the cell has been partially depolarized by a separate stimulus. It therefore causes a prolongation of the excitation. This receptor also binds phencyclidine (PCP). Because this drug of abuse can cause psychotic symptoms, it is possible that dysfunction of pathways involving NMDA receptors causes some of the symptoms of schizophrenia.

The hippocampus (Fig. 41.1.6) is an area of the limbic system of the brain that is involved in emotion and memory. Certain synaptic pathways there become more active when chronically stimulated, a phenomenon known as long-term potentiation. This represents a possible model of how memory is laid down, and it requires activation of the NMDA receptor and the consequent influx of calcium.

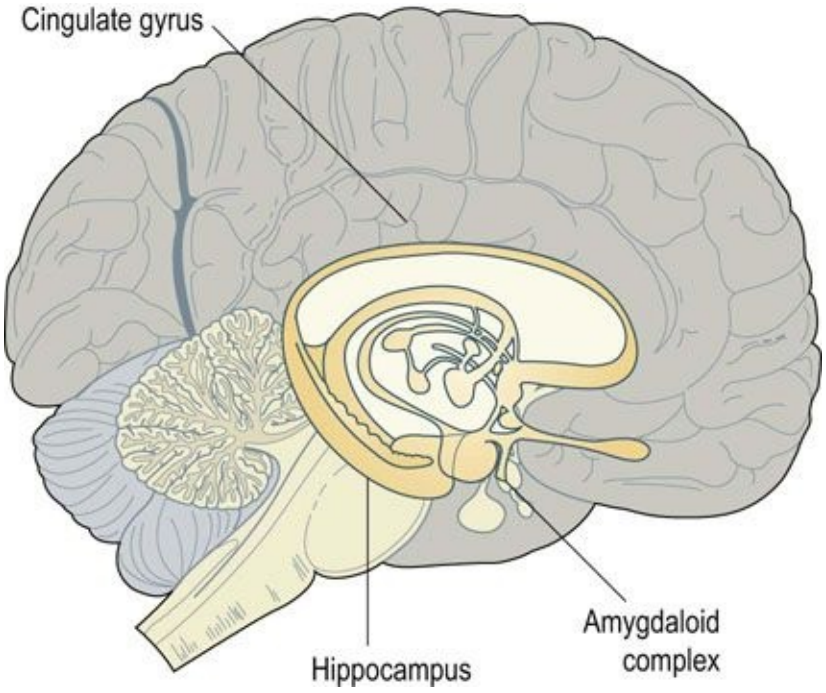


FIG. 41.1.6 Limbic system.
The limbic system of the brain is involved in emotions and memory. It consists of various areas surrounding the upper brainstem, including the hippocampus, the amygdaloid body, and the cingulate gyrus. Removal of the hippocampus prevents the laying down of short-term memory, while intact amygdaloid function is required for the emotion of fear.

Glutamate is recycled by high-affinity transporters into both neurons and glial cells. The glial cells convert it into glutamine, which then diffuses back into the

neuron. Mitochondrial glutaminase in the neuron regenerates glutamate for reuse.

Glutamate and excitotoxicity

Extracellular glutamate concentration is increased after **trauma** and **stroke**, during severe convulsions, and in some organic brain diseases such as **Huntington's chorea, AIDS-related dementia**, and **Parkinson's disease.** This is because of release of glutamate from damaged cells and damage to the glutamate uptake pathways.

Excess glutamate is toxic to nerve cells

The activation of NMDA receptor allows calcium entry into cells. This activates various proteases, which in turn initiate the pathway of programmed cell death or **apoptosis** (see Chapter 42). There may, in addition, be changes in other ionotropic glutamate receptors that also cause aberrant calcium uptake. Uptake of sodium ions is also implicated and causes swelling of cells. Activation of NMDA receptors also increases the production of nitric oxide, which may in itself be toxic. Cell death in some models of excitotoxicity can be prevented by inhibitors of nitric oxide production, but the mechanism of toxicity is not clear.

Attempts are being made to develop drugs to inhibit NMDA activation and suppress excitotoxicity. The hope is that damage caused by stroke can be limited or even reversed. Unfortunately, many of the drugs have side effects because they bind to the phencyclidine-binding site and have unpleasant psychologic effects such as paranoia and delusions.

y-Amino butyric acid (GABA)

GABA is synthesized from glutamate by the enzyme glutamate decarboxylase

GABA (Fig. 41.1.7) is the major inhibitory transmitter in the brain. There are two known GABA receptors: the $GABA_A$ receptor is ionotropic and the $GABA_B$ receptor is metabotropic. The $GABA_A$ receptor consists of five subunits that arise from several gene families, giving an enormous number of potential receptors with different binding affinities. This receptor is the target for several useful therapeutic drugs. **Benzodiazepines** bind to it and cause a potentiation of the response to endogenous GABA; these drugs reduce anxiety and also cause muscle relaxation. **Barbiturates** also bind to the GABA receptor and stimulate it

directly in the absence of GABA; because of this lack of dependence on endogenous ligand, they are more likely to cause toxic side effects in overdose.

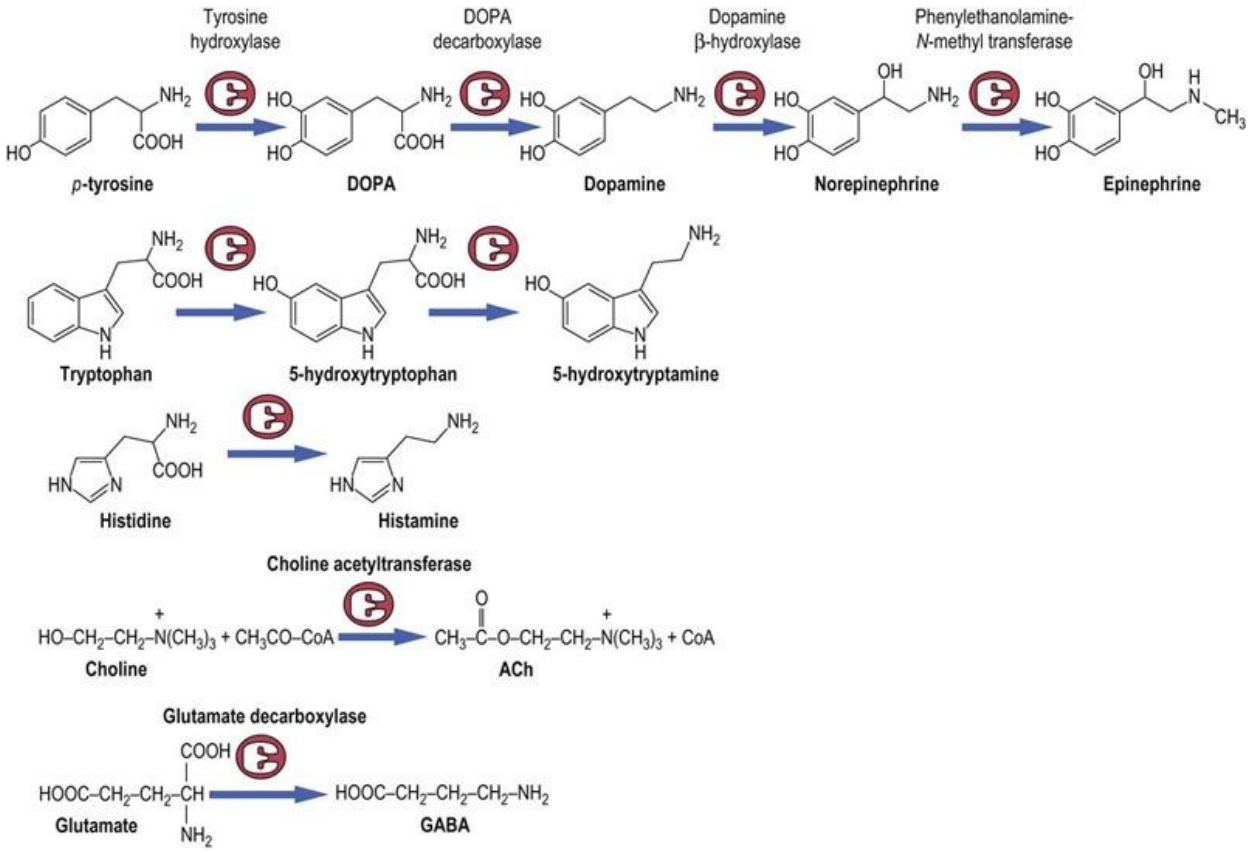


FIG. 41.1.7 Synthesis of neurotransmitters and their precursors.
The amino acid tyrosine is the precursor of dopamine,norepinephrine and epinephrine.Tryptophan is the precursor of serotonin (5-hydroxytryptamine), and histamine derives from the amino acid histidine. Choline, an amino alcohol is the precursor of acetylcholine, and the common amino acid, glutamic acid, is the precursor of the GABA.

Glycine

Glycine is primarily found in inhibitory interneurons in the spinal cord, where it blocks impulses traveling down the cord in motor neurons to stimulate skeletal muscle. The glycine receptor on motor neurons is ionotropic and is blocked by strychnine; motor impulses can then be passed without negative control, which accounts for the rigidity and convulsions caused by this toxin.

Catecholamines

Norepinephrine, epinephrine, and dopamine, known as catecholamines, are all derived from the amino acid tyrosine (Fig. 41.1.7). In common with other compounds containing amino groups, such as serotonin, they are also known as biogenic amines. Nerves that release catecholamines have varicosities along the axon, instead of a single area of release at the end. Transmitter is released from the varicosities and diffuses through the extracellular space until it meets a receptor. This allows it to affect a wide area of tissue, and these compounds are believed to have a general modulatory effect on overall brain functions such as mood and arousal.

Norepinephrine and epinephrine

Norepinephrine (also known as noradrenaline) is a major transmitter in the sympathetic nervous system

Sympathetic nerves arise in the spinal cord and run to ganglia situated close to the cord, from which postganglionic nerves run to the target tissues. Norepinephrine (Fig. 41.1.7) is the transmitter for these postganglionic nerves, whereas the transmitter at the intermediate ganglia is ACh. Stimulation of these nerves is responsible for various features of the ‘fight or flight’ response, such as stimulation of the heart rate, sweating, vasoconstriction in the skin, and bronchodilation.

There are also norepinephrine-containing neurons in the CNS, largely in the brainstem (Fig. 41.1.8). Their axons extend in a wide network throughout the cortex and alter the overall state of alertness or attention. The stimulatory effects of **amphetamines** are caused by their close chemical similarity to catecholamines.

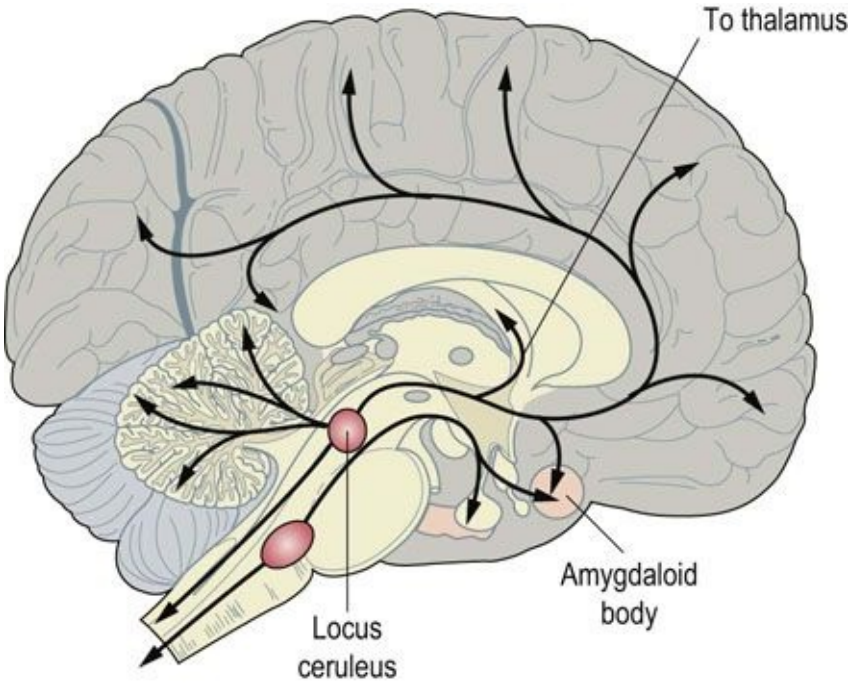


FIG. 41.1.8 Norepinephrine neurons in the CNS.
Norepinephrine-containing neurons arise in the locus ceruleus in the brainstem and are distributed throughout the cortex.

Clinical box A man with a severe headache and hypertension

A 50-year-old man had been suffering from depression for some years. His condition was treated with tranylcypromine, an inhibitor of monoamine oxidase types A and B. He developed a severe, throbbing headache and his blood pressure was found to be 200/110 mmHg. The only unusual occurrence had been that he had attended a cocktail party the previous evening at which he ate cheese snacks and drank several glasses of red wine.

Comment.

The patient was experiencing a hypertensive crisis caused by an interaction between the food he had eaten and the drug he was treated with – a MAO inhibitor. This drug inhibits the main enzyme that catabolizes catecholamines. Several foods, including cheese, pickled herring and red wine, contain an amine called tyramine, which is similar in structure to natural amine transmitters

and is also broken down by MAO. If this enzyme is not functional, the concentrations of tyramine increase and it starts to act as a neurotransmitter. This can cause a hypertensive crisis, as it did in this patient.

Epinephrine (also known as adrenaline) is produced by the adrenal medulla under the influence of ACh-containing nerves, analogous to the sympathetic preganglionic nerves

Epinephrine is more active than norepinephrine on the heart and lungs, causes redirection of blood from the skin to skeletal muscle, and has important stimulatory effects on glycogen metabolism in the liver. In response to epinephrine, a sudden extra supply of glucose is delivered to muscle, the heart and lungs work harder to pump oxygen round the circulation, and the body is then prepared to run or to defend itself (Chapter 21). Epinephrine is not essential for life, however, as it is possible to remove the adrenal medulla without serious consequences.

The receptors for norepinephrine and epinephrine are called **adrenoceptors** (see Fig. 13.5). They are divided into α-and β-receptor classes and subclasses on the basis of their pharmacology. Epinephrine acts on all classes of the receptors but norepinephrine is more specific for α-receptors. **β-Blockers**, such as atenolol, are used to treat **hypertension** and chest pain (angina) in **ischemic heart disease** because they antagonize the stimulatory effects of catecholamines on the heart. Nonspecific **α-blockers** have limited use, although the more specific β_1-blockers, such as prazosin, and α_2-blockers, such as clonidine, can be used to treat **hypertension**. Certain subclasses of β-receptors are found in particular tissues; for instance, the β_2-receptor is present in lung and **β_2-receptor agonists** such as **salbutamol** are therefore used to produce bronchial dilatation in asthma without stimulating the β_1-receptor in the heart.

Norepinephrine is taken up into cells by a high-affinity transporter and catabolized by the enzyme monoamine oxidase (MAO). Further oxidation and methylation by catecholamine-*O*-methyltransferase (COMT) convert the products to **metanephrines and vanillylmandelic acid (4-hydroxy-3-**methoxymandelic acid) (Fig. 41.1.9), which can be measured in the urine as

indices of the function of the adrenal medulla. They are particularly increased in patients who have the tumor of the adrenal medulla known as pheochromocytoma. This tumor causes hypertension because of the vasoconstrictor action of the catecholamines it produces.

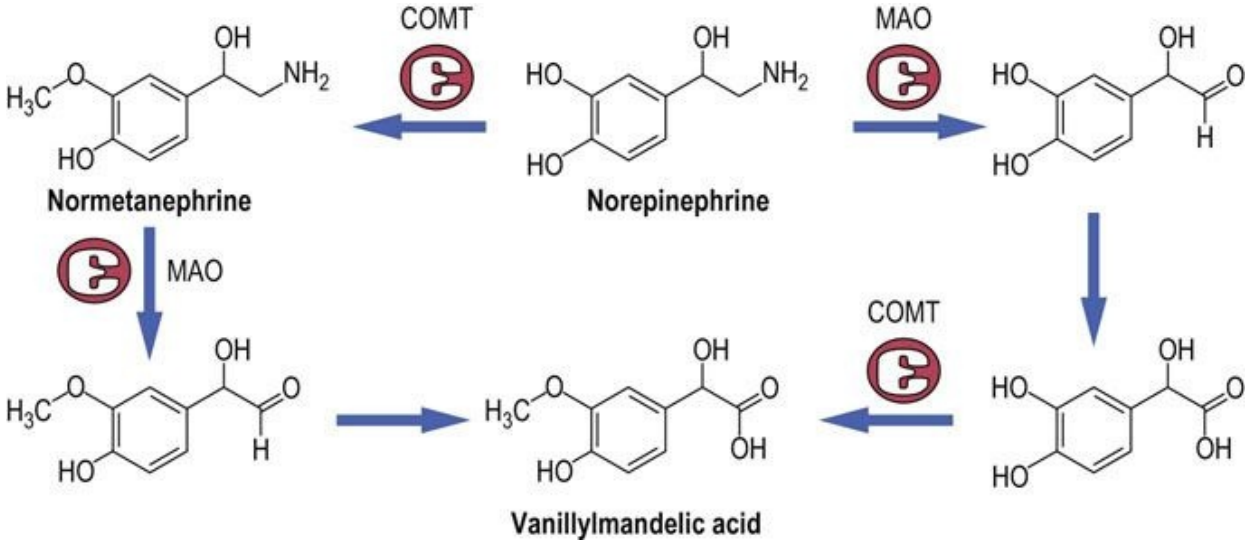


FIG. 41.1.9 Catabolism of catecholamines.
Catecholamines are degraded by oxidation of the amino group by the enzyme monoamine oxidase (MAO), and by methylation by catecholamine-*O*-methyltransferase (COMT). The pathway shown is for norepinephrine but the pathways for epinephrine, dopamine, and 5-HT are analogous.

Advanced concept box Depression as a disease of amine neurotransmitters: the antidepressants

Monoamine oxidase (MAO) inhibitors prevent the catabolism of catecholamines and serotonin. They therefore increase the concentrations of these compounds at the synapse and increase the action of the transmitters. Compounds with this property are **antidepressants**. Reserpine, an antihypertensive drug that depletes catecholamines, caused depression and is no longer in use. These findings gave rise to the ‘amine theory of depression’: this states

that depression is caused by a relative deficiency of amine neurotransmitters at central synapses, and predicts that drugs **which** increase amine concentrations should improve symptoms of the condition.

In support of this theory, tricyclic antidepressants inhibit transport of both norepinephrine and serotonin into neurons, thereby increasing the concentration of amines in the synaptic cleft. **Selective serotonin reuptake inhibitors (SSRIs),** such as fluoxetine (Prozac), are also highly effective **antidepressants**. However, as the symptoms of depression do not resolve for several days after treatment is started, it is likely that long-term adaptations of concentrations of transmitters and their receptors are at least as important as acute changes in amine concentrations in the synaptic cleft.

This role of monoamines in depression is undoubtedly an oversimplification. Thus, cocaine is also an effective reuptake inhibitor but is not an antidepressant, and amphetamines both block reuptake and cause release of catecholamines from nerve terminals, but cause mania rather than relief of depression.

Clinical box A 56-year-old woman presented with severe hypertension: pheochromocytoma

A 56-year-old woman presented with severe hypertension. She suffered from attacks of sweating, headaches, and palpitations. Her high blood pressure had not responded to treatment with an angiotensin-converting enzyme inhibitor and a diuretic. A sample of urine was taken for measurement of catecholamines and metabolites. The rate of excretion of norepinephrine was 1500 nmol/24 h (253 mg/24 h) (reference range <900 nmol/24 h, <152 mg/24 h), that of epinephrine 620 nmol/24 h (113 mg/24 h) (reference range <230 nmol/24 h, <42 mg/24 h) and that of

vanillylmandelic acid 60 mmol/24 h (11.9 mg/24 h) (reference range <35.5 mmol/24 h <7.0 mg/24 h) (see Fig. 41.1.9).

Comment.

The patient had a pheochromocytoma which is a tumor of the adrenal medulla that secretes catecholamines. Both norepinephrine and epinephrine may be secreted: norepinephrine causes hypertension by activating α_1-adrenoceptors on vascular smooth muscle, and epinephrine increases heart rate by activating β_1-adrenoceptors on the heart muscle. Hypertension may be paroxysmal and severe, leading to stroke or heart failure.

Diagnosis is made by measuring catecholamines in plasma or urine, or their metabolites, such as metanephrines and vanillylmandelic acid, in urine. The tumor is usually localized by radiologic techniques such as nuclear magnetic resonance (NMR) or computed tomography (CT) scanning.

Although this is a rare cause of hypertension, comprising only about 1% of cases, it is very important to remember it, as the condition is dangerous and often amenable to surgical cure.

Dopamine

Dopamine is both an intermediate in the synthesis of norepinephrine and a neurotransmitter

It is a major transmitter in nerves that interconnect the nuclei of the basal ganglia in the brain and control voluntary movement (Fig. 41.1.10). Damage to these nerves causes Parkinson's disease, which is characterized by tremor and difficulties in initiating and controlling movement. Dopamine is also found in pathways affecting the limbic systems of the brain, which are involved in emotional responses and memory. Defects in dopaminergic systems are implicated in **schizophrenia**, because many antipsychotic drugs used to treat this disease have been found to bind to dopamine receptors.

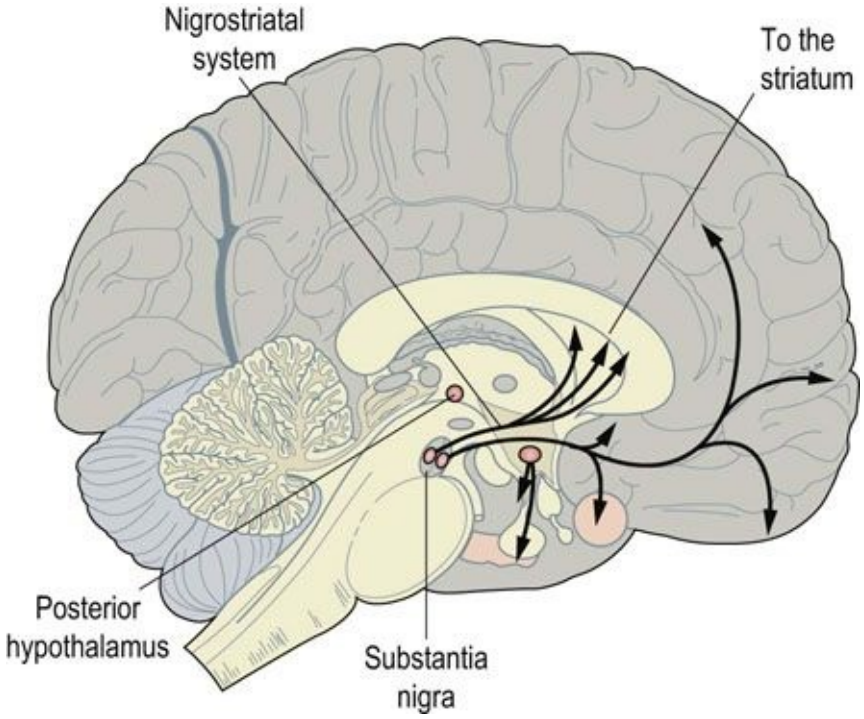


FIG. 41.1.10 Dopamine in the nigrostriatal tract.
Nerves containing dopamine run in well-defined tracts. One of the most important tracts, the nigrostriatal, connects the substantia nigra in the midbrain with the basal ganglia below the cortex. Damage to this causes Parkinson's disease, with loss of fine control of movement.

In the periphery, dopamine causes vasodilatation and it is therefore used clinically to stimulate renal blood flow, and is important in the treatment of **renal failure** (Chapter 23). The catabolism of dopamine is comparable to norepinephrine. However, the major metabolite formed is homovanillic acid (HVA).

Clinical box Tyrosine hydroxylase and aromatic amino acid decarboxylase (AADC) deficiencies: inherited causes of impaired biogenic amine metabolism

Tyrosine hydroxylase is the first step in dopamine biosynthesis, and inherited disorders affecting the activity of this enzyme result in brain dopamine deficiency. A number of clinical phenotypes

have been described and include a progressive gait disorder and infantile parkinsonism. Treatment of tyrosine hydroxylase deficiency is by the administration of L-dopa. In order to prevent the decarboxylation of L-dopa to dopamine in the blood (by peripheral aromatic amino acid decarboxylase – AADC), an inhibitor (which does not affect the activity of the brain AADC enzyme) is given at the same time as the L-dopa. Such inhibition optimizes the transport of L-dopa across the blood–brain barrier. Within the brain, AADC can then convert the L-dopa to dopamine.

AADC catalyzes the conversion of L-dopa to dopamine and 5-hydroxytryptophan to serotonin. Consequently, an inborn error of metabolism affecting the activity of this enzyme results in a brain deficiency of both dopamine and serotonin. Patients with AADC deficiency have a clinical picture that includes a severe movement disorder, abnormal eye movements and neurologic impairment. Treatment of AADC deficiency consists of preventing the degradation of any dopamine and serotonin that may be produced by residual AADC activity, *i.e.* by the use of monoamine oxidase inhibitors. In addition, dopamine agonists such as pergolide and bromocriptine are used to 'mimic' the effects of dopamine.

Clinical box Loss of activity of a dopamine transporter leads to a clinical picture suggestive of a dopamine deficiency state

Dopamine released into the synaptic cleft is taken back, via the dopamine transporter (DAT; SLC6A3) into presynaptic neurons where it can be recycled. Autosomal recessive mutations are now documented that affect the DAT. This results in an intracellular neuronal dopamine deficiency and a marked increase in extracellular levels of the neurotransmitter. **This excess dopamine is metabolized to homovanillic acid (HVA)** via non-neuronal

monoamine oxidase and catechol-O-methyltransferase. A markedly elevated CSF concentration of HVA is a strong indicator of DAT deficiency. Serum prolactin may also be elevated in this disorder. Clinically, patients with DAT mutations can present with parkinsonism-dystonia, associated with an eye movement disorder and pyramidal tract features. Currently, there is not an adequate treatment.

Clinical test box Serum hormone concentration can point to central neurotransmitter deficiency: prolactin and dopamine

Hypothalamic dopamine is an inhibitor of the release of prolactin from the pituitary. Consequently, a profound deficiency of central dopamine can lead to elevations in the concentration of serum prolactin. However, critical to the use of this peripheral biomarker is the adoption of appropriate age-related reference intervals, because,for instance, the serum prolactin concentration declines markedly during the first year of life. Whilst serum prolactin may not be elevated in all cases of central dopamine deficiency, documented elevations have been noted in the inherited disorders of tetrahydrobiopterin metabolism, and in tyrosine hydroxylase and aromatic amino acid decarboxylase deficiency states. Furthermore, correction of the central dopamine deficit can be accompanied by a lowering of the serum prolactin concentration, thereby enabling the monitoring of treatment efficacy.

Serotonin (5-hydroxytryptamine)

Serotonin, also called 5-hydroxytryptamine (5-HT), is derived from tryptophan (Fig. 41.1.7)

In addition, serotonin biosynthesis has a number of biochemical similarities to dopamine synthesis. Thus, tryptophan hydroxylase, like tyrosine hydroxylase, displays a cofactor requirement for tetrahydrobiopterin (BH_4) (see below). Furthermore, 5-hydroxytryptophan is converted to serotonin by dopa decarboxylase (also known as aromatic amino acid decarboxylase).

Serotoninergic neurons are concentrated in the raphe nuclei in the upper brainstem (Fig. 41.1.11), but project up to the cerebral cortex and down to the spinal cord. They are more active when subjects are awake than when they are asleep, and serotonin may control the degree of responsiveness of motor neurons in the spinal cord. In addition, it is implicated in so-called vegetative behaviors such as feeding, sexual behavior, and temperature control.

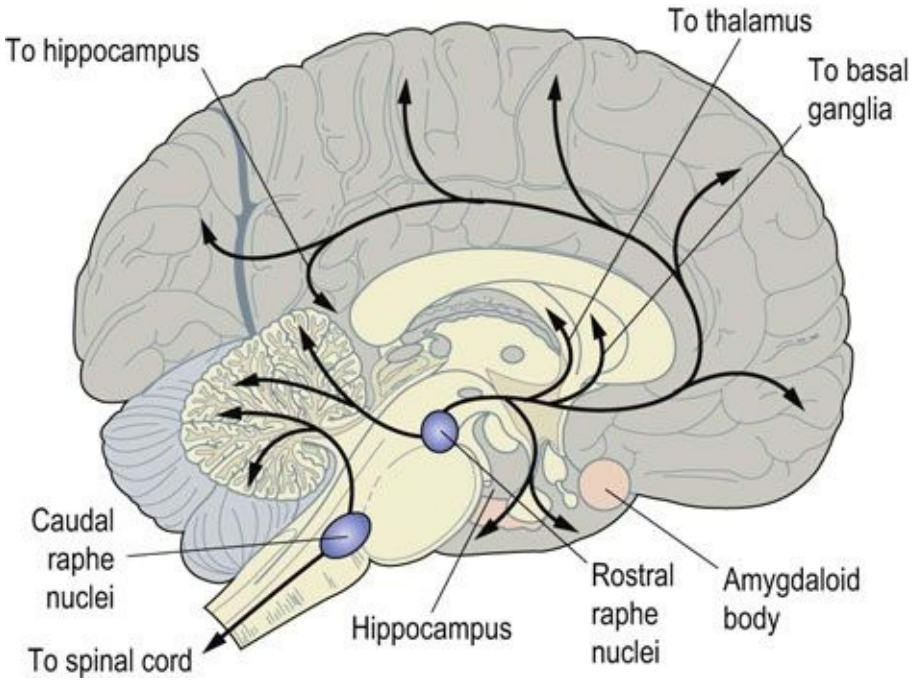


FIG. 41.1.11 Serotoninergic nerves in the CNS.
Serotonin-containing nerves arise in the raphe nuclei, part of the reticular formation in the upper brainstem. In common with those containing norepinephrine, they are distributed widely.

Advanced concept box Dopamine and serotonin receptors

Multiple receptors have been isolated for dopamine and serotonin. Not all those that have been cloned have yet been shown to be functional, but the possible relevance in terms of drug development is obvious. In some cases, specific manipulation of particular receptors can be exploited therapeutically.

There are five known dopamine receptors, falling into two main groups (**D_1-like**: D_1 and D_5, and **D_2-like**: D_2, D_3, and D_4) that differ in their signaling pathways. D_1 receptors increase the production of cAMP, whereas D_2 receptors inhibit it. Antipsychotic drugs such as phenothiazines and haloperidol tend to inhibit D_2-like receptors, suggesting that excessive dopamine activity may be important in causing the symptoms of schizophrenia.

The D_2 receptor is a major receptor in the nerves that interconnect the basal ganglia. As it is known that destruction of these nerves causes **Parkinson's disease**, it is not surprising that antipsychotic drugs that inhibit the D_2 receptor tend to have the side effect of causing abnormal movements. Drug, such as clozapine that bind preferentially to the D_4 receptor appear to be free of such side effects, although that particular drug also binds to several other receptors.

More than a dozen **serotonin (5-HT) receptors** have been isolated using molecular biological techniques. They have been divided into classes and subclasses on the basis of their pharmacologic properties and their structures. Most are metabotropic, although the 5-HT_3 receptor is ionotropic and mediates a fast signal in the enteric nervous system. The 5-HT_{1A} receptor is found on many presynaptic neurons, where it acts as an autoreceptor to inhibit the release of 5-HT.

In general, increasing the brain concentration of 5-HT appears to increase **anxiety**, whereas reducing its concentration is helpful in treating the condition. The antidepressant buspirone acts as an agonist at 5-HT_{1A} receptors, and presumably causes a decrease in

production of 5-HT. In addition to its effects on the D_4 dopamine receptor, clozapine binds strongly to the 5-HT_{2A} receptor, and it may be that a combination of a high level of 5-HT_{2A} antagonism and low D_2-binding activity is desirable for drugs that can be used to treat schizophrenia with the minimum frequency of side effects. The 5-HT_3 blocker ondansetron is an antiemetic, extensively used to prevent vomiting during chemotherapy. Migraine can be treated with sumatriptan, a 5-HT_{1D} agonist.

The central role of 5-HT in controlling brain function and the huge number of associated receptors suggest that it may possible to tailor a large number of drugs to treat specific disorders, and that pharmacologic manipulation of the function of the nervous system is probably still in its infancy.

Clinical box A 60-year-old man who suffered attacks of flushing and diarrhea: carcinoid syndrome

A 60-year-old man complained of attacks of flushing, associated with an increased heart rate. He also had troublesome diarrhea and abdominal pain, and had lost weight. The symptoms suggested a diagnosis of carcinoid syndrome caused by excessive secretion of serotonin and other metabolically active compounds from a tumor. To confirm this, a urine sample was taken for measurement of **5-hydroxyindoleacetic acid (5-HIAA),** the major metabolite of 5-HT; the concentration was found to be 120 mmol/24 h (23 mg/24 h) (reference range 10–52 mmol/24 h, 3–14 mg/24 h).

Comment.

The patient had the carcinoid syndrome, which is caused by tumors of enterochromaffin cells, usually originating in the ileum, that have metastasized to the liver. These cells are related to the catecholamine-producing chromaffin cells in the adrenal medulla

and convert tryptophan to serotonin (5-HT). Serotonin itself is believed to cause diarrhea, but other mediators, such as histamine and bradykinin, may be more important in the flushing attacks. The urinary concentration of 5-HIAA provides a useful diagnostic test and can be used to monitor the response of the cancer to treatment.

Acetylcholine

Acetylcholine (ACh) is the transmitter of the parasympathetic autonomic nervous system and of the sympathetic ganglia (Fig. 41.1.3)

Stimulation of the parasympathetic system produces effects that are broadly opposite to those of the sympathetic system, such as slowing of the heart rate, bronchoconstriction, and stimulation of intestinal smooth muscle. ACh also acts at neuromuscular junctions, where motor nerves contact skeletal muscle cells and cause them to contract. Apart from these roles, ACh may be involved in learning and memory, as neurons containing this transmitter also exist in the brain.

ACh is synthesized from choline by the enzyme choline acetyl transferase. After it is secreted into the synaptic cleft, it is largely broken down by acetylcholinesterase. The remainder is taken back up into the nerve cell by transporters similar to those for amines.

There are two main classes of ACh receptors: **nicotinic** and **muscarinic** (see Chapter 41.2, Fig. 41.2.3).

Both respond to ACh but can be distinguished by their associated agonists and antagonists; they are quite different structurally and differ in their mechanisms of action.

■ **Nicotinic receptors are ionotropic.** They bind nicotine and are found on ganglia and at the neuromuscular junction. When ACh or nicotine binds, a pore opens, which allows both Na^+ and K^+ to pass through. Because the action of the ligand on the channel is direct, action is rapid.

■ **Muscarinic receptors, responding to the fungal toxin muscarine, are metabotropic.** They are much more widespread in the brain than are nicotinic

receptors, and are also the major receptors found on smooth muscle and glands innervated by parasympathetic nerves. **Atropine** specifically inhibits these receptors. There are several separate muscarinic receptors, differing in their tissue distribution and signaling pathways. As yet, no clear pattern has emerged as to their specific functions.

Clinically, ACh agonists, in common with acetylcholinesterase inhibitors, are used to treat **glaucoma**, an eye disease characterized by high intraocular pressure, by increasing the tone of the muscles of accommodation of the eye. They are also used to stimulate intestinal function after surgery. On the other hand, when acetylcholinesterase is inhibited by **organophosphate insecticides** or **nerve gases**, a toxic syndrome is caused by the resulting excess of ACh. There may be diarrhea, increased secretory activity of several glands, and bronchoconstriction. This syndrome can be antagonized by atropine, although longer-term treatment involves the use of drugs that can remove the insecticide from the enzyme, such as pralidoxime.

Clinical box A woman with occasional double vision and a change in her voice: myasthenia gravis

A 35-year-old woman noticed that she had difficulty in keeping her eyes open. She also had periods of double vision when her voice was indistinct and nasal and she had difficulty swallowing. Her physician suspected myasthenia, a disease of nerve–muscle conduction. The serum titer of **antiacetylcholine receptor** antibodies was measured and found to be elevated.

Comment.

The patient was suffering from myasthenia gravis. This is a disease that manifests itself as weakness of voluntary muscles and is corrected by treatment with acetylcholinesterase inhibitors. It is caused by autoantibodies directed against the nicotinic acetylcholine receptor, which circulate in serum. Because of these autoantibodies, transmission of nerve impulses to muscle is much less efficient than normal.

Drugs that inhibit acetylcholinesterase increase the concentration of acetylcholine in the synaptic space, which

> compensates for the reduced number of receptors. Improvement in nerve–muscle conduction in response to edrophonium can be used as a diagnostic test but requires several precautions; long-acting acetylcholinesterase inhibitors such as pyridostigmine can be used to treat the disease, but corticosteroids are often effective.

Nitric oxide gas

In autonomic and enteric nerves, nitric oxide (NO) is produced from arginine by the tetrahydrobiopterin-dependent nitric oxide synthases

NO has a number of attributed physiologic functions including relaxation of both vascular and intestinal smooth muscle, and the possible regulation of mitochondrial energy production. Furthermore, within the brain, NO may have a role in memory formation. However, excessive NO formation has been implicated in the neurodegenerative process associated with **Parkinson's** and **Alzheimer's** disease. Whilst the exact mechanism whereby excessive NO causes neuronal death is not known, a growing body of evidence suggests that irreversible damage to the mitochondrial electron transport chain may be an important factor.

NO is not stored in vesicles, but released directly into the extracellular space

Consequently, NO does not, in the strictest sense, meet all the current criteria to be labeled as a neurotransmitter. NO itself diffuses comparatively easily between cells, and binds directly to heme groups in the enzyme guanylate cyclase, stimulating the production of cyclic guanosine monophosphate.

Other small molecules

ATP and other purine-containing molecules derived from it are now known to have transmitter functions

ATP is present in synaptic vesicles of sympathetic nerves, along with norepinephrine, and is responsible for rapid excitatory potentials in smooth muscle. Adenosine receptors are widespread in the brain and in vascular tissue. Adenosine is largely inhibitory in the CNS, and inhibition of adenosine receptors is believed to underlie the stimulatory effects of **caffeine**.

Study of histamine in nerves is complicated by the large amounts that are present in mast cells

Histamine is found in a small number of neurons, mainly in the hypothalamus, although their projections are widespread throughout the brain. It has been shown to control the release of pituitary hormones, arousal, and food intake. **Antihistamines** designed to control **allergies** caused by release from mast cells act on the H_1 receptor and tend to be sedative, suggesting that other central functions also probably exist. The **histamine receptor in the stomach** is of the H_2 class; therefore, the H_2 inhibitors, such as **cimetidine** and **ranitidine**, that are used to treat peptic ulcers have no effect on allergy.

Clinical box Deficiency of pyridoxal phosphate: a cause of neonatal epilepsy

Pyridoxal phosphate (PLP), the biologically active form of vitamin B_6 Chapter 11), is utilized as a cofactor by more than 100 enzymes, including reactions catalyzed by aromatic amino acid decarboxylase (AADC), threonine dehydratase and the glycine cleavage system. Vitamin B_6 is present in the human body as a number of ‘vitamers’ that are precursors to PLP. A pivotal enzyme in the formation of PLP is pyridox(am)ine-5′-phosphate oxidase (PNPO). This enzyme catalyzes the conversion of the precursors pyridoxine phosphate and pyridoxamine phosphate to PLP. Deficiency of PNPO results in decreased availability of PLP and, such patients, when presenting in the neonatal period, have a clinical picture that includes severe epilepsy. Biochemical analysis of CSF reveals elevated threonine, glycine and evidence of impaired AADC activity. In addition, the CSF concentration of PLP is decreased. Treatment, which can be particularly effective, is

by the administration of PLP.

Peptides

Many peptides act as neurotransmitters

It is an open question whether all the peptides that have been described are really true neurotransmitters. Nevertheless, more than 50 small peptides have now been shown to influence neural function. All known peptide receptors are metabotropic and coupled to G-proteins (Chapter 40), and so act comparatively slowly. There are no specific uptake pathways or degradative enzymes, and the main route of disposal is simple diffusion followed by cleavage by a number of peptidases in the extracellular fluid. This allows a peptide to affect a number of neurons before it is finally degraded.

Vasoactive intestinal peptide (VIP) is one of many peptides that affect the function of the intestine through the enteric nervous system. It was originally described as a gut hormone that affected blood flow and fluid secretion, but it is now known to be an important enteric neuropeptide, inhibiting smooth muscle contraction. It also causes vasodilatation in several secretory glands, and potentiates stimulation by ACh.

Many neuropeptides belong to a multigene family

The **opioid peptides** and opioid receptors provide a good example of a multigene family. They are the endogenous ligands for opiate analgesics such as morphine and codeine. The control of pain is complex, and opioid peptides and receptors are found both in the spinal cord and in the brain itself. There are at least three genes that code for these peptides, and each contains the sequences for several active molecules:

- **proopiomelanocortin** contains β-endorphin, which binds to opiate μ-receptors, and also adrenocorticotropic hormone (ACTH) and the melanocyte-stimulating hormones (MSH), which are pituitary hormones (Chapter 39);
- **proenkephalin A** contains the sequences for Met-and Leu-enkephalins, which bind to δ-receptors and are involved in pain regulation at local levels in the brain and spinal cord;

■ **prodynorphin** contains sequences for dynorphin and several other peptides, which bind to the κ-class of receptors.

Opiates affect pleasure pathways in the brain, which explains their euphoriant effects, and they also have side effects, such as respiratory depression, that limit their use. In excess, they cause contraction of the muscles of the eye, resulting in 'pinpoint' pupils. It has been shown that **endorphins** are released after strenuous exercise, giving the so-called **'jogger's high'.** It is hoped that increased knowledge of the specific opioid receptors and neural opioid pathways will allow the development of analgesics with fewer side effects and less likelihood of abuse.

Substance P is another example of a member of a multigene family, known as the tachykinin family. It is present in afferent fibers of sensory nerves and transmits signals in response to pain. It is also involved in so-called neurogenic inflammation stimulated by nerve impulses, and is an important neurotransmitter in the intestine.

Neuropeptides can act as neuromodulators

Some peptides do act as true neurotransmitters, but they also have many other actions. They often alter the action of other transmitters, acting as neuromodulators, but have no action of their own. For instance, VIP enhances the effect of ACh on salivary gland secretion in cat submandibular glands (glands located under the jawbone) by causing vasodilatation and potentiating the cholinergic component. NPY causes inhibition of the release of norepinephrine at autonomic nerve terminals, acting at presynaptic autoreceptors, and potentiates the action of norepinephrine in certain arteries while having only weak actions itself. Opioid peptides also are capable of modulating neurotransmitter release.

Summary

■ Neurons communicate at synapses by means of neurotransmitters.
■ A large number of compounds, whether of low molecular weight, such as the biogenic amines, or larger peptides, can act as neurotransmitters.
■ They act on specific receptors and there is normally more than one receptor for each neurotransmitter.
■ The presence of several transmitters in the same nerves and the identification of multiple receptors suggest that there is a high degree of flexibility and complexity in the signals that can be produced in the nervous system.

Active learning

1. Does nitric oxide meet all the criteria to be defined as a true neurotransmitter?
2. Explain how a neurotransmitter such as serotonin can have so many diverse effects within the central nervous system.
3. Explain how neurotransmitters can be excitatory or inhibitory.
4. What neurotransmitter types are likely to become deficient in the brain of a patient with an inborn error affecting tyrosine hydroxylase, aromatic amino acid decarboxylase and tetrahydrobiopterin metabolism?
5. Discuss the factors that need to considered when establishing a diagnostic method for disorders of dopamine and serotonin metabolism.
6. Explain the concept of ionotropic and metabotropic receptors.

Further reading

Aitkenhead, H, Heales, SJ. Establishment of paediatric age-related reference intervals for serum prolactin to aid in the diagnosis of neurometabolic conditions affecting dopamine metabolism. *Ann Clin Biochem*. 2013; 50:156–158.

Clayton, PT. B6-responsive disorders: a model of vitamin dependency. *J Inherit Metab Dis*. 2006; 29:317–326.

Hyland, K. Inherited disorders affecting dopamine and serotonin: critical neurotransmitters derived from aromatic amino acids. *J Nutr*. 2007; 137(6 suppl 1):1568S–1572S.

Kurian, MA, Zhen, J, Meyer, E, *et al*. Clinical and molecular characterisation of hereditary dopamine transporter deficiency syndrome: an observational cohort and experimental study. *Lancet Neurol*. 2011; 10:54–62.

Lam, AAJ, Hyland, K, Heales, SJR. Tetrahydrobiopterin availability, nitric oxide metabolism and glutathione status in the hph-1 mouse: implications for the pathogenesis and treatment of tetrahydrobiopterin deficiency states. *J Inherit Metab Dis*. 2007; 30:256–262.

Websites

The AADC Research Trust. www.aadcresearch.org.

Databases of PKU and PND variations this website hosts databases of pediatric neurotransmitter disorders (PND). www.BioPKU.org.

The PND Association is a disease organization representing children and families who are affected by a pediatric neurotransmitter disease. www.pndassoc.org.

CHAPTER 41.2

Neurochemistry

Hanna Bielarczyk, Andrzej Szutowicz and Edward J. Thompson

Learning objectives

After reading this chapter you should be able to:

- Describe the cellular components of the central nervous system.
- Discuss the function of the blood–brain barrier in health and disease.
- Describe the basic principles of neuronal signaling and receptors.
- Describe catecholamine, cholinergic, glutamatergic and GABAergic transmission.
- Describe the role of ion channels in nerve transmission.
- Comment on the role of sodium, potassium and calcium ions in nerve transmission.
- Discuss the process of vision as an example of a chemical process underlying neuronal function.

Introduction

The brain is, in many ways, a chemist's delight. This is so because it illustrates various general principles of biology applied to a highly specialized tissue that ultimately regulates all the other tissues of the body. This chapter highlights the differences between the central nervous system – that is, the brain and spinal cord – and the peripheral nervous system, which is outside the dura (the thick fibrous covering that contains the cerebrospinal fluid (CSF)).

Brain and peripheral nerve

The distinction between brain and peripheral nerve essentially reflects the division between the central nervous system (CNS) and the peripheral nervous system (PNS): a convenient dividing line being the confines of the dura, within which watertight compartment is the CSF, partially produced (about one-third of the total volume) through the action of the blood–brain barrier. Myelin insulates the axons of nerves; the chemical composition of CNS myelin is quite distinct from that of PNS myelin, not least because the two forms are produced by two different types of cells: the oligodendrocytes within the CNS and the Schwann cells within the PNS. The distinction between the functions of the CNS and those of the PNS is fundamental to differential diagnosis in neurology. A typical example is the difference between the demyelination of the CNS that occurs **in multiple sclerosis,** and the demyelination of the PNS that occurs **in Guillain–Barré syndrome.**

The blood–brain barrier

The term blood–brain 'barrier' is a slight misnomer, in that the barrier is not absolute but relative: its permeability depends on the size of the molecule in question

Initially, experiments based upon use of a dye (Evans blue) bound to albumin showed that, over a period of hours, an animal progressively turned blue in all tissues, with the notable exception of the brain, which remained white. It subsequently became clear that 1 molecule in 200 of serum albumin passed normally into the CSF, which is analogous to lymph. It also became obvious that, for any given protein, the ratio of its concentrations in CSF and serum was a linear function of the molecular radius of the molecules in solution.

There are a total of six sources of the CSF

Under normal and pathologic conditions, proteins pass from these cellular or tissue sources into the CSF, and their degree of filtration and/or rate of local synthesis vary. The total quantity of the CSF therefore constitutes the algebraic summation of these six sources (Fig. 41.2.1).

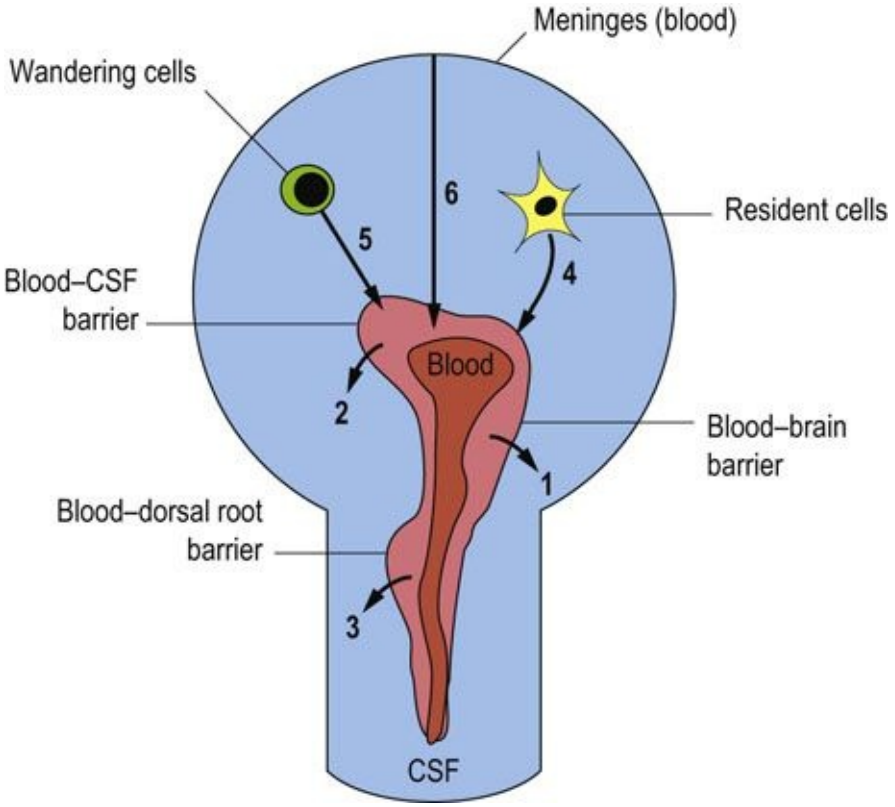
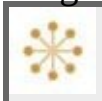


FIG. 41.2.1 The six main sources of cerebrospinal fluid (CSF).
The involved processes comprise passage across barriers (from the blood, *i.e.* 1, 2, 3) and direct sources of local production (CNS cells, *i.e.* 4, 5, 6).

■ **The blood–brain barrier** (the parenchymal capillaries) gives rise to about one-third of the volume of CSF, and is known as the interstitial fluid source.
■ **The blood–CSF barrier** provides the bulk of CSF (almost all of the remaining two-thirds), termed choroidal fluid because it is principally provided by the choroid plexi (capillary tufts) situated in the lateral ventricles and, to a lesser degree, the plexi situated in the third and fourth ventricles.
■ **The dorsal root ganglia** contain capillaries that have a much greater degree of permeability.
■ **The brain parenchyma of the CNS** produces a number of brain-specific proteins. These include prostaglandin synthase (formerly called β-trace protein), and transthyretin (a protein formerly called prealbumin).
■ **The CSF circulating cells**, mainly lymphocytes within the CNS, synthesize local antibodies. However, in the CNS there is strong presence of immune suppressor cells. Because of this, in brain infections such as **meningitis,** steroids are given in addition to antibiotics, to suppress the potentially devastating effects, within this confined space, of inflammation associated with the

intrathecal immune response.

■ **The meninges** represent a sixth source of CSF under pathologic conditions; they can give rise to dramatic increase in the concentrations of CSF proteins.



Advanced concept box Differential diagnosis of nasal discharge (rhinorrhea)

It is essential, on clinical grounds, to distinguish CSF rhinorrhea from local nasal secretions caused by, say, influenza infection. The ENT surgeon must know whether the fluid present is CSF, as any leak must be surgically repaired, lest it remain a chronic potential source of meningitis as a result of the migration of nasal flora into the subarachnoid space. One characteristic and useful marker protein in the CSF is **asialotransferrin**, which is transferrin lacking sialic acid. In the systemic circulation, this absence of sialic acid gives a molecular signal for the protein to be recycled, and it is thus immediately removed from the systemic circulation by all reticuloendothelial cells. The brain has no true reticuloendothelial cells along the path of CSF flow, and hence asialotransferrin is present in quite high concentrations. The aqueous humor of the anterior chamber of the eye also produces the characteristic asialotransferrin, and the same asialotransferrin can also be found in the perilymph of the semicircular canals in the inner ear.

Clinical box A 65-year-old man with progressive weakness of limbs: the guillain–barré syndrome

Three weeks after an acute diarrheal illness, a 65-year-old man presented with progressive ascending weakness of the limbs followed by respiratory muscle weakness requiring assisted

ventilation. On examination, he was hypotonic and areflexic, with profound general weakness. Isoelectric focusing of CSF and parallel serum samples showed a similar abnormal pattern of oligoclonal bands in both.

Comment.

This predominantly motor neuropathy is Guillain–Barré syndrome and the patient has antibodies developed as a result of infection with the bacterium *Campylobacter jejuni*. The organism contains the antigen, ganglioside sugar GM_1, which is shared with a ganglioside on peripheral nerves. Antibodies bind to peripheral motor nerves and cause the neuropathy. It is an example of **molecular mimicry**.

Clinical box A 75-year-old woman with dizziness, intermittent diarrhea and numbness in both feet: amyloidosis

A 75-year-old woman complained of postural dizziness, dry mouth, intermittent diarrhea, and numbness in both her feet. On examination, there was a marked decrease in blood pressure on assuming the upright posture. A chest radiogram revealed lytic lesions in the sternum. Her urine contained Bence Jones protein. A bone marrow examination demonstrated increased numbers of plasma cells (see also Chapter 4).

Comment.

Her neurologic condition was caused by amyloidosis in which the free light-chain component of myeloma globulin produced by tumor of plasma cells in the bone marrow accumulates in peripheral nerves. The light chains adopt the configuration of a β-pleated sheet, with multiple copies that are intercalated and

resistant to normal proteolysis.

It is essential to make a diagnosis of a disease which relapses and remits since the patient may show no abnormalities at the time of physical examination by the clinician. Lumbar puncture therefore plays a major role with the demonstration of **oligoclonal bands in CSF**, which are absent from the parallel serum specimen. This means that there is an intrathecal rather than a systemic immune response. The converse is seen in, *e.g.* neurosarcoidosis, in which the systemically synthesized immunoglobulins are transferred passively into the spinal fluid, giving rise to a so-called 'mirror' pattern, where the oligoclonal bands are the same in both CSF and serum. The test involves isoelectric focusing of CSF with a parallel serum sample. The separated immunoglobulins are exposed to anti-IgG to identify bands which are present in CSF but absent from the corresponding serum. Such patterns indicate local synthesis of IgG within the brain.

Cells of the nervous system

Fewer than 10% of the cells of the nervous system are large neurons. The three major cell types in the nervous system (which each constitute about 30%) are **astrocytes**, which also make up part of the blood–brain barrier; **oligodendrocytes,** which are principally composed of fat and serve to insulate the axons; and **microglia**, which are essentially resident macrophages (scavengers).

These different cell types are associated with predominant protein molecules that are important in various brain pathologies (Table 41.2.1). Other minor constituents of the nervous system include the ependymal cells, which are ciliated cells secreting brain-specific proteins such as prostaglandin synthase. The brain endothelial cells, unlike other tissue capillaries, have tight junctions that bind them together; this feature is believed also to contribute to the blood–brain barrier, although it is the basement membrane which is the major source of molecular sieving of the different-sized proteins.

Table 41.2.1

The different cells of the CNS, and their protein markers indicating brain pathologies

Cell	Protein	Pathology
Neuron	Neuron-specific enolase	Brain death
Astrocyte	GFAP	Plaque (or scar)
Oligodendrocyte	Myelin basic protein	De/remyelination
Microglia	Ferritin	Stroke
Choroid plexi	Asialotransferrin	CSF leak (rhinorrhea)

GFAP, glial fibrillary acidic protein.

Neurons

The significant features of neurons are their length, their many interconnections, and the fact that they do not divide postpartum

There is an archetypal notion of the electrical activity of the nervous system – in particular, of the electrical activity of neurons. However, three other biological

features of neurons are particularly worthy of note: their length, their prolific interconnections, and the fact that they do not divide postpartum.

Because of their great length, neurons depend upon an efficient system of axonal transport

Neurons can typically be 1 m long; thus the nucleus, the source of information for the synthesis of neurotransmitters, is typically quite remote from the synaptic terminal, the site of release of those transmitters. Because of this extensive length, a crucial requirement is the neuron's ability to transport material both from the nucleus towards the synapse (anterograde transport) and from the synapse to the nucleus (retrograde transport). Neurons have evolved special characteristics to deal with this separation of their two functional sites, and to maintain electrical activity at the nodes of Ranvier (Fig. 41.2.2).

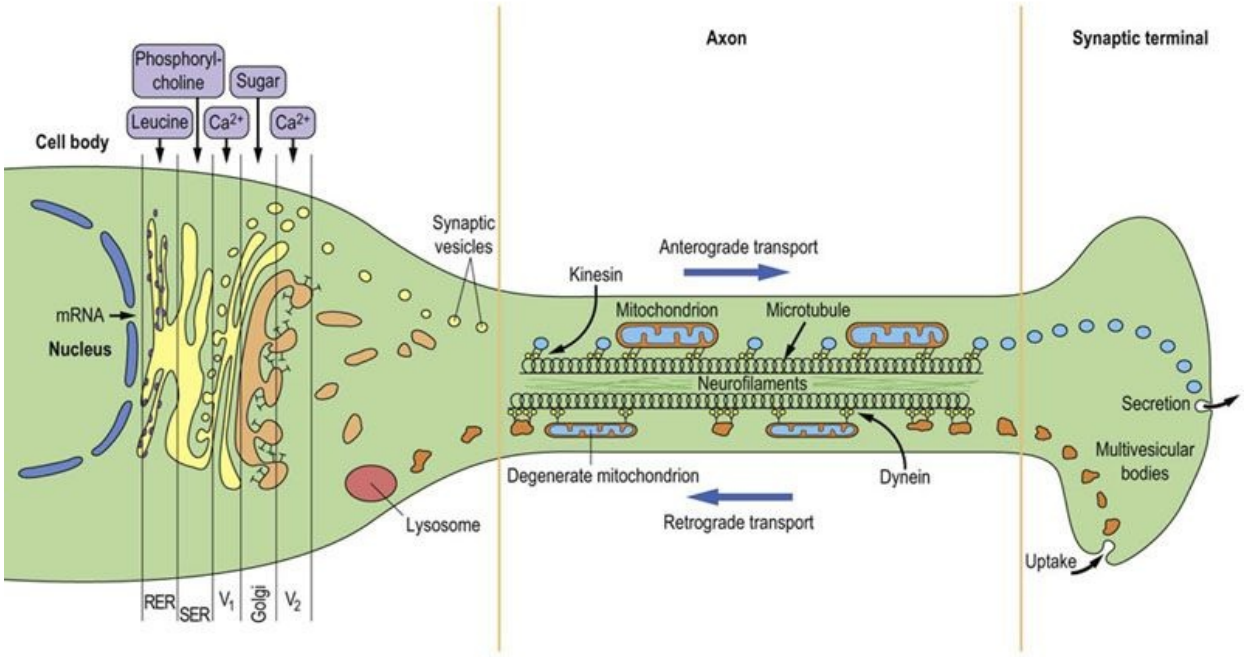


FIG. 41.2.2 The functional structure of a neuron.
Within the cell body, there is specialized movement through the Golgi stack by the components required to form synaptic vesicles (V_1, V_2). In the axon, there is fast axonal transport along microtubules via the motile proteins, kinesin (in anterograde transport) or dynein (in retrograde transport). RER, rough endoplasmic reticulum; SER, smooth endoplasmic reticulum.

The normal ‘resting’ movement within the axon is mediated by separate **molecular ‘motors’** (motile proteins): kinesin in the case of anterograde transport and dynein in retrograde transport. The materials being transported in each direction are also rather different, and the different components of axonal structure shown in Figure 41.2.2 possess the capacity for different speeds of transportation. During growth, a separate form of transport (toward the synapse) occurs that takes place at the rate of about 1 mm/day; this flow constitutes bulk movement of the building blocks such as the filamentous proteins.

Neurotransmission is an energy-demanding process

Brain constitutes 2% of body mass, yet under resting conditions it responsible for 20% of the overall glucose consumption. Glucose is almost exclusive energy substrate for the brain. The brain does not utilize fatty acids for energy production. Under certain conditions it may utilize β-hydroxybutyrate or lactate derived from circulation. Large amount of energy is required to maintain plasma membrane resting potentials of the neurons, that are subject to continuous (10–60 Hz frequency) depolarization cycles (action potentials). To meet these demands, over 70% of brain energy is produced and utilized by neurons. Astrocytes utilizing glucose release significant amounts lactate to extracellular brain compartment, and it serves as complementary energy source for the neurons. However, neither exogenous nor endogenous lactate, can fully replace glucose as the principal energy source.

Neuroglial structures

Astrocytes and oligodendrocytes comprise the neuroglial structures

In the cortex, or gray matter, one typically finds a protoplasmic astrocyte with one set of processes surrounding the endothelial cells, thereby helping to ‘filter’ materials from the blood, and a separate set of processes surrounding the neurons, which are thereby being ‘fed’ selected substances that have been extracted from the blood for passage to the neurons.

In the white matter, the astrocytes have a rather more fibrous appearance and have more of a structural role. When there is injury to the CNS, astrocytes can play a major part in the reaction, synthesizing large amounts of the glial

fibrillary acidic protein (GFAP). This is the cellular equivalent of scar tissue and is found in diseases such as **multiple sclerosis**, in which it is the major constituent of the characteristic plaques. Astrocytes are not present in the PNS.

The oligodendrocytes present in the CNS can wrap round as many as 20 axons, forming the myelin sheath that insulates these neuronal processes from one another and stops cross-talk between neurons. There is also intense oligodendrocyte mitochondrial activity at the nodes of Ranvier, which are parallel to the sites of depolarization within the underlying axon. In the PNS, the Schwann cells form the myelin and, typically, wrap round only a single axon.

Synaptic transmission

One of the unique chemical characteristics of the brain is the massively high density of synapses between different neurons

Thus a locally acting neurohormone is released by one axon onto many other cell bodies. On the receiving end, a given cell body will typically receive myriad cellular products via its profusely branched dendritic tree: each branch can be smothered in **synapses**. The 'neurotransmitter' (neurohormone) is released by the axon's nerve terminals of the first neuron onto the dendrite of the second neuron or a nonneuronal cell. This is mediated by a neurotransmitter **receptor** on the respondent cell. There is usually a **second messenger,** such as a cyclic nucleotide, which may stimulate protein phosphorylation. Typically, **G-proteins** are found just under the neurotransmitter receptor protein spanning the cell membrane, where they act to 'couple' the first messenger (e.g. norepinephrine) to a second messenger (e.g. cyclic AMP, cAMP) (Chapter 40). Some neurotransmitter receptors are coupled with **ion channels.**

Neurotransmitters are normally inactivated after their actions on the target cell. Hydrolysis is a major mechanism by which this is achieved. The best-studied example is that of the enzyme acetylcholinesterase. There can also be regulation at the level of the second messenger, such as cAMP, which is broken down by the enzyme phosphodiesterase. Phosphodiesterase is inhibited by caffeine and other methylxanthines, and they thereby mimic many of the effects of adrenergic neurotransmission.

Synaptic transmission involves the recycling of membrane components

In addition to release of a specific neurohormone, there is also an extensive system for recycling of membrane constituents associated with this process. The synaptic vesicles contain a very high concentration of the relevant neurotransmitter, which is bounded by a membrane (Chapter 42.1). During synaptic release of the transmitter, there is fusion of the synaptic vesicle membrane (containing the neurotransmitter) with the presynaptic membrane. This increase in total membrane mass is redressed by invagination of the lateral

aspects of the nerve terminals, where an inward puckering movement of the membrane is effected by contractile movements of the protein clathrin. There then follows a form of pinocytosis of the excess membrane, which is transported in retrograde fashion toward the nucleus, to be digested in lysosomes.

Types of synapse

Because of the multitude of different synaptic inputs to a given neuron, the final algebraic summation results in a 'decision' at the level of the axon hillock (the site of origin of the axon from the cell body) as to whether or not to transmit an action potential down the axon as an all-or-nothing phenomenon. However, even before this decision is made, the input of a particular neurotransmitter can essentially be classified as **excitatory** or **inhibitory**.

In addition to the relatively short-term decisions concerning action potentials (Chapter 42.1), there is a longer-term modulation of the resting membrane potential, moving it either closer to (excitation) or further from (inhibition) the critical membrane potential, which is the level at which the resting membrane potential will finally trigger an action potential at the axon hillock. Many drugs have a longer-term effect on modulation, in addition to the short-term effect, which partially explains their addictive effect; this can be seen with alcohol or the opioid drugs. There are also long-term effects during treatment with various drugs: for example, those used to treat endogenous depression, such that it may be weeks before any beneficial effects are seen.

Cholinergic transmission

The best-studied neurotransmitter is acetylcholine

Acetylcholine (ACh) is synthesized in the cytoplasmic compartment of cholinergic nerve terminals from acetyl-CoA and choline by choline acetyltransferase. The enzyme is expressed exclusively in the cholinergic neurons. Acetyl-CoA is synthesized from pyruvate derived from glycolysis, whereas choline is taken up from extracellular compartment by plasma membrane-potential-driven high-affinity choline uptake system.

As a model system, this transmitter can have two rather different effects, depending upon its site of origin within the nervous system (i.e. central or peripheral). The effects originally demonstrated by experiments with **nicotine** are characteristic of the nicotinic receptors, whereas those demonstrated with

muscarine characterize the muscarinic receptors. The nicotinic type of transmission is exerted by motor neurons located in the brainstem and anterior horns of the medulla oblongata. Another group of central cholinergic neurons located in the brain septum plays a key role in the basic and higher cognitive functions through activation of postsynaptic muscarinic receptors. There is a complex picture of the agonists and antagonists associated with the regional actions of ACh (Fig. 41.2.3). The classic antagonist of the muscarinic effect is **atropine** and the best-studied blocker for the nicotinic receptor is the poisonous snake venom **α-bungarotoxin**.

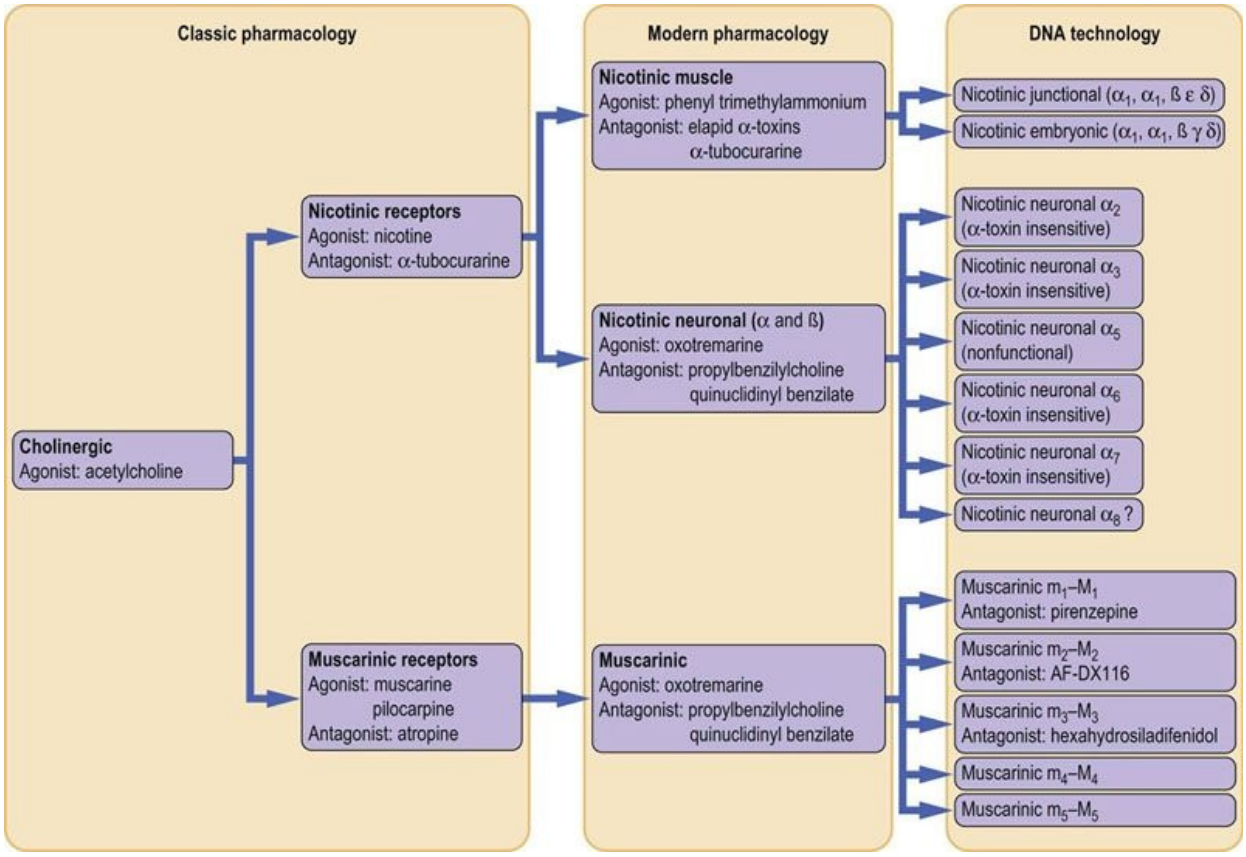


FIG. 41.2.3 The history of naming: acetylcholine agonists and antagonists.
The changes in nomenclature, for the agonists and antagonists of the different central (neuronal) versus peripheral (muscle) regional actions of acetylcholine (ACh).

In **Alzheimer's disease,** amyloid-β(1–42), in combination with other neurotoxic factors, causes preferential impairment of cholinergic neurons in the brain septum, yielding progressive loss of cognitive function, which leads to

dementia. In the early stages of this disease, the inhibitors of acetylcholinesterase with M_2 receptor agonist properties improved cognitive functions but had no effect on disease progress. The antagonists of glutamatergic NMDA receptors are employed to reduce excitotoxic effects of excessive activation of glutamatergic neurons.

In **myasthenia gravis,** autoantibodies are formed against the nicotinic receptor. However, by blocking the hydrolysis of ACh, for example by means of the drug edrophonium (which inhibits acetylcholinesterase), the concentration of ACh can be effectively increased (Chapter 42.1).

Peripheral cholinergic neurons are located in parasympathetic ganglia and innervate all visceral tissues. They dilate blood vessels of the gatrointestinal tract and enhance salivation and peristalsis. They also constrict airways, control heart function, constrict the pupils and regulate lens accommodation, and stimulate sexual arousal and genital erection.

Catecholamine transmission

Catecholamines, epinephrine and norepinephrine are synthesized from L-tyrosine in the sequence of reactions catalyzed by tyrosine hydroxylase/aromatic amino acid decarboxylase, then dopamine β-hydroxylase and phentolamine-*N*-methyltransferase, yielding dopamine, norepinephrine, and epinephrine, respectively. Dopamine is a precursor of norepinephrine and epinephrine.

Dopamine is a transmitter in the brain dopaminergic neurons located in its several areas, including substantia nigra. They are involved in reward-driven learning, regulation of mood, attention, learning, and prolactin release through different classes of dopamine receptors ($D_{1–5}$). Disturbances of dopamine metabolism are associated with several central nervous system pathologies including **Parkinson's disease, schizophrenia,** and **restless legs syndrome. To** overcome dopamine deficits in some of these diseases L-DOPA, its precursor, is administered, as it easily crosses blood–brain barrier. Dopamine is given to patients in **shock and heart failure** to elevate cardiac output and increase blood pressure and renal filtration. Several drugs, including amphetamines, cocaine, and nicotine, exert their behavioral and addictive effects through excessive stimulation of the release and the increase of dopamine level in the synaptic cleft. They also stimulate serotoninergic and norepinephrinergic transmission in the brain.

Norepinephrine and epinephrine are synthesized in the brain and peripheral

sympathetic ganglia by respective groups of neurons acting as neurotransmitters. On the other hand, catecholamines released from chromaffin cells into circulation exert endocrine effects. In the brain, they exert regulatory functions in decision-making processes. Peripherally, they increase blood pressure (they cause vasoconstriction and increase the rate and force of cardiac muscle contraction), cause bronchial and pupil dilatation, inhibit peristalsis, increase sweating and renin secretion, and promote ejaculation. **Their actions are mediated through two separate receptors: α-adrenergic receptor, blocked by phentolamine, and β-adrenergic receptor, blocked by propranolol.** The latter drug was commonly used by cardiologists (other β-blockers are the mainstay of treatment in coronary heart disease), but neurologists also use it as part of the treatment of Parkinson's disease. Many adrenergic effects are mediated by cAMP (Chapter 42.1).

The action of catecholamines is terminated by their reuptake and degradation to aldehydes by mitochondrial monoamine oxidases and subsequent methylation by catechol-*O*-**methyltransferase to** homovanillic or vanillylmandelic acids, which are excreted with urine. Excess of these compounds in urine may indicate the presence of adrenal medullar tumor, **pheochromocytoma**.

Glutamate: glutamatergic transmission

Depending on the brain region, 50–80% of the neuronal population is glutamatergic

The mean L-glutamate level in the brain is in the range of 5–10 mmol/L. It is synthesized from α-ketoglutarate by glutamate dehydrogenase and aminotransferases or from glutamine by phosphate-activated glutaminase. The L-glutamate/glutamate–zinc complex is taken up by synaptic vesicles of glutamatergic presynaptic nerve terminals, where it reaches concentrations exceeding 100 mmol/L. Glutamate is released upon depolarization, transiently reaching high concentrations in the synaptic clefts. Its binding to different classes of receptors, including NMDA (the principal one), causes depolarization/activation of postsynaptic recipient neurons. Glutamatergic receptor stimulation is subject to multiple regulations, which play an important role in synaptic plasticity, termed long-term potentiation. This phenomenon takes place in the hippocampus and in the different regions of brain cortex, and is involved in learning, memory formation, and other cognitive functions.

Glutamate is quickly taken up from the synaptic space by specific transporters expressed mainly on the adjacent astroglial cells. There, glutamine synthetase converts glutamate to glutamine, which is subsequently transported back to glutamatergic neurons.

Excessive glutamate release or its impaired uptake, which takes place, among others, in ischemia, hypoglycemia and exposure to neurotoxic xenobiotics, may cause its excessive accumulation in the extracellular space. This in turn causes prolonged depolarization of the recipient cells and, consequently, excitotoxic injury. **Epilepsy** is the pathologic condition caused by excessive glutamate release by pathologically stimulated glutamatergic neurons and/or the deficiency of inhibitory GABAergic transmission (see below).

y-Aminobutyric acid (GABA): GABAergic transmission

GABA is the chief inhibitory neurotransmitter in the brain

The inhibitory effect of GABA on postsynaptic neurons results from binding to specific $GABA_A$ receptors. GABA concentration remains in the range of 4–6 mmol/L. It is the ligand for gated chloride channels. Their opening upon GABA binding causes flow of Cl^- ions into the neuron, causing its hyperpolarization and inhibition of transmitter function. GABA is synthesized by L-glutamate decarboxylase present in the cytoplasm of GABAergic neurons. GABA action is terminated mainly by its uptake by presynaptic terminals through high-affinity GABA transporter. GABA may then be either loaded again into vesicles or metabolized to succinate – a TCA cycle intermediate. Several $GABA_A$ receptor agonists and GABA uptake or GABA-transaminase inhibitors are used as sedatives, tranquilizers or anxiolytic drugs. The most common groups include **barbiturates, benzodiazepines, chloral hydrate and valproate.** Ethanol also acts as the $GABA_A$ receptor agonist.

Ion channels

Even at rest, the neuron is working to pump ions along ionic gradients

The ‘resting’ neuron is, nevertheless, continually pumping sodium out of the cell and potassium in, through ion channels. During an action potential, there is a momentary reversal of these ionic movements: sodium enters the cell and

potassium then leaves, effectively repolarizing the resting membrane (Chapter 41.1). Mutations of sodium channels can occur at different sites and give rise to **hyperkalemic periodic paralysis**. The negative ion chloride moves through separate channels, which are implicated in specific pathologic states such as **myotonia**.

Calcium ions have an important role in the synchronization of neuronal activity

The movement of calcium ions within cells often provides a 'trigger' for the cells to synchronize an activity such as synaptic release of neurotransmitter; this synchronization also has a prominent role in the sarcoplasmic reticulum of muscle (Chapter 20). Within the central nervous system, the **Lambert–Eaton syndrome** is a disease that affects predominantly the P/Q subtype of calcium channels, in an example of **molecular mimicry.** The patient may have a primary oat cell carcinoma of the lung; the immune system responds by making antibodies against these malignant cells. However, the malignant cells and the calcium channels possess a common epitope, the effect of which is that the immune response causes the release of neurotransmitter to be blocked at the presynaptic site. This is analogous to, but nevertheless can be clearly distinguished from, the condition in **myasthenia gravis,** in which the block is postsynaptic.

It is also worth noting that blockade of the presynaptic release of neurotransmitter may be usefully exploited by therapeutic application of botulinum toxin (a protein derived from anerobic bacteria), which contains enzymes to hydrolyze the presynaptic proteins involved in release of neurotransmitters. This toxin is used in special cases of spasticity such as **torticollis**, in which the patient can be relieved of the excessive contractures of the neck muscles, which turn the head chronically to one side and thus cause pain and distraction if untreated.

Clinical box An 18-year-old man with weakness of arms and legs: familial periodic paralysis

An 18-year-old male awoke in the night with intense weakness of the proximal muscles of his arms and legs. Before retiring he had consumed a meal of pasta and cake. His brother and father had

previously been similarly affected. He was taken to the emergency room of the local hospital where the weak limbs were noted to be hypotonic with depressed tendon reflexes. Serum concentration of potassium was mildly reduced at 2.9 mmol/L (normal 3.5–5.3). By the next day he had fully recovered and serum potassium had risen spontaneously to normal levels. A further attack of paralysis was induced by an infusion of intravenous glucose, thus confirming a diagnosis of familial periodic paralysis.

Comment.

Different molecular lesions at various sites of the sodium channel pores can give rise to hyperkalemic periodic paralysis. As the name suggests, the patient has intermittent muscle weakness, during which time the serum potassium concentration is increased. This is caused by an imbalance of cationic movements in which sodium enters the cell and potassium leaves it. In these patients, the abnormal flux of sodium into the muscle is not correctly regulated with its counterflux of potassium ions (see also Chapter 8). Hypokalemic periodic paralysis is inherited as a Mendelian dominant and results from a mutation in the gene encoding the L-type calcium channel. Genetic diseases that affect ion channel function are called **channelopathies** (see also Chapter 8).

Clinical box A woman with progressive blurred vision, dysphagia and limb weakness: botulism

Twenty-four hours after eating home-preserved vegetables, a healthy young woman experienced progressive onset of blurred vision, severe vomiting, dysphagia, and advancing limb weakness starting in the shoulders. Her doctor admitted her to hospital, and electrophysiologic studies confirmed the clinical diagnosis of botulism. Trivalent antiserum, made from inactivated toxin, was administered immediately and, with the help of assisted

ventilation, the patient recovered within a few weeks.

Comment.

The vegetables contained the exotoxin of the anaerobe, *Clostridium botulinum* which had not been destroyed during the preservation process. The toxin hydrolyzes the presynaptic proteins involved in the release of neurotransmitter, and thus the blockade is similar to the functional lesion in Lambert–Eaton myasthenic syndrome; however, in botulism the blockade can be lethal, especially at the level of the phrenic nerve, which is essential for appropriate respiratory lung movement.

Mechanism of vision

The mechanism by which the human eye can detect a single photon of light provides a fascinating example of the chemical processes underlying neuronal function

The mechanism of vision involves both trapping of photons and the transducer effect, whereby the energy of light is converted into a chemical form, which is then ultimately transmuted into an action potential by a retinal ganglion neuron. A number of the intermediates are as yet not precisely known, but the underlying hypothesis is that the receptor protein, **rhodopsin**, is coupled to the G-protein. There are several sequence homologies of rhodopsin with the adrenergic β-receptor and with the muscarinic ACh receptor. The main steps (Fig. 41.2.4) take place in the following order):

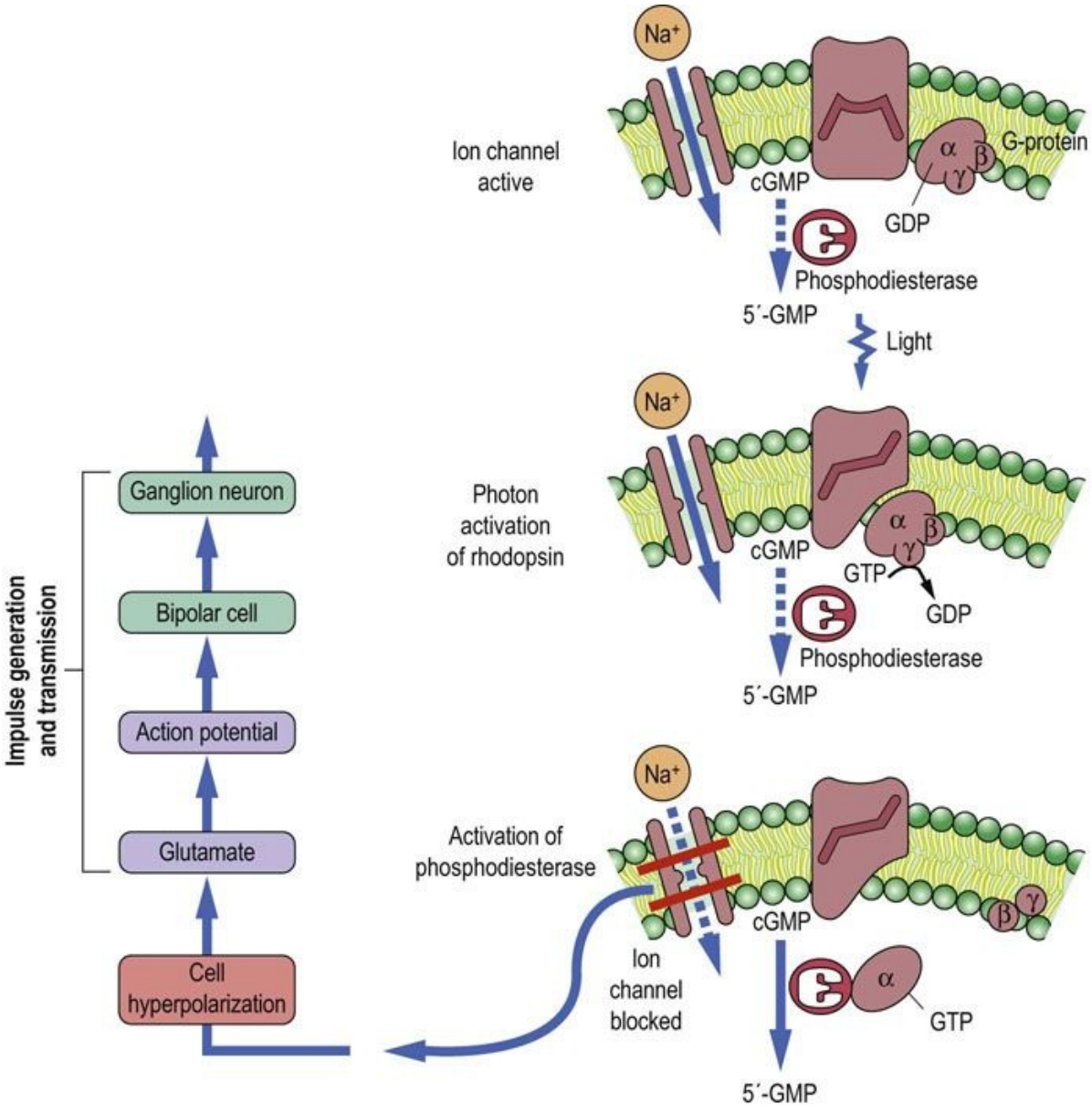


FIG. 41.2.4 Neurochemistry of synaptic transmission in the mechanism of vision. The figure shows the consequences of photon activation of rhodopsin via G-protein coupling in a rod cell. Phosphodiesterase is activated and hydrolyzes the second messenger, cGMP, thereby blocking the entry of sodium and causing hyperpolarization of the cell. Currently, the steps through which neurotransmission subsequently proceeds to produce the final action potential in the ganglion neuron are not known in detail.

Compare the activities of the G-protein coupled receptors. Dotted lines indicate an inactive process.

- *Cis*-retinal is converted to *trans*-retinal.
- Rhodopsin becomes activated.
- The level of cGMP decreases.
- Na^+ entry into the cell is blocked.
- The rod cell hyperpolarizes.
- There is release of glutamate (or aspartate).
- An action potential depolarizes the adjacent bipolar cell.
- This depolarizes the associated ganglion neuron, to send an action potential out of the eye.

Summary

■ The nervous system contains a number of distinct cells, each of which synthesizes its own individual proteins.
■ The specialized functions of the nervous system mean that these proteins are effectively compartmentalized in different loci.
■ In order to facilitate communication within the brain, there are two specialized methods of moving cells, organelles and proteins: the cerebrospinal fluid and axonal transport.
■ The blood–brain barrier is diverse in anatomic origin and is not absolute but relative (specifically based on molecular size of the transferred molecules).
■ Synthesis of antibodies within the CSF, with no parallel presence in plasma is unequivocal evidence for the brain as the source of antigenic stimulation.
■ Neurotransmitters in the brain are synthesized and released into synaptic clefts from axonal terminals of specific groups of neurons (glutamatergic, GABAergic, catecholaminergic, cholinergic etc).
■ The activity of neurons depends on the function of ion channels both in the central nervous system and the peripheral nerves.

Active learning

1. G-proteins are widely used throughout the body as ‘coupling’ agents between the first extracellular messenger and the second intracellular messenger. Discuss some of the roles for which the G-protein has been adapted amongst various cell types.
2. Mitochondria play an important role in providing for the metabolic requirements at the nodes of Ranvier, along the considerable length of the axon. Discuss the role of the two molecular motors in recycling the mitochondria required to support this function.
3. Chloride is an important anion which forms part of the complex movements with cations such as sodium and potassium during depolarization. Discuss the consequences of congenital abnormalities in chloride transport.
4. Describe the reactions that occur in the process of vision.
5. Give an example of molecular mimicry.

Further reading

Barry, DM, Millecamps, S, Julien, JP, *et al.* New movements in neurofilament transport, turnover and disease. *Exp Cell Res*. 2007; 313:2110–2120.

Bettens, K, Sleegers, K, Van Broeckhoven, C. Genetic insights in Alzheimer's disease. *Lancet Neurol*. 2013; 12:92–104.

Bos, JL, Rehmann, H, Wittinghofer, A. GEFs and GAPs: critical elements in the control of small G proteins. *Cell*. 2007; 129:865–877.

Cannon, SC. Physiologic principles underlying ion channelopathies. *Neurotherapeutics*. 2007; 4:174–183.

de Leon, MJ, Mosconi, L, Blennow, K, *et al.* Imaging and CSF studies in the preclinical diagnosis of Alzheimer's disease. *Ann N Y Acad Sci*. 2007; 1097:114–145.

George, DR, Whitehouse, PJ, D'Alton, S, *et al.* Through the amyloid gateway. *Lancet*. 2012; 380:1986–1987.

Honig, LS. Translational research in neurology. Dementia. *Arch Neurol*. 2012; 69:969–977.

Stein-Streilein, J, Taylor, AW. An eye's view of T regulatory cells. *J Leukoc Biol*. 2007; 81:593–598.

CHAPTER 42

Cellular Homeostasis

Cell Growth and Cancer

Alison M. Michie, Verica Paunović and Margaret M. Harnett

Learning objectives

After reading this chapter you should be able to:

- Define the stages of the mammalian cell cycle.
- Outline how the cell cycle is regulated by cyclins and cyclin-dependent kinases.
- Describe the molecular events that enable growth factors to regulate cellular proliferation.
- Discuss the different mechanisms that enable cells to cease proliferation or die.
- Define the cellular and molecular events that define apoptosis and autophagy.
- Explain how subversion of normal physiologic growth can lead to the development of tumors.
- Distinguish between oncogenes and tumor suppressor genes and describe their role in tumor progression/suppression.

Introduction

The development and survival of multicellular organisms such as human beings is reliant on the appropriate regulation of growth, differentiation and death of individual cell types to maintain the integrity of the organism

While cells have evolved a complex set of control mechanisms to deter replication of damaged cells and to enable repair, if the growth control mechanisms become damaged, this can lead to the development of cancer. The majority of cells in an adult are not dividing. However, under certain conditions such as tissue repair and senescence, regulated growth and proliferation are promoted. Thus, as cells die, either by senescence or as a result of tissue damage, they must be replaced in a strictly regulated manner. Cellular homeostasis maintains organ integrity through controlled cell survival, proliferation and cell death, to ensure that healthy cells, unlike cancerous (transformed) cells, generally stop dividing when they contact neighboring cells.

Research in transformed cancerous cells has highlighted the important mechanisms that regulate cellular growth and cell division in normal cells

Through investigation of genetic alterations in cancer cells, it has been possible to identify a large number of genes that are critical for the regulation of normal cell proliferation. It is perhaps unsurprising to find that molecular mechanisms that favor cell survival and proliferation processes are commonly upregulated in human cancers. Mutated proliferative genes are called **oncogenes** (cancer-causing genes), the normal cellular counterparts of which are called **proto-oncogenes**. Proto-oncogenes are predominantly signal transducers that act to regulate normal cell growth and division; aberrant regulation of these processes leads to cellular transformation. Conversely, the proteins involved in suppressing proliferation or **tumor suppressor genes** are generally inhibited during oncogenesis, leading to uncontrolled proliferation. Exploring the role of these proteins under normal physiologic conditions can assist in the understanding of how they can be subverted when they become dysregulated during oncogenesis.

Cell cycle

Individual cells multiply by duplicating their contents and then dividing into two daughter cells

Cell division is tightly controlled by a complex mechanism called cell cycle. The duration of cell cycle varies between organisms, as well as between cell types within a single organism. In mammals, for example it can last from minutes to years. However, in immortalized cell lines, which are widely used as experimental model systems, one round of cell cycle is typically completed within 24 hours.

In recent years, extensive research of the cell cycle has defined a number of key control points

Traditionally, the cell cycle is divided into several phases (Fig. 42.1). Mitosis (**M phase**) is the stage of cell division, which is usually completed within an hour. The remainder of the cell cycle, during which cells prepare for division and duplicate deoxyribonucleic acid (DNA), is referred to as interphase. Nuclear DNA replication occurs in the synthesis (**S**) phase of interphase. The period between M and S phase is called **G1** phase, while the interval between S and M phase is called **G2** phase. During G1 and G2 phases, cells undergo several checkpoints to ensure appropriate cellular growth and accurate DNA synthesis occurs prior to cell division, thus preventing the incorporation of mutated DNA into daughter cells. As mentioned above, the duration of an individual cell cycle varies enormously and most of this variability can be attributed to the different length of the G1 phase. This is due to the fact that some cells, which are not stimulated to duplicate their DNA, can enter into the specialized form of G1 phase, called **G0** phase.

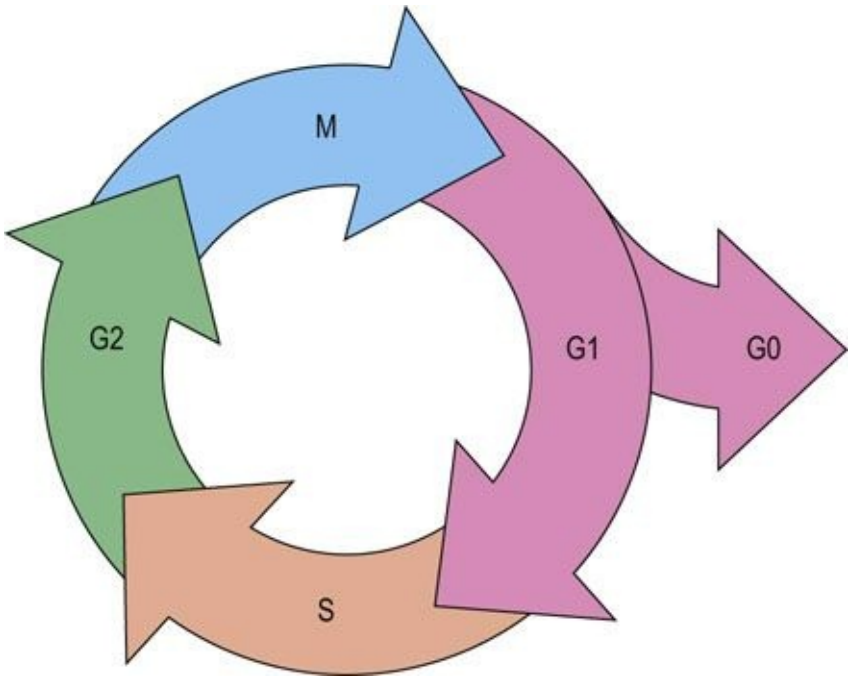


FIG. 42.1 The phases of the cell cycle.
Cell cycle is divided into interphase and mitotic phase. In interphase, composed of G1, S and G2 phases, cells grow, prepare for division and duplicate their DNA. Mitosis is the stage of cell division into two daughter cells. M, mitosis; S, synthetic phase of interphase; G1, interval between M and S phases; G2, interval between S and M phases; G0, resting, or quiescent, phase.

The G0 phase is a form of the resting state, or quiescence, in which cells reside until they receive appropriate signals: for example, from growth factors, urging them to re-enter and progress through the cell cycle

In mammals, the time required for a cell to transit from the beginning of S phase through mitosis is typically 12–24 hours, irrespective of the duration of G1 phase. Thus, the majority of variation in proliferation rates observed between different cell types is due to the amount of time spent in the G0/G1 phase. In conditions that favor cell growth, the total ribonucleic acid (RNA) and protein content of the cell increases continuously, except in M phase when the chromosomes are too condensed to enable transcription to occur.

Of note, the majority of cells in the human body have irreversibly withdrawn from the cell cycle into either a terminally differentiated state (neurons, myocytes or surface epithelial cells of skin and mucosa), or they are in the

reversible quiescent G0 phase (stem cells, glial cells, hepatocytes or thyroid follicular cells). Only a minority of cells are normally actively cycling and these cells are located mainly in the stem/transit compartments of self-renewing tissues, which include the bone marrow and epithelia.

Regulation of cell proliferation and growth: growth factors

Cells of a multicellular organism have to receive positive signals in order to grow and divide

Many of these signals are in the form of polypeptide hormones (e.g. insulin), growth factors (e.g. epidermal growth factor, EGF) or cytokines (e.g. interleukins IL-1 to IL-36). These growth factors bind to specific cell surface receptors, initiating an intricate network of intracellular signaling cascades that counteract negative regulatory controls present in resting cells to block cell cycle progression and division.

In most cell types, proliferation is controlled by signals generated from a specific combination of growth factors rather than stimulation by a single growth factor

In this way, a relatively small number of growth factors can selectively regulate the proliferation of many cell types. In addition, some of the factors may induce cell growth without providing a signal for division. Indeed, neurons in the G0 phase of cell cycle grow very large without dividing. Moreover, although proliferating cells can stop growing when they are deprived of growth factors, they continue their progress through the cell cycle until they reach the point in the G1 phase at which they can enter G0 phase (the resting state) or undergo senescence.

Growth factors bind to their specific cell surface receptors. There are about 50 known growth factors

Growth factors bind to their specific cell surface receptors expressed on effector cells, which are generally transmembrane proteins with a growth factor (or ligand) binding domain and cytoplasmic protein tyrosine kinase (PTK) domain. There are about 50 known growth factors, of which platelet-derived growth factor (PDGF) was the first identified. The growth and proliferative responses to

PDGF are prototypic for many growth factors, and these responses include:
■ Immediate increase in intracellular Ca^{2+} levels – indicating the initiation of transmembrane signaling.
■ Reorganization of actin stress fibers – to enable anchorage dependence of cell attachment, a requirement for cell cycle progression.
■ Activation and/or nuclear translocation of transcription factors that bind to regulatory regions of the DNA encoding genes responsive to a specific growth factor. These genes are referred to as immediate early genes, which usually code for transcription factors themselves that mediate expression of components of the cell cycle machinery, such as cyclins.
■ DNA synthesis and cell division.

Growth factors initiate selective signaling cascades

Individual growth factors activate distinct cohorts of signaling molecules and transcription factors that in turn induce a unique repertoire of gene expression. In this way specific growth factors initiate characteristic differential responses that have a unique impact on cell behavior.

Growth factors initiate selective signaling cascades by binding to their receptors

The binding to a receptor causes receptor dimerization or oligomerization and activation of the intracellular tyrosine kinase domain, which in turn mediates receptor transphosphorylation at specific amino acids within the cytoplasmic domain. The phosphorylated region of the receptor is then able to act as a **‘docking site’** for the binding of specific proteins, thus enabling protein–protein interactions (Fig. 42.2). This in turn leads to the recruitment and activation of additional signaling molecules such as enzymes and **‘adapter’ molecules,** which mediate the propagation of the intracellular signaling cascade from the surface of plasma membrane inside the cell. Transphosphorylation of receptor cytoplasmic domains creates a **scaffold for binding of signal transduction elements** such as phospholipase Cγ (PLC-γ), GTPase-activating proteins (GAPs), nonreceptor PTKs (Src, Fyn Abl), phosphotyrosine phosphatases (PTPases) and adapter molecules (Shc or Grb2), via their phosphotyrosine recognition domains.

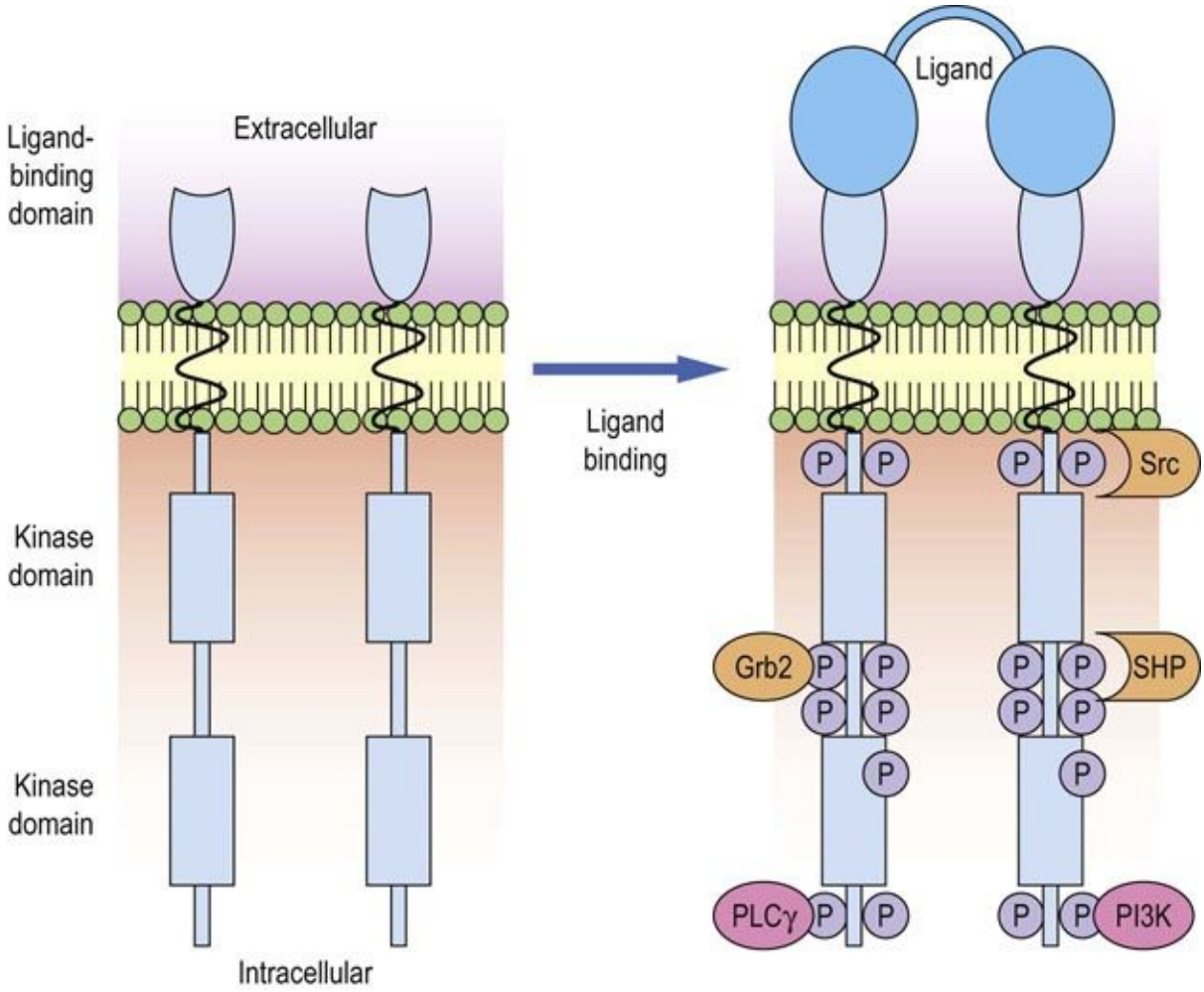


FIG. 42.2 Activation of growth factor receptor by ligand binding and recruitment of signaling molecules.
The binding of a growth factor (such as PDGF or EGF) to its receptor causes receptor dimerization and activation of the tyrosine kinase, intrinsic to the receptor cytoplasmic domains. This leads to the tyrosine phosphorylation of the dimerized receptors at specific sites within the cytoplasmic domains, by the process of transphosphorylation. The phosphorylation events create docking sites that enable protein–protein interactions between the receptor and downstream signaling components such as PLC-γ, protein tyrosine kinase Src, tyrosine phosphatase SHP, PI3K and adapter molecules like Grb2, which in turn recruits Ras/MAPK pathway. PLC-γ, phospholipase Cγ; PI3K, phosphatidylinositol 3-kinase; Grb2, growth factor receptor-bound protein 2; SHP, SH2 domain-containing phosphatase.

Advanced concept box protein tyrosine kinases

Role in signal transduction.

Protein tyrosine kinases (PTKs) are enzymes that transfer the γ-phosphate group of ATP to tyrosine residues on target substrate proteins. The term ‘protein tyrosine kinase’ is a generic term for a large superfamily of enzymes, which includes both transmembrane-spanning receptors with an intrinsic tyrosine kinase activity in their cytoplasmic domains (e.g. some growth factor receptors), as well as a wide range of cytoplasmic tyrosine kinase subfamilies such as Src, Abl, Syk, Tec or Janus kinase (JAK) families.

Tyrosine phosphorylation is a covalent modification, which provides a rapid and reversible (by the action of protein tyrosine phosphatases) mechanism of changing the enzymatic activity of the target proteins, or modifying these proteins, so that they can act as adapters to recruit other signaling molecules.

For example, the tyrosine phosphorylation of receptors or signaling molecules creates ‘docking sites’ that enable protein–protein interactions via protein–protein interaction domains, such as those referred to as SH2 domains of other signaling transducers. SH2 stands for Src-homology region 2, from the cytoplasmic *Src* tyrosine kinase, in which this domain was first characterized. SH2 domains contain approximately 100 amino acids and specifically recognize a phosphotyrosine in the context of the three amino acids immediately *C*-terminal to that phosphotyrosine.

Role in the regulation of cellular proliferation, survival and differentiation.

PTKs play a critical role in the regulation of cellular proliferation, survival and differentiation, and the importance of these regulatory events is highlighted by the defects that are observed upon dysregulation of genes encoding PTKs. Indeed, dysregulated expression of growth factor receptors carrying intrinsic PTK activity can lead to the constitutive activation of the Ras/Raf/MEK/ERK and Ras/PI3K/Akt/mTor signaling pathways leading to enhanced cell growth, survival and proliferation and a subversion of molecular events that regulate apoptosis, events that are dysregulated in cancer. Mutations have been identified in

PDGFR, EGFR, Kit and Flt3, in specific cancer types. Indeed, mutation in the EGFR family of receptors are responsible 30% of all epithelial cancers, including lung and brain cancers.

Nonreceptor tyrosine kinases also play critical roles in cellular responses, with mutations leading to a loss in kinase activity, resulting in serious abnormalities in B and T lymphocyte development. For example, loss of expression/activity of ZAP-70, a PTK that is essential for antigen-dependent T cell activation, can lead to **severe combined immunodeficiency (SCID)**, due to a lack of T cell effector function during an immune challenge. Similarly, **X-linked agammaglobulinemia,** an immunodeficiency caused by a lack of IgG antibody production occurs as a result of loss of function mutations of Btk, a PTK that is important for B cell effector functions.

Epidermal growth factor (EGF) receptor signaling

Upon ligand binding, the EGF receptor (EGFR) leads to the activation of signals through phosphatidylintositol-, Ras/Raf/MAPK-and the PI3K/Akt/mTor-mediated signaling cascades

The EGF receptor (EGFR) can be bound and activated by a number of ligands including EGF and transforming growth factor-α (TGF-α). Ligation of the EGFR leads to the recruitment and activation of Src family of PTKs, which catalyze phosphorylation of PLC-γ, leading to the activation of this enzyme. Activated PLC-γ then catalyzes the hydrolysis of phosphatidylinositol 4,5-bisphosphate (PIP_2) to generate the intracellular second messengers inositol 1,4,5-trisphosphate (IP_3) and diacylglycerol (DAG). IP_3 stimulates release of Ca^{2+} from intracellular stores (mainly ER) and DAG activates an important signal transducer family of proteins, the protein kinase C (PKC) family. Ligation of the EGFR also induces activation of another lipid modifying enzyme, phosphatidylinositol 3-kinase (PI3K). This enzyme mediates phosphorylation of

PIP_2, generating the lipid second messenger phosphatidylinositol 3,4,5-trisphosphate (PIP_3), which contributes to the activation of certain members of the PKC family (Fig. 42.3). Moreover, PIP_3 can activate another kinase referred to as PIP_3-dependent kinase (PDK1) or serve as docking site for proteins that contain so-called pleckstrin homology (PH) domains.

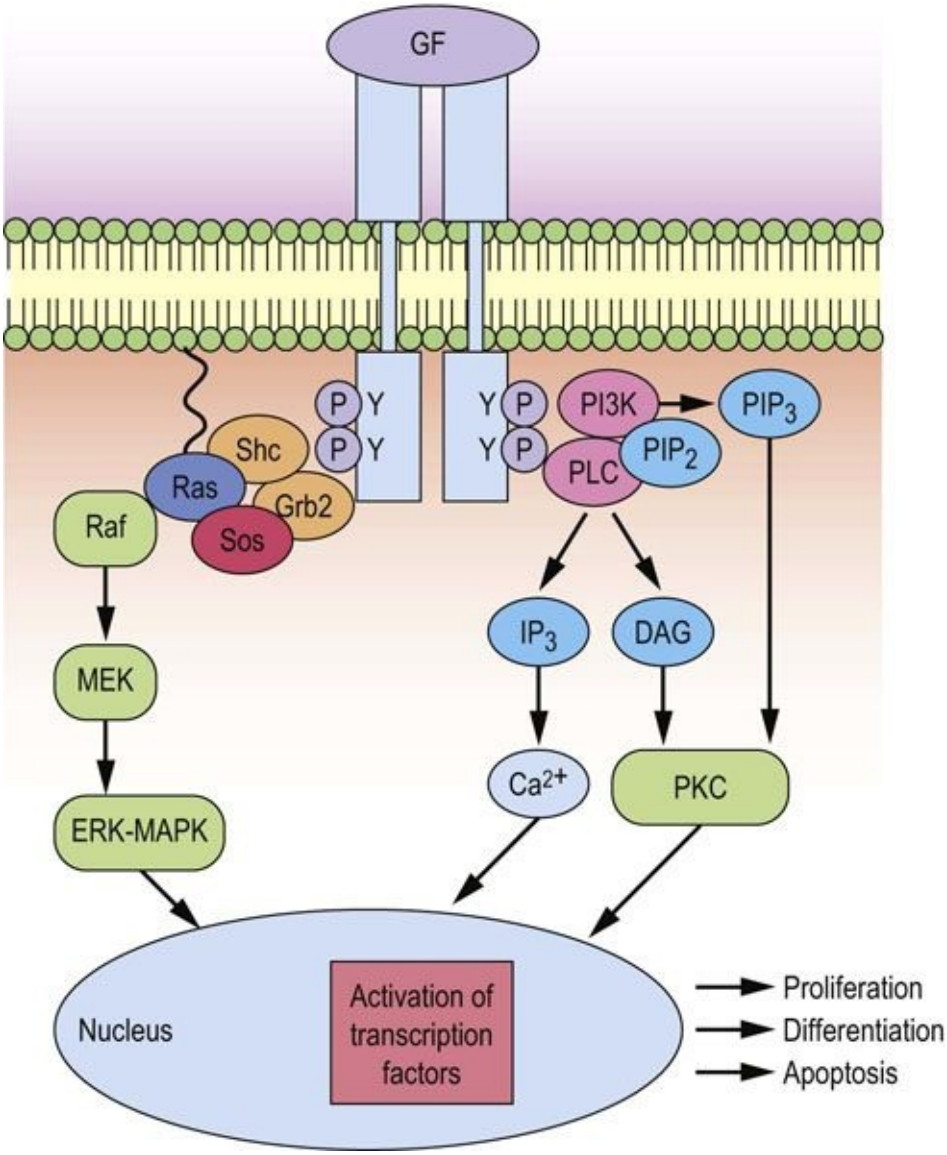


FIG. 42.3 Activation of ERK-MAPK kinase and PKC signaling cascades upon growth factor binding.
Growth factor signals activated at the plasma membrane as a result of ligand binding can regulate gene transcription, cell cycle progression, proliferation, differentiation or

apoptosis. Growth factor receptor signals activate proximal signals to recruit the adapter molecules Grb2 and Shc, which lead to the activation of ERK-MAPK and PI3K/Akt and PKC pathways. Activation of these pathways leads to the activation of transcription factors, such as Jun and Fos (which dimerize to form AP1), NFAT, and Myc. These regulate the induction of components of the cell cycle machinery that controls cell cycle progression and identify the need for DNA repair. If DNA damage is detected, then cell cycle is arrested, and if the damage is too great, then apoptosis is induced. GF, growth factor; DAG, diacylglycerol; IP_3, inositol 1,4,5-trisphosphate; PI3K, phosphatidylinositol 3-kinase ; PIP_2, phosphoinositol 4,5-bisphosphate; PIP_3, phosphoinositol 3,4,5-trisphosphate; PKC, protein kinase C; PLC, phospholipase C; Shc, Src homology and collagen domain protein.

Signaling cascade involving Ras GTPase is important in regulating cell division

A quarter of all tumors have constitutively active mutations in the signaling component Ras, which is critical for the transmission of proliferation and differentiation signals from the extracellular receptors to the nucleus. Ras is constitutively bound to the plasma membrane via a post-translational modification involving the addition of a lipophilic farnesyl group. Ligation of the EGFR recruits Ras by its binding to the adaptor protein Grb2. Ras is a GTPase, which cycles between an active GTP-bound and inactive GDP-bound form (Fig. 42.4). Its intrinsic catalytic activity is low, and is enhanced by binding to a GTPase-activating protein (GAP). GDP/GTP exchange is promoted by binding to a guanine nucleotide exchange factor, called Sos, which returns Ras into an active state. One of the main functions of active Ras is to act as an allosteric regulator of the mitogen-activated protein kinase (MAPK) signaling cascade. Ras transduces signals from EGFR via activation of two intermediary kinases, Raf and MEK kinase. MEK is a MAPK-type kinase that activates MAPK by mediating dual-specificity phosphorylation of tyrosine and threonine residues within its TEY activation motif. MAPK exists in two isoforms, extracellular-signal regulated kinase (ERK) 1 and 2. Upon activation, they translocate into the nucleus and phosphorylate (on serine and threonine) key transcription factors involved in the regulation of transcription of genes that regulate DNA synthesis and cell division.

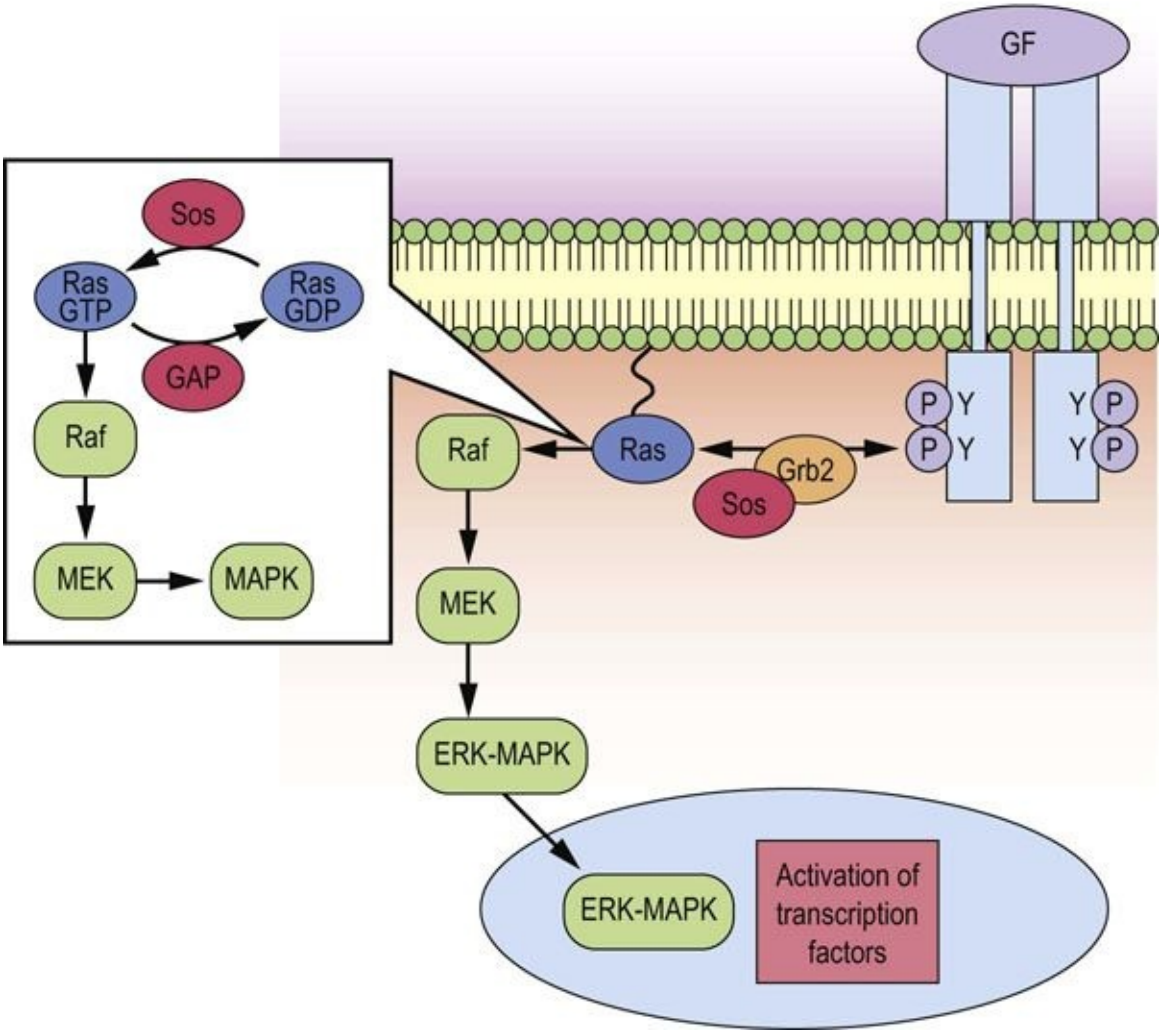


FIG. 42.4 Recruitment and activation of Ras by growth factor receptors. Ras is anchored in the plasma membrane and recruited to the activated growth factor receptor via interaction with the Grb2-Sos complex. Receptor stimulated GTP/GDP exchange, and thus activation of Ras, is promoted by Sos, whereas GAP inactivates Ras by stimulating its intrinsic GTPase activity. Ras couples growth factor receptors to the MAPK signaling cascade, via stimulation of the intermediary kinases Raf and MEK. MAPK translocates into the nucleus and phosphorylates key transcription factors involved in the regulation of DNA synthesis and cell division. GF, growth factor; GAP, GTPase-activating protein; MAPK, mitogen-activated protein kinase; Sos, son of sevenless.

mTORC-1 and mTORC-2 complexes integrate mitogen and nutrient signals

In addition to acting on selective members of the PKC family, PI3K is also

responsible for activating PDK1, which in turn activates Akt. This enzyme mediates the activity of mammalian target of rapamycin (mTor), a serine/threonine protein kinase. mTor can participate in two distinct signaling complexes mTORC-1 and mTORC-2, integrating mitogen and nutrient signals to promote cell survival, growth and proliferation. PI3K/Akt activation by receptor ligation leads to phosphorylation of the tuberous sclerosis protein 1/2 (TSC1/2), resulting in mTORC-1 activation. The best-characterized downstream effectors of mTORC-1 are translational regulators within the protein synthesis pathway, initiation factor 4E (eIF4E)-binding protein 1 (4E-BP1) and ribosomal S6 kinase 1 (S6K1), which are stimulated by mTORC1-mediated phosphorylation (Fig. 42.5).

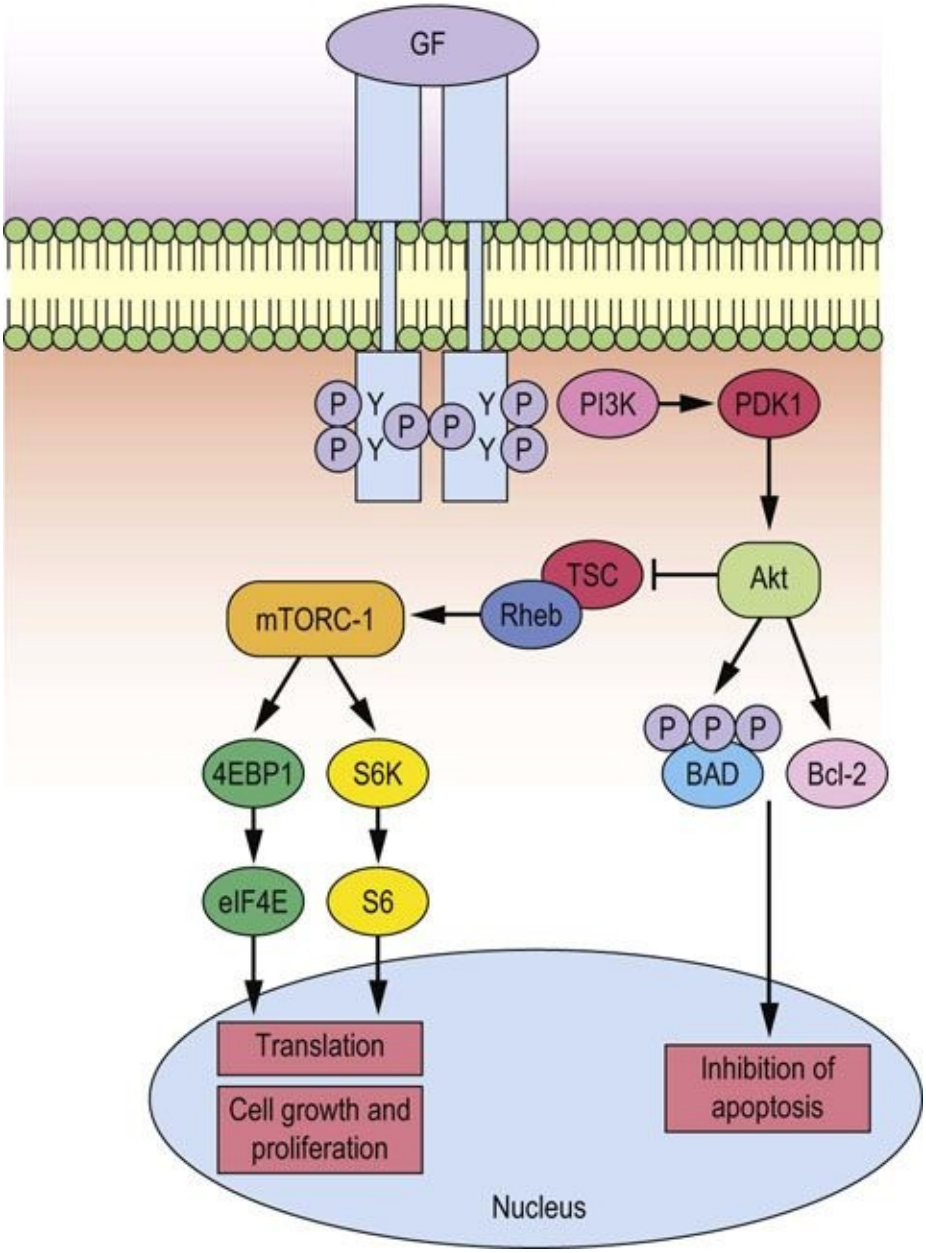


FIG. 42.5 Growth factor stimulation of mTor signaling.
Growth factor signals activated at the plasma membrane as a result of ligand binding can result in the recruitment of PI3K, thus activating PDK1/AKT, which enhances protein synthesis via mTOR (which is the catalytic subunit of the mTORC-1 protein complex shown), thereby contributing to cell cycle progression and inhibition of apoptosis by increasing the expression of cyclins and anti-apoptotic Bcl-2 family members. In addition, Akt inhibits apoptosis by mediating hyperphosphorylation of pro-apoptotic BAD, leading to the stabilization of mitochondria. GF, growth factor, PI3K, phosphatidylinositol 3-kinase; mTOR, mammalian target of rapamycin; BAD, Bcl-2-associated death promoter; TSC, tuberous sclerosis factor; Rheb, Ras homologue enriched in brain.

While the signaling pathways described above are linear, a significant amount of crosstalk occurs amongst these cascade elements

For example the ERK-MAPK signaling cascade is capable of regulating mTORC-1, by activating the kinase RSK1, which in turn phosphorylates and inhibits TSC1/2, thus resulting in activation of mTORC-1. Examples such as this illustrate the complexity that exists in intracellular signaling pathways downstream of receptors. Moreover, crosstalk goes some way to highlighting how a mutation in a single signaling component can impact on a wide array of biological responses within a single cell.

Cytokine receptor signaling

Cytokines are growth factors that mainly coordinate the development of hematopoietic cells and the immune response, although they also have multiple effects on nonhematopoietic cell types

As described above for growth factors, cytokines also exert their effects on cells by binding to cell surface receptors. There are many classes of cytokine receptors, many of which belong to a superfamily called the **hematopoietic receptors.** These are transmembrane glycoprotein receptors, characterized by the conserved extracellular ligand-binding domains, which contain characteristic cysteine pairs and a pentapeptide motif WSXWS (tryptophan-serine-X-tryptophan-serine, where X is any amino acid). Many consist of multisubunit receptors that contain a unique ligand-binding subunit that gives specificity, and a common signal transducer subunit that is often shared by several related cytokines. The sharing of the signal transducing subunit gives the basis for classification of cytokines into distinct subfamilies and helps to explain the severe immunodeficiencies that result from naturally occurring defects in these receptors.

Janus kinases (JAKs) link the hematopoietic receptors with the downstream signaling and gene transcription

In contrast to growth factor receptors, cytokine receptors do not possess an intrinsic catalytic activity. However, a family of cytosolic PTKs, called Janus kinases (JAKs), are essential for linking the hematopoietic receptors with the downstream signaling and gene transcription. Thus, after ligand engagement with the receptor, tyrosine kinase activity is induced, and most cytokines, in common with classic growth factors, can signal through PLC, PI3K and Ras-MAPK signaling cascades, JAKs associate with the receptors by binding to the conserved regions near the transmembrane domain. Upon cytokine binding, which causes receptor oligomerization, JAKs become phosphorylated and activated to mediate phosphorylation of their downstream targets, which are transcription factors called signal transducer and activators (STATs) (Fig. 42.6). In unstimulated cells, STATs are found in the cytoplasm in monomeric forms.

Cytokine stimulation leads to the JAK-mediated STAT phosphorylation and dimerization. STAT dimers then translocate to the nucleus, where they mediate transcription of the target genes by binding to their specific DNA sequences. Recently, growth factors such as EGF and PDGF have also been shown to induce activation of JAK/STAT pathways; therefore, JAK/STAT signaling may be a universal mechanism utilized by growth factors to regulate gene induction and cellular responses.

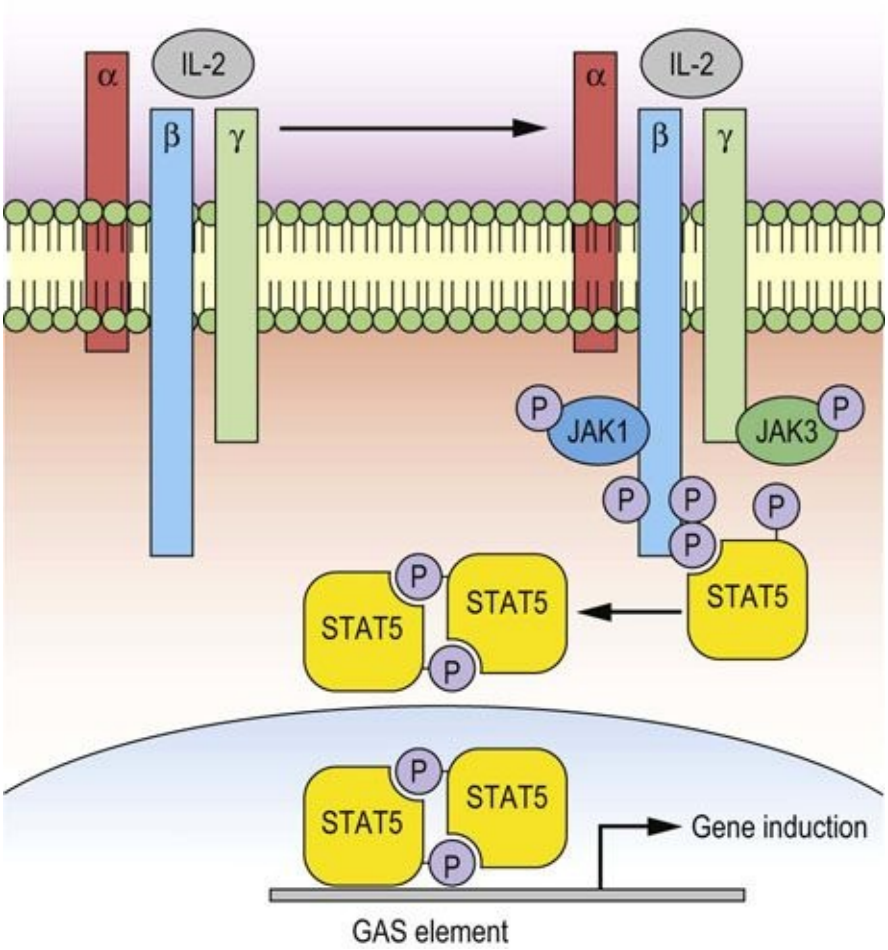


FIG. 42.6 Cytokine receptor signaling: IL-2 receptor.
The receptor is composed of the IL-2 receptor α-chain (IL-2Rα), IL-2/15Rß and the common cytokine-receptor γ-chain (γc). Binding of IL-2 to the α subunit causes association of this subunit with ß-and γ-subunits, thus forming a stable heterotrimer. JAK molecules associate with β-subunit (JAK1) and γc-subunit (JAK3), phosphorylate themselves and the key tyrosine residues within β-and γc-intracellular domains, thus enabling STAT5 recruitment and phosphorylation. -Phosphorylated STAT5 molecules

dissociate from the receptors and form dimers, which rapidly translocate into the nucleus and act as transcription factors, binding to GAS elements. Different JAKs and STATs can be utilized to achieve the specific response of individual cytokines. These pathways control transcription, cell growth, proliferation and survival. JAK, Janus kinase; STAT, signal transducer and activator of transcription; GAS, gamma interferon activation site.

Clinical box X-SCID: interleukin common gamma-chain deficiency

Severe combined immunodeficiency (SCID) is a group of rare, sometimes fatal, inherited congenital disorders characterized by little or no immune response. SCID patients share the common characteristic of a block in T cell development, associated with a variable impairment in B/NK immunity depending on the mutation that has taken place (Chapter 38). Without a functional immune system, SCID patients are susceptible to recurrent infections such as pneumonia, meningitis and chickenpox, and can die before the first year of life. Though invasive, new treatments such as bone marrow and stem-cell transplantation save as many as 80% of SCID patients.

Comment.

Around 60% of SCID patients have defective cytokine signaling. The gene responsible for SCID was initially reported to code γ-chain of the IL-2 receptor. It is now clear that this γ-chain is a **shared common signaling subunit** of receptors for IL-4, IL-7, IL-9, IL-13, and IL-15, highlighting the reason for such severe immunodeficiency, as the patient is unable to respond to any of these cytokines. The mutation is recessive, so heterozygous females are normal carriers, whereas the males who inherit the defective X chromosome develop the disease, hence being referred to as X-SCID. The intracellular domain of these proteins normally associates with the PTK, JAK3, whose activation leads to gene induction, cell growth and proliferation. In agreement with this, JAK3 mutations that lead to the absence or defective function of JAK3, are the cause of an autosomal recessive form of SCID, and result in a similar phenotype to X-SCID.

While treatment of this severe disorder can be through bone

marrow transplantation, gene therapy has been trialled as an option to restore the immune system with the introduction of the patient's own bone marrow-derived B and T cell precursors. In these cells the γ-chain defect was repaired by the viral-mediated gene therapy, rendering the cells responsive to the cytokines, thus enabling restoration of the adaptive immune response. The finding that JAK3 mutations also resulted in SCID suggested that JAK3 may be good therapeutic target for the development of novel immunosuppressants. Indeed, the JAK3 inhibitor, tofacinib, is used as an immunosuppressant to prevent **organ transplant rejection** as well as autoimmune diseases such as **psoriasis, psoriatic arthritis, inflammatory bowel syndrome and rheumatoid arthritis.** Moreover, JAK3 inhibitors are potential **antitumor therapeutics** for cancers in which increased JAK3 activity has been shown, such as acute **myeloid leukemia (AML), and colorectal and lung cancers**.

Regulation of cell cycle

Cyclin-dependent kinase (CDK) family of serine/threonine kinases, and cyclins, regulate cell cycle transition points

As cells progress through the cell cycle in response to growth factor stimulation, they must pass through three switch-like transition/restriction points, positioned at the G1/S phase boundary and at the entry and exit of M phase. Master regulators of these transition points include members of the cyclin-dependent kinase (CDK) family of serine/threonine kinases and a family of proteins known as cyclins.

CDKs are expressed as heterodimers comprising a protein kinase subunit and a regulatory cyclin subunit. Their activity is tightly regulated by different mechanisms, including the phosphorylation status of the kinase subunit, levels of cyclins and/or interaction with inhibitory proteins (CDKIs) that block their catalytic activity. While CDK expression levels are relatively constant throughout the cell cycle, the expression levels of cyclins are strictly controlled at both the mRNA and the protein level (transcriptional and translational control). Indeed, cyclins were originally defined as proteins that were specifically degraded during every mitosis.

The traditional model of cell cycle states that a specific cyclin-CDK partner drives distinct parts of the cell cycle: D-type cyclins and CDK4/6 regulate the events in early G1 phase, cyclin E-CDK2 triggers S phase, cyclin A-CDK2 and cyclin A-CDK1 regulate the completion of the S phase, while cyclin B-CDK1 controls mitosis (Fig. 42.7).

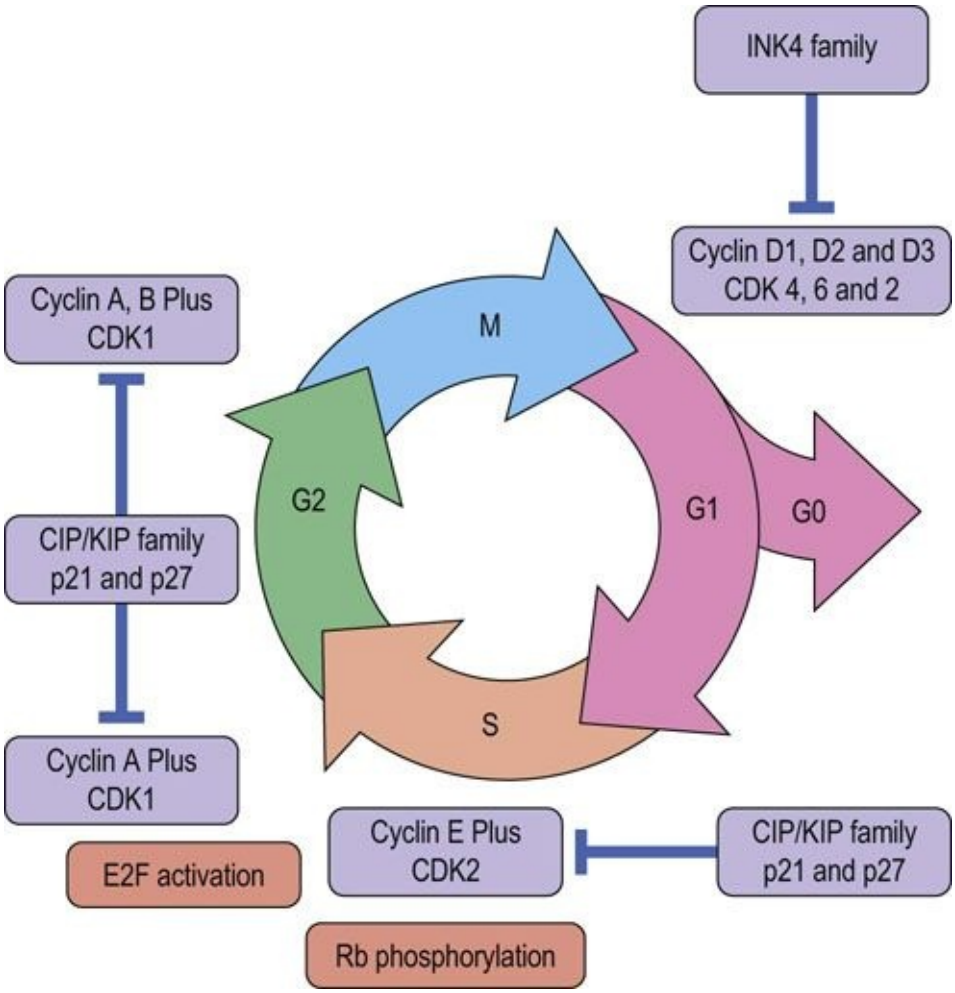


FIG. 42.7 Regulation of the cell cycle.
Progression through the cell cycle is regulated by interplay amongst specific cyclin-CDK partners and their inhibitors. Cyclin expression, particularly of those controlling G1 and S phase is regulated by growth factor signaling. INK4 family members inhibit cyclin D specific CDKs (CDK4 and 6), while CIP/KIP family inhibits all CDKs. CDK, cyclin-dependent kinase; CDKI, CDK inhibitor; E2F, transcription factor; Rb, retinoblastoma protein; INK4, inhibitors of CDK4 family.

Advanced concept box Changing the classical cell cycle paradigm

Pioneering studies assessing cells at different stages of the cell cycle, using classical biochemical approaches such as overexpression of kinase-dead CDK mutants and selective pharmacologic CDK inhibitors, revealed the preferred cyclin-CDK partners, leading to the formulation of the traditional model of cell

cycle. However, results from the recent genetic studies, employing specific deletions of cyclins and CDKs in mice and yeast, demonstrate that this model should be revised.

Interestingly, mouse embryos are viable even in the absence of CDK2, CDK4 or CDK6, indicating that these CDKs have redundant functions and are dispensable for controlling cell proliferation. The absence of these CDKs, however, affects the proliferation and/or differentiation of some specific cell types. CDK2 knockout mice survive for up to two years of age, but they are sterile, indicating that CDK2 is absolutely required for meiosis.

Moreover, CDK4 controls proliferation/differentiation of pancreatic β-cells, or pituitary hormone producing cells, whereas CDK6 regulates proliferation/differentiation of some hematopoietic cells. In contrast, CDK1 is essential for driving the cell cycle in most cell types, at least until halfway through gestation.

From these studies, a new model has emerged, referred to as the **minimal threshold model**. In this model, either CDK1 or CDK2 coupled with cyclin A or E is sufficient to control interphase, while CDK1 coupled with cyclin B drives cells into mitosis. The differences in activity of the same cyclin-CDK complexes in interphase and mitosis may be attributed not only to the substrate specificity, but also to the different localization within the cell, and higher activity threshold for mitosis than for interphase. In addition to being regulated by binding of cyclins, CDK activation is also modulated by phosphorylation. CDK activating complex (CAK), composed of CDK7, cyclin H and MAT1 (ménage a trois), mediates phosphorylation/activation of CDK1, CDK2, CDK4 and CDK6, when bound to their cyclin partners. Moreover, CAK plays a role in gene transcription as a part of TFIIH, which is a general transcription factor. In this context, CAK phosphorylates the RNA polymerase II large subunit *C*-terminal domain (CTD). This event is part of the process of promoter clearance and progression from the pre-initiation to the initiation phase of transcription.

Mitogenesis

Mitogenic signals activated by growth factors exert their effects between the onset of G1 phase and a point late in the G1 phase, called the restriction point

The key event for the initiation of the cell cycle in G1 phase is the phosphorylation of **retinoblastoma protein**, Rb, at various residues. Rb controls the expression of genes that commit cells that have reached restriction point late in G1 phase to enter the S phase (DNA synthesis) of the cell cycle. In the early G1 phase, Rb molecules are in the hypophosphorylated state. This allows them to bind and repress the DNA-binding activity of the main regulators of the G1/S phase transition, members of E2F family of transcription factors, thus inhibiting cell cycle progression (Fig. 42.7). Additional molecules that play an important role at this stage are histone deacetylases and chromatin remodeling complexes that epigenetically regulate gene transcription. Stimulation with growth factors and/or mitogens affects the entry into the cell cycle by triggering the expression and/or activation of proto-oncogenes, such as *Ras and Myc*, which results in the induction of expression of cyclins from the D-type family (D1, D2 and/or D3), followed by the cyclins from E-type family (E1 and E2). The D-type cyclins partner with CDK4/6 and stimulate their activity, whereas E-type cyclins increase the kinase activity of CDK2. These cyclin-CDK partners modulate the phosphorylation state of Rb by converting it from hypophosphorylated into hyperphosphorylated state, thus inactivating Rb, and promoting the release of E2F from Rb inhibitory complex. Free E2F proteins mediate gene transcription, the products of which are important for entry into S phase, and beyond, such as A-and B-type cyclins (Fig. 42.7). From this point onward, cell cycle progression is independent of growth factors.

When E-type cyclins are degraded, CDK2 binds to the A-type cyclins and these complexes phosphorylate many protein targets necessary for the proper completion and exit from the S phase. At the end of S phase, A-type cyclins also associate with CDK1, and these complexes share substrates with cyclin A-CDK2. The importance of the existence of both cyclin A-CDK2 and cyclin A-CDK1 complexes is still not clear.

Nevertheless, during G2 phase A-type cyclins are degraded by ubiquitin-mediated proteolysis and B-type cyclins are synthesized and interact with CDK1. It is estimated that cyclin B-CDK1 complexes phosphorylate more than

70 target proteins, which are important mediators of both regulatory and structural processes (chromosomal condensation, fragmentation of the Golgi network and breakdown of the nuclear envelope) during G2/M transition. Finally, the inactivation of cyclin B-CDK1 complexes is necessary for the exit from mitosis. This is achieved by the ubiquitin-labeling and subsequent proteasomal degradation of B-type cyclins regulated by the anaphase-promoting complex (APC). The main phosphatase that mediates return to interphase after mitosis is a form of protein phosphatase-2A (PP2A), whose activity is increased after degradation of the mitotic cyclins.

Monitoring for DNA damage

Molecular checkpoints that mediate appropriate progression through the cell cycle sense problems that may occur during DNA synthesis and chromosome segregation

The ultimate role of these checkpoints is to inhibit cyclin-CDK activity and thus delay or arrest cell division. During replication, DNA is less condensed and therefore protected to a lesser degree from the attack of exogenous and endogenous genotoxic agents, which may cause DNA damage.

In the case of damage occurring, the DNA damage checkpoints sense alterations and activate signaling pathways that mediate DNA repair

If the damage is beyond repair, they induce apoptosis. The core molecules of the DNA damage checkpoint are **sensor kinases,** ataxia-telangiectasia mutated (ATM) and ataxia-telangiectasia Rad3-related (ATR), which detect double-strand breaks and replication stress, respectively, and **checkpoint kinases,** CHK1 and CHK2, which relay the signals from the sensor kinases. These molecules prevent G1/S and G2/M phase transitions by inhibiting the CDK activity through increasing the expression of CDK inhibitor (CDKI) protein, p21 via p53 stabilization and/or inhibiting the CDK activator Cdc25 phosphatases.

The tumor suppressor protein p53 is predominantly a DNA damage-sensing protein which monitors DNA damage

throughout the cell cycle

If DNA damage is detected, ATM and subsequently CHK2 kinases are activated, action that contributes to the stabilization of p53 and thus enables the induction of DNA repair mechanisms. One of the roles of **p53** is to act as a transcription factor, increasing the expression of the CDKI p21 (WAF1). In turn, p21 inhibits cyclinD-CDK4 complexes, preventing Rb phosphorylation and thus promoting Rb/E2F binding and suppression of E2F-mediated gene transcription. Moreover, type-E cyclin-CDK2 complexes are inhibited, enabling cell cycle arrest at the G1/S transition. This allows the cell time to repair the DNA damage, and hence prevents the incorporation of mutated genetic material into daughter cells. However, if the DNA damage is beyond repair, p53-dependent programmed cell death by **apoptosis** is triggered.

A p53-independent pathway involving the INK4 family of proteins can also induce cell cycle arrest in the G1 phase in response to DNA damage

These proteins, including p16 (INK4A), p15 (INK4B), p18 (INK4C) and p19 (INK4D), mediate cell cycle arrest by binding to CDK4/6 or their binding partner, cyclin D, thereby causing the inactivation of the cyclin cyclin D-CDK4/6 complexes. Another important checkpoint is the spindle assembly checkpoint (SAC), which ensures proper alignment and segregation of chromosomes in the metaphase of mitosis. The SAC signal is generated by the presence of unattached or improperly attached kinetochores (protein complexes on chromatids which mediate attachment to the mitotic spindle), ultimately leading to the inhibition of the anaphase-promoting complex (APC) and thus preventing the onset of anaphase.

Defects in the DNA damage checkpoints allow accumulation of the DNA alterations, thus contributing to the **genomic instability,** whereas defective SAC may lead to the unequal segregation of the genetic material among the daughter cells, thus creating **chromosomal aberrations.** Both genomic instability and chromosomal aberrations are major culprits in cell transformation and oncogenesis.

Clinical box Ataxia telangiectasia

Ataxia telangiectasia is a rare autosomal recessive disease caused by the mutations in the ataxia telangiectasia-mutated (ATM) gene.

Patients suffering from this disease have the following symptoms and signs: an unsteady gut, telangiectasia, skin pigmentation, infertility, immune deficiencies, and increased incidence of cancer, especially lymphoreticular tumors.

Comment.

ATM is a serine/threonine protein kinase involved in the induction of DNA damage checkpoint in response to ionizing radiation, anticancer treatment, or programmed DNA breaks during meiosis. Ataxia telangiectasia cells exhibit chromosomal instability, hypersensitivity to reagents that induce DNA strand breaks and defects in the G1/G2 phases and altered regulation of p53 and p21 expression. ATM is constitutively expressed during the cell cycle and is involved in induction of p53-mediated cell cycle arrest and/or apoptosis. In ataxia telangiectasia, cells with mutated ATM cannot properly activate p53, and hence induce cell cycle arrest or apoptosis in response to ionizing radiation. Moreover, when p53 is also mutated, the cell cycle is not adequately regulated, which increases the risk of tumor formation as a result of accumulation of mutations.

Cell death

Cell death is a fundamentally important part of a cell's life cycle, and appropriate regulation of this process is critical to maintain the homeostatic regulation of a multicellular organism

Cell death may be accidental or programmed, initiated and executed through distinct biochemical pathways. **Programmed cell death (PCD)** is genetically regulated and its role is to remove superfluous, damaged or mutated cells. For many years, apoptosis was a synonym for PCD; however, this concept is changing due to the recent findings identifying different modes of controlled cell death.

Both initiation and execution of cell death are complex processes, with scientists classifying the various forms of cell death on the basis of morphological and/or biochemical features as the most common criteria. According to morphological criteria, cell death can be classified as:

- **Apoptosis** – the rounding up of the cell, retraction of pseudopods, reduction of cellular and nuclear volume (pycnosis), chromatin condensation and fragmentation (karyorrhexis), plasma membrane blebbing, formation of apoptotic bodies and engulfment by resident phagocytes in vivo.
- **Autophagy** (often referred to as autophagic cell death) – massive vacuolization of the cytoplasm, accumulation of the double membrane autophagic vacuoles, without chromatin condensation and little or no uptake by phagocytes in vivo.
- **Necrosis** or necrotic cell death – an enlargement of the cell volume, swelling of the organelles and plasma membrane bursting with concomitant loss of the intracellular content. Necrosis is often considered to be an accidental and uncontrollable form of cell death that occurs after severe insult to the cell. However, recent research has shown that necrosis can be controlled and initiated through specific signaling pathways involving mainly RIP1 serine/threonine kinase. This type of necrosis is called necrosis-like PCD or necroptosis and it has been observed in cancer cells, proliferating cells that have suffered DNA damage as well as cells infected with certain viruses (such as Vaccinia).
- **Cornification** (also known as ‘cornified envelope’ formation or

keratinization): an epidermis-specific form of PCD whose role is to create a barrier between the body and environment. It is characterized by elimination of cytosolic organelles, modifications of plasma membrane, accumulation of specific lipids and proteins both inside and outside of the cells to create skin elasticity, mechanical resistance and water resistance.

Apoptosis

Apoptosis is initiated and executed through either perturbation of intracellular homeostasis by the intrinsic (mitochondrial) or extrinsic pathways, the latter of which involves ligation of death receptors such as FAS, TNFR, TRAIL and TWEAK

For both of these pathways, two families of proteins are considered quintessential regulators: cysteine proteases called caspases, and the B cell lymphoma protein 2 (Bcl-2)-related family members, which interact to serve as modulators of life versus death decisions (Fig. 42.8). However, there is a growing body of evidence for other PCD pathways that sense stress and damage in other cellular organelles (such as endoplasmic reticulum and lysosomes), and result in the initiation of death programs. These programs can occur in association with, or independently of, the intrinsic mitochondrial pathway.

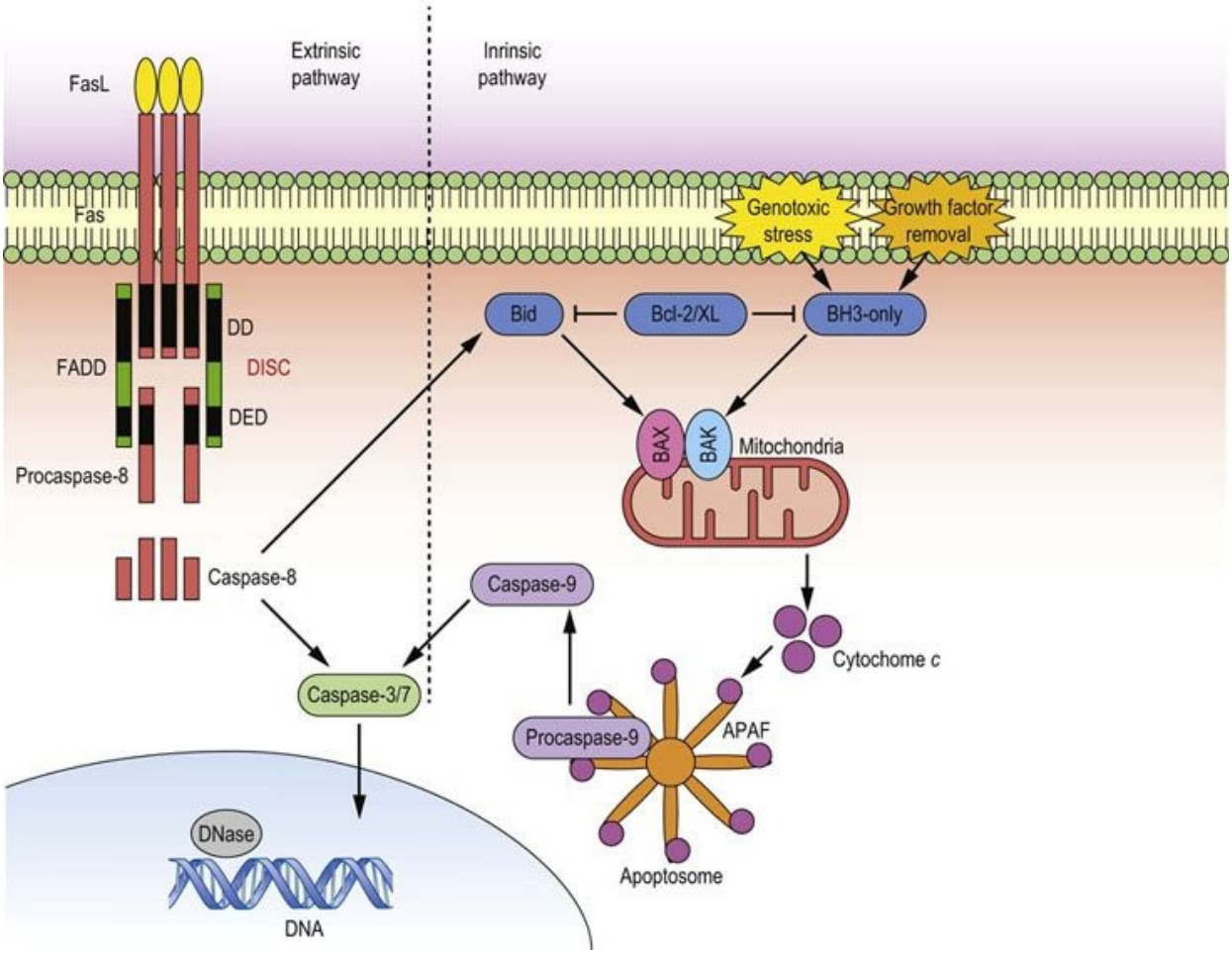


FIG. 42.8 Regulation of apoptosis.
Two main modes of apoptosis are induced through the ligation of death receptors, such as Fas, by growth factor deprivation, or genotoxic stress. Fas receptor ligation induces the extrinsic death pathway: the main initiator caspase is caspase-8. FasL binding causes receptor trimerization and formation of a macromolecular structure called DISC, which acts as a platform for caspase-8 activation. When activated, caspase-8 may directly cleave and activate effector caspase-3, or it may cleave BH3-only protein Bid, which acts directly on mitochondria, thus feeding into the second death pathway referred to as an intrinsic pathway. This pathway is initiated by the upregulation or activation of BH3-only family members, which relieve the inhibition mediated by Bcl-2/xL, thus allowing formation of the Bax/Bak pore on the mitochondrial outer membrane. This causes leakage of cytochrome *c* into cytoplasm, where it forms the apoptosome with APAF-1 and procaspase-9. Apoptosome is a platform for caspase-9 activation which, when cleaved, adopts active conformation, and can cleave effector caspase-3. Active caspase-3 then mediates bulk proteolysis either by activating other hydrolases, or by directly cleaving structural components. DISC, death-inducing signaling complex; Bid, BH3-interacting-domain death agonist; Bax, Bcl2-associated X protein; Bak, Bcl-2 homologous antagonist/killer; APAF-1, apoptotic protease activating factor 1; FADD, FAS-associated death domain protein; DD, death domain; DED death edffector domain.

Caspases

Caspases are cysteine proteases with aspartate substrate specificity

They are synthesized as inactive proenzymes (zymogens) referred to as procaspases. According to their role in the death pathways, caspases can be categorized as initiator or effector caspases.

Initiator caspases (caspase-2, -8, -9, -10) are synthesized as monomers and upon cell receiving a death signal they undergo activation, resulting from proximity-induced conformational changes and dimerization within multimeric complexes, as well as auto-proteolytic cleavage, which induces full enzymatic activity. Once cleaved/activated, caspases mediate proteolytic cleavage of other caspases in the death pathway cascade.

Effector caspase proenzymes are expressed as pre-formed dimers, which are activated by the direct proteolytic attack of initiator caspases. These effectors then execute the cell demise program by cleaving many vital cellular proteins (such as lamins, gelsolin), inducing cell cycle arrest, and disabling the initiation of homeostatic and repair mechanisms. All these events lead to the detachment of the cell from its surrounding tissue, dismantling structural components and ultimately flagging the phosphatidylserine (PS) ‘eat-me’ signal for phagocytosis. Overexpression of active caspases is sufficient to induce cellular apoptosis.

IAP gene family: its main function is to inhibit apoptosis

The inhibitor of apoptosis (IAP) gene family, composed of nine family members (X-linked IAP, cIAP1, cIAP2, melanoma IAP, IAP-like protein, neuronal apoptosis inhibitory protein, survivin, livin and apollon), is evolutionarily conserved from *Drosophila* through to humans. The main function of this gene family is, as their name indicates, to inhibit apoptosis by either directly blocking caspases and/or activating survival pathways via NFκB function. For example, X-linked IAP inhibits caspase-3, -7 and -9 by direct binding to the active pocket of caspase-3 and -7, while in the case of caspase-9 it prevents the dimerization necessary for full activation. On the other hand, cIAP1 and cIAP2 are positive regulators of both canonical and non-canonical pathways of NFκB activation. Survivin, in addition to inhibiting apoptosis, plays a role in cell cycle progression by blocking the activity of caspase-3, thus preserving the integrity of p21 within the survivin–caspase-3–p21 complex, and by mediating proper

chromosome segregation as a part of the complex that binds chromosomes to kinetochores. It is not surprising, therefore, that IAP family members are found to contribute to tumor cell survival, cell invasion and metastasis in many human cancers. Interestingly, loss of cIAP1/cIAP2 is implicated in the development of **multiple myeloma.**

Bcl-2 gene family is composed of structurally related proteins that form homo-or heterodimers and act as positive or negative regulators of apoptosis

The *Bcl-2* gene was initially discovered in a follicular B cell lymphoma as a protein that is constitutively expressed due to a t(14;18) chromosome translocation, which placed the *Bcl-2* gene under the control of the immunoglobulin heavy chain promoter. Bcl-2 family members have traditionally been classified into three groups: pro-survival family members (Bcl-2, Bcl-xL, Bcl-W, Mcl-1); the pro-apoptotic BAX/BAK family; and the pro-apoptotic BH3-only proteins (BIM, BID, PUMA, NOXA, BAD, BIK). While the function of Bcl-2 appears to be to preserve the integrity of the outer mitochondrial membrane, pro-apoptotic family members BAX and BAK are responsible for inducing mitochondrial outer membrane permeabilization and subsequent release of apoptotic mediators (such as cytochrome *c*), leading to the caspase activation. Bcl-2 and Bcl-xL prevent apoptosis induction by inhibiting BAX and BAK, while BH-3-only family members prevent this inhibition by a direct binding to the Bcl-2 and other anti-apoptotic family members (Fig. 42.8).

There are alternative routes to apoptosis

The rupture of lysosomal membranes causes release and activation of lysosomal proteases (i.e. cathepsins), which either mediate direct proteolytic cleavage of cellular components or activation of the intrinsic pathway. The endoplasmic reticulum (ER) stress is generally caused by the accumulation of unfolded proteins in the lumen, giving rise to the unfolded protein response or irregular intracellular Ca^{2+} flux. These perturbations usually induce apoptosis through the intrinsic pathway. Accumulation of unfolded or misfolded protein aggregates is associated with many neurodegenerative disorders, including **Alzheimer's, Huntington's** and **Parkinson's disease.** For example, amyloid-β-protein aggregates and mutations in the ER-associated presenilin 1 have been associated with the development of the familial Alzheimer's disease. Interestingly, all these

pathways may crosstalk and perform rather complex death programs in both the morphologic and biochemical sense.

Advanced concept box Intrinsic and extrinsic death pathway

The extrinsic apoptotic pathway is activated by the engagement of death receptors such as TNF family members (Fas, TNFR, TRAIL or TWEAK). For example, binding of homotrimeric FasL to Fas causes receptor oligomerization and assembly of an intracellular 'death-inducing signaling complex' (DISC). DISC contains procaspase-8, its adaptor/activator Fas-associated death domain (FADD) and its modulator cFLIP. Activation of caspase-8 occurs firstly by conformational change, which enables full enzymatic activity, and then by auto-proteolytic cleavage of the procaspase form. Cleaved caspase-8 molecules then leave the DISC and gain access to downstream targets, which include either effector caspases, such as caspase-3 and -7, or pro-apoptotic Bcl-2 family member BH3-interacting-domain death agonist (Bid). Cleaved Bid then feeds into the intrinsic pathway to amplify the death signal.

The intrinsic apoptotic pathway is also referred to as a Bcl-2 regulated pathway, due to the complex interplay between the pro- and anti-apoptotic Bcl-2 family members, which determines cell fate. It is usually activated either by the developmental cues, viral infections, DNA damage, and growth factor deprivation or other cytotoxic insults. These stress conditions increase the expression of BH3-only family members or, alternatively, their post-translational activation, depending on the context of death induction. Activated pro-apoptotic Bcl-2 family members relieve the inhibition of Bcl2-associated X protein (BAX) and Bcl-2 homologous antagonist/killer (BAK) at the pores of the outer mitochondrial membrane, enabling cytochrome *c* release from the mitochondrial intermembrane space into the cytoplasm. Cytoplasmic cytochrome *c* then binds to the apoptotic protease activating factor 1 (APAF1), allowing the formation of an apoptosome, which serves as a platform for caspase-9 activation (Fig. 42.8). Active caspase-9 then

cleaves caspase-3 and/or -7, which, in turn, mediate a bulk proteolysis of vital cellular proteins, activate DNases, and orchestrate the demolition of the cell.

Autophagy

Autophagy is a degradation process of cellular components in which a part of cytoplasm is engulfed by a specific membrane, and the contents are subsequently degraded by lysosomal enzymes

Autophagy is a highly regulated homeostatic process, which plays a role in the turnover of long-lived proteins or in the elimination of damaged organelles such as mitochondria (mitophagy) or ER (reticulophagy). In addition to being associated with cell death, autophagy can also enable cells to survive starvation conditions in the circumstances of decreased availability of extra-and intracellular nutrients. In this case, autophagy induces catabolic processes, which generate metabolic substrates from 'self'-components, thus allowing cells to meet their bioenergetic needs and initiate so-called adaptive protein synthesis in the time of scarcity.

Clinical box Deregulated apoptosis and autophagy can cause distinct pathologic conditions

Deregulated apoptosis.

Excessive apoptosis is linked to **neurodegenerative diseases** and **immunodeficiency**, while evasion of apoptosis is an important contributor to **oncogenesis** and the development of **autoimmune diseases**. Mutations (deletions/additions of one or a few nucleotides in the coding exons or splice sites) in the death receptor Fas lead to the defect in Fas-mediated apoptosis, which leads to the increased survival of activated lymphocytes causing

autoimmune lymphoproliferative syndrome (ALPS). This rare inherited disorder usually presents in early childhood. Patients with ALPS have lymphoadenopathy, splenomegaly, and autoimmune cytopenias, and have increased risk for development of lymphomas. Aberrant expression and function of Bcl-2 family members is also implicated in the development of autoimmune diseases and cancer. Deletion of the Bcl-2 family member Bim in mice causes a systemic lupus erythematosus-like (SLE-like) disease, while chromosomal translocation of the Bcl-2 to the immunoglobulin heavy chain locus (t(14;18)) results in the constitutive expression of Bcl-2, leading to the development of follicular lymphoma.

Deregulated autophagy.

Perturbations in autophagy induction and execution may also lead to a series of disorders and diseases. Monoallelic deletions in *BECN1/ATG* gene are tumorigenic in mice, and a decrease in expression of Beclin 1 is observed in **human breast carcinoma.** Loss of function mutations in the Pink1 and Parkin genes, which are regulators of mitophagy, are linked to **familial Parkinson's disease** in humans. Similar to the common neurodegenerative diseases are **lysosomal storage disorders,** which include over 40 genetic conditions, most of which are linked to the deficiency of lysosomal hydrolases. Reduced or impaired function of these enzymes leads to the accumulation of otherwise degraded macromolecules, while **Huntington's** and **Parkinson's diseases** are associated with accumulation of a mutant form of protein α-synuclein.

Autophagy is induced by a variety of stress stimuli, including nutrient and energy stress as well as hypoxia, redox stress, infections, ER stress, and mitochondrial damage

All of these stresses induce distinct signaling pathways that regulate autophagy.

A characteristic signaling event that induces autophagy as a result of nutrient deprivation is the inhibition of mTORC-1 signaling and/or AMPK activation. The process of autophagy involves the formation of a membrane structure, termed a phagophore (most likely by the de novo synthesis), which encases part of the cytoplasm or a whole organelle, forming a double-membrane structure called the autophagosome. The autophagosome then may fuse with an endosome, creating an amphisome. Subsequently, these structures fuse with a lysosome, generating an autolysosome where acid hydrolases break down the inner membrane and the cargo. Digested macromolecule building blocks are then recycled back into the cytoplasm through protein channels referred to as permeases (Figs 42.9 and 42.10).

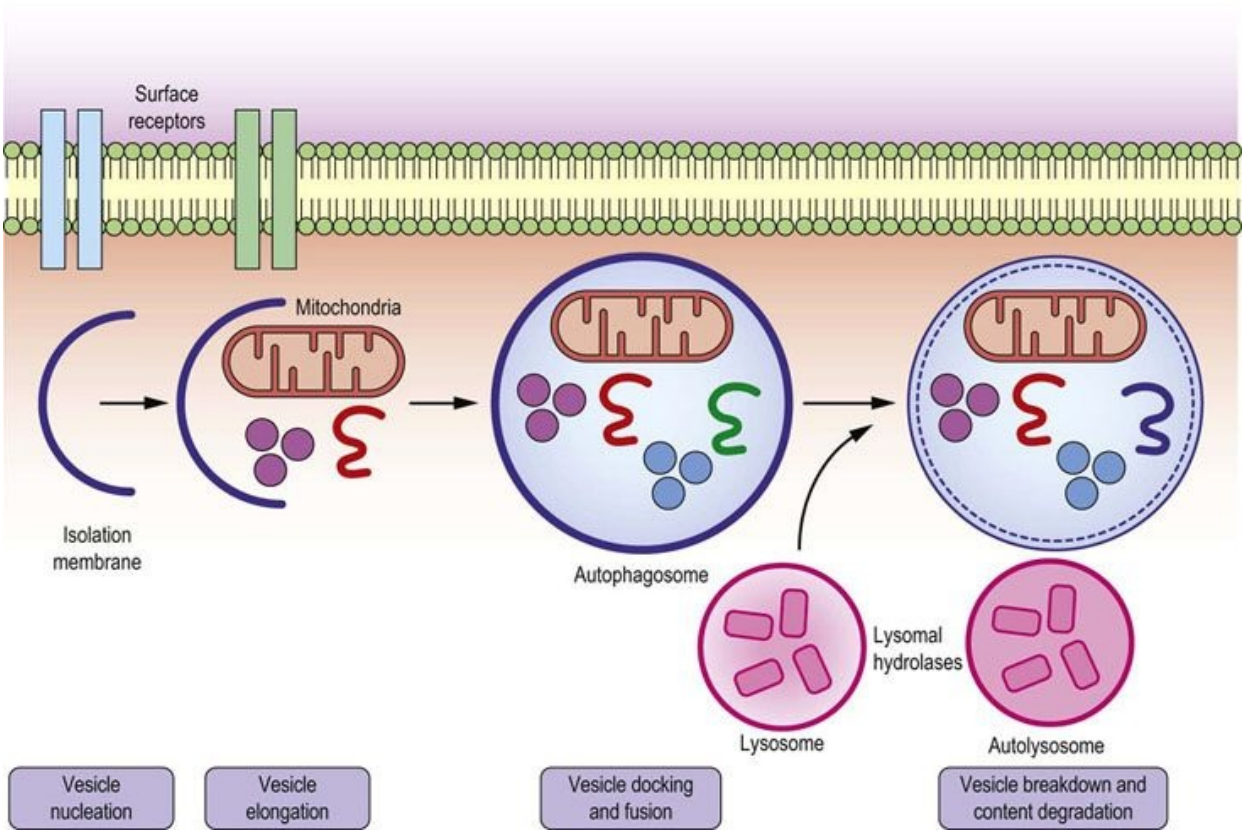


FIG. 42.9 Autophagy.
The process of autophagy starts by formation of isolation membrane (phagophore), which engulfs damaged mitochondria and/or misfolded proteins and forms the double-membraned vesicle autophagosome (autophagic vacuole). Autophagosomes then mature and fuse with lysosomes, thus creating autolysosomes, in which the inner membrane of autophagosome and its luminal content are degraded by the action of

lysosomal acid hydrolases, such as cathepsins.

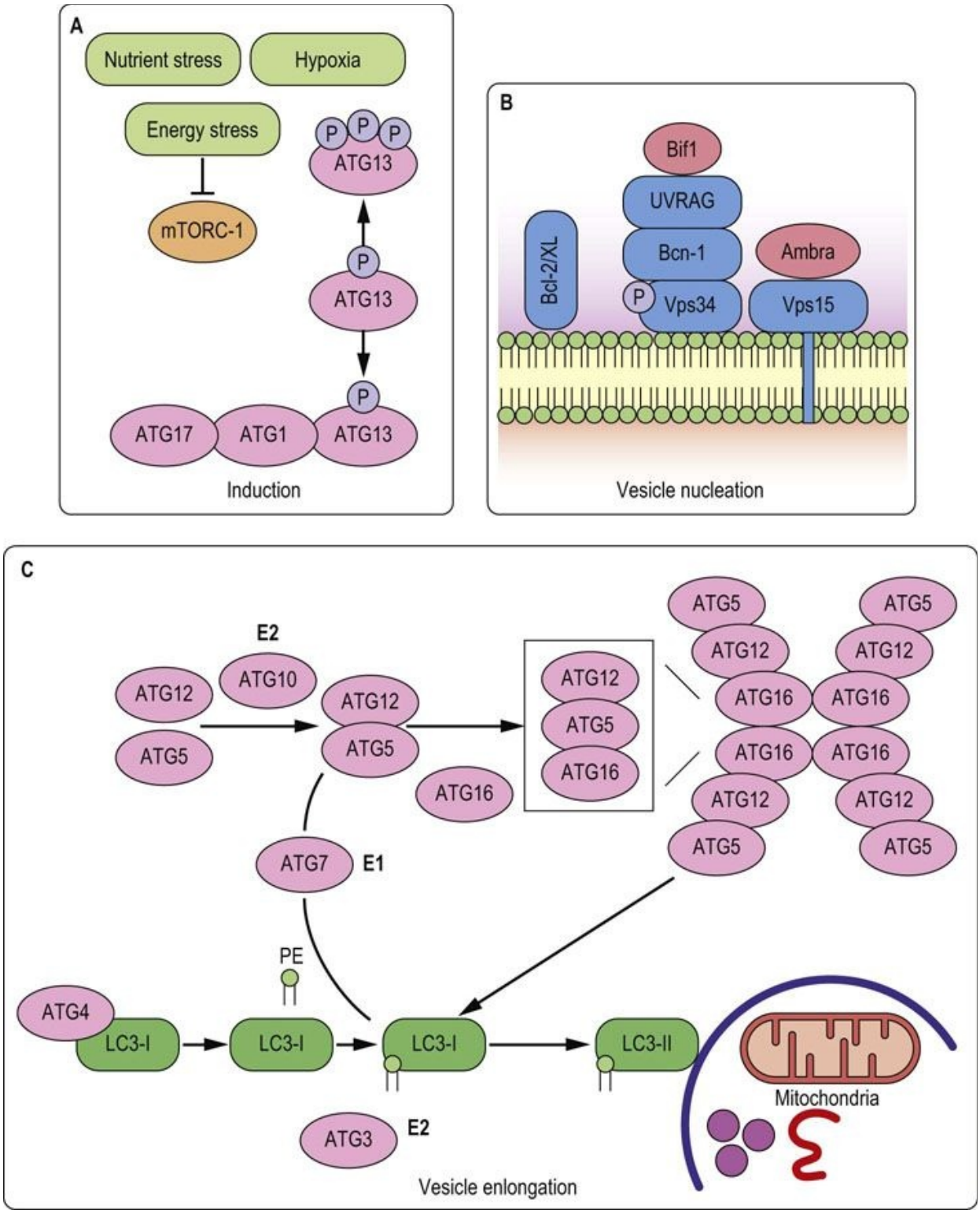


FIG. 42.10 Autophagy: important molecules.
A. Autophagy is usually induced by the inhibition of mTOR pathway, which relieves ATG13 protein from inhibitory phosphorylation and allows its association with ATG1. Within the ATG1–ATG13–ATG17 complex, ATG1 catalytic activity is increased and

contributes to autophagy induction by an as yet unknown mechanism. **B.** Vesicle nucleation is initiated by the activation of Vps34, a mammalian class III PI3K, within a complex with Bcn-1, UVRAG and myristylated kinase Vps15. Active Vps34 catalyzes generation of PIP3 necessary for the autophagosomal membrane formation. **C.** Vesicle elongation and completion of autophagosome is regulated by the two ubiquitin-like conjugation systems: LC3 (ATG8) and ATG12. Phosphatidylethanolamine (PE) is conjugated to LC3 in a stepwise process controlled by the ATG4 protease, the E1-like enzyme ATG7 and E2-like enzyme ATG3. Binding of the PE to LC3 converts LC3 (LC3-I) from a soluble form into autophagosomal membrane-associated form (LC3-II). This conversion process is the foundation for many methods of autophagy detection. In the second pathway, ATG12 is conjugated to the ATG5 in a process mediated by the **E1-like enzyme ATG7** and **E2-like enzyme ATG10**. ATG12-ATG5 then forms a multimeric complex with ATG16, which helps in targeting LC3 to the autophagosomal membrane, and accelerates lipid conjugation. ATG, autophagy-related gene; PI3K, phosphatidylinositol 3-kinase; UVRAG, UV radiation resistance-associated gene protein; LC3, microtubule-associated protein light chain 3. For more details see Advanced Concepts Box on the next page.

Apart from its roles in maintenance of cell homeostasis, autophagy plays a role in both innate and adaptive immunity. For example, autophagy is used for the elimination of intracellular bacteria such as *Streptococcus pyogenes* and *Mycobacterium tuberculosis*. Moreover, the Epstein–Barr virus nuclear antigen 1 (EBNA1) is processed through an autophagic pathway, and loaded on MHC class II molecules for presentation to the $CD4^+$ T lymphocytes.

Advanced concept box Regulation of autophagy by atg gene family

Autophagy is a stepwise process whose main regulators are members of the autophagy-related gene (ATG) family, composed of 31 genes to date. Since their initial discovery in *Saccharomyces cerevisiae*, orthologues (genes in different species that evolved from a common ancestral gene) have been identified in mammals, indicating this gene family is conserved from yeast to humans.

Comment.

The inhibition of mTOR, more precisely mTORC-1, relieves the inhibition of ATG13, which can associate with ATG1 kinase, thereby increasing its catalytic activity and inducing autophagy by a mechanism that is not well defined. Initial steps of autophagosome formation are mediated by the PI3K complex, which contains mammalian Vps34 (class III PI3K, which generates

phosphatidylinositol-3-phosphate PIP_3), *BECN-1* and UV irradiation resistance-associated tumor suppressor gene (UVRAG; Fig. 42.10). The next step, vesicle elongation, is mediated by the two ubiquitin-like conjugation systems. Firstly, E1-like enzyme ATG7 and E2-like enzyme ATG10 conjugate ATG12 to ATG5. Secondly, protease ATG4, E1-like enzyme ATG7 and E2-like enzyme ATG3 conjugate phosphatidylethanolamine (PE) to microtubule-associated protein light chain 3 (LC3/ATG8). Binding of PE to LC3 is crucial for its translocation to the autophagosomal membrane (Fig. 42.10). The conversion of soluble (LC3-I) to the membrane-bound form (LC3-II) is widely used as an experimental marker for autophagy. After their formation, autophagosomes mature into the autolysosomes by fusion with lysosomes, which allows the process of degradation to be completed.

Cancer

Cells that develop mutations affecting normal regulation of the cell cycle are able to undergo unchecked proliferation, resulting in a loss of homeostatic regulation and the development of a tumor or neoplasm

As long as the neoplastic cells remain as an intact tumor, the tumor is considered **benign** and can be removed surgically. However, if further mutations allow such tumor cells to invade and colonize other tissues, creating widespread secondary tumors or **metastases,** the tumor is described as **malignant** and classified as a **cancer**. Each cancer is derived from a single cell that has undergone some germline mutation that allows it to outgrow its surrounding cells; by the time they are first detected, tumors typically contain a billion cells. Cancers are classified according to the tissue and cell type that they are derived from: those from epithelial cells are **carcinomas**, those from connective tissue or muscle cells are termed **sarcomas,** and those from the hematopoietic system are called **leukemias**. About 90% of human cancers are carcinomas, the five most common being those of the **lung, stomach, breast, colon/rectum, and uterine cervix**.

In the majority of cases, a single mutation is not sufficient to convert a healthy cell to a cancer cell; several rare mutations have to occur together

Mutations in DNA occur spontaneously at a rate of 10^{-6} mutations per gene, per cell division (even more in the presence of mutagens). Thus, since approximately 10^{16} cell divisions occur in the human body over an average lifetime, every human gene is likely to undergo mutation on about 10^{10} occasions. Clearly then, a single mutation is not normally sufficient to convert a healthy cell to a cancer cell; several rare mutations have to occur together, as demonstrated by epidemiologic studies showing that, for any given cancer, the incidence increases exponentially with age. It has been estimated that 3–7 independent mutations are usually required, leukemias apparently needing the fewest mutations and carcinomas the most.

Mutations need to occur in the appropriate cells in order to enable the neoplasm to develop, indicating that cell context has an important bearing on the type of cancer that subsequently develops

In addition to developing cancer-promoting mutations, the cell in which a mutation occurs must then be permissive to being a cancer-initiating cell. This relates to the cellular context in which an oncogene is expressed, and the properties bestowed upon a cancer cell by expression of a particular oncogene. In order to become oncogenic and promote cancer cell growth, the cell must possess self-renewal capacity. If oncogene expression occurs in a stem cell, then oncogene expression could inhibit the negative regulatory networks in place that would ordinarily stop cell growth and proliferation, thus generating a cancer stem cell. Cancers such as chronic **myeloid leukemia (CML)** and **acute myeloid leukemia (AML)** can arise from mutations in stem cells. However, it is not essential for the cell of origin of individual cancers to arise from stem cells. Indeed, cancers can arise from mutations arising in committed progenitor cells that enable those cells to acquire the potential to self-renew, therefore providing a cellular source for cancer.

Mutations that lead to the expression of established oncogenes do not necessarily lead to the development of

cancer if it occurs in nonsusceptible cells

For example, the Philadelphia chromosome (t(9;22)) that generates a fusion protein BCR-Abl, encoding a constitutively active from of the protein tyrosine kinase (PTK) c-Abl (which is the causative mutation of the development of >95% CML cases), is detected at a very low level in circulating peripheral blood cells isolated from around 30% of healthy individuals. Studies have demonstrated that expression of BCR-Abl alone does not confer self-renewal properties on committed progenitor cells, indicating that secondary mutations are required to make committed progenitor cells cancerous. This finding suggests that the expression of oncogenes such as BCR-Abl within a stem cell environment can enable it to hardwire into the initiation of the neoplastic program.

Advanced concept box Oncogenic events are cell-context dependent

While genetic mutations are common, and expected, with loss of function in p53, PTEN and Rb, and gain of function with *Ras*, genome sequencing has also identified mutations in 'unexpected' genes in individual cancers. For example the *Notch* gene is mutated to form a constitutively active Notch protein, and is linked to the generation of around 50% of **T lymphocyte-acute lymphoblastic leukemia (T-ALL)** cases. However, activating mutations in Notch1 have also been identified in B lineage diseases including **B lymphocyte chronic lymphocytic leukemia (B-CLL)** and mantle cell lymphoma (MCL). Interestingly, an inactivating mutation of Notch has recently been identified in head and neck squamous cell carcinoma, indicating that while in hematopoietic lineages Notch acts as a tumor promoter, it can behave as a tumor suppressor in squamous cell carcinogenesis.

Alterations in gene expression have also been identified as a mechanism for tumor development/promotion. Indeed, expression patterns of specific protein kinase C (PKC) isoforms are dysregulated, possibly through alterations in epigenetic regulation of gene expression, in a number of cancers. In particular, PKCα is upregulated in breast, gastric, prostate and brain cancers, suggesting that it contributes to tumorigenesis. Moreover, PKCα

expression levels have also been linked with the aggressiveness and invasive capacity of breast cancer cells. However, PKCα expression is downregulated in epidermal, pancreatic, colon cancers and B-CLL, suggesting that PKCα can also function as a tumor suppressor. Taken together, these findings indicate that the cell in which the mutation or modulation of expression of a protein occurs has a direct bearing on whether cancer formation is a likely outcome.

Tumor promoters: oncogenes

Mutations that lead to the uncontrolled proliferation of cancer cells can result either from disruption of the control of normal cell division or, alternatively, from a reduction in the normal processes of terminal differentiation or apoptosis. This distinction is reflected by the two major groups of genes targeted for mutation in cancer: **oncogenes** and **tumor suppressor genes**.

Oncogenes were first identified as viral genes that infect normal cells and transform them into tumor cells

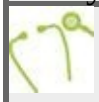
The Rous sarcoma virus, which is a retrovirus that causes connective tissue tumors in chickens, will infect and transform fibroblast cells grown in cell culture. The transformed cells outgrow the normal cells and exhibit a number of growth abnormalities, such as a loss of cell contact-mediated inhibition of growth and loss of anchorage dependence of growth. In addition, the cells have a rounded appearance and can proliferate in the absence of growth factors. Moreover, the cells are immortal, do not senesce, and can induce tumor formation when injected into a suitable animal host, confirming their ability to self-renew.

The key to understanding cell transformation lies in the mutation of a normal cellular gene that controls cell growth

The use of mutant Rous sarcoma viruses that, despite multiplying normally, have lost the ability to transform host cells, showed that it was the *Src* gene that was responsible for such cell transformation. The breakthrough in our understanding

of how this single gene could transform cells in culture came when it became apparent that the viral oncogene was a mutated homologue of a normal cellular gene. This gene is now called the c-*Src* proto-oncogene, and has been identified as a PTK signal transducer involved in the normal control of cell growth. As expression of this gene is not essential to the survival of the retrovirus, it is likely that *Src* was accidentally incorporated by the virus from a previous host genome and was somehow mutated in the process. In the case of the Rous sarcoma virus, the introns normally present in c-*Src* are spliced out and, in addition, there are a number of mutations causing amino acid substitutions, resulting in a constitutively active PTK.

Cell transformation can, however, also result from oncogenes that are not constitutively activated but, rather, are overexpressed in an abnormally high number of copies, as a consequence of the gene being under the control of powerful promoters or enhancers in the viral genome. Alternatively, for retroviruses, DNA copies of the viral RNA can be inserted into the host genome at or near sites of proto-oncogenes (insertional mutation), causing abnormal activation of these proto-oncogenes. In this situation, the altered genome is inherited by all progeny of the original host cell.



Clinical box mTOR in cancer and metabolic disorders

Owing to its vital role in regulating cell proliferation/survival and its intricate relationship with the phosphatidylinositol-3-kinase(PI3K) pathway, components of the mTOR pathway are often dysfunctional in many types of cancers and metabolic disorders. **Tuberous sclerosis complex** (TSC) is an autosomal dominant genetic disorder that occurs as a result of inactivating mutations within the *TSC1* or *TSC2* genes. TSC is characterized by multiple benign tumors such as angiofibroma of the skin, lymphangioleiomyoma of the lungs, renal angiomyolipoma and astrocytoma of the brain.

In the context of the mTORC-1 axis, Tsc1/2 serve as a relay center for tumor microenvironmental cues. Under normal circumstances, hypoxia (via Hif1α), DNA damage (via p53) and nutrient deprivation (via LKB1 transcription factor) activate Tsc1/2 to regulate mTORC-1, thus controlling biosynthetic

processes. These pathways are inactivated during tumorigenesis, usually by the cooperative action of oncogenic PI3K/PDK1 and Ras/MAPK pathways to reduce Tsc1/2 activity. Upregulation of mTORC-1 signaling leads to overactivated protein and lipid biosynthesis, which supports the bioenergetic needs of energy-demanding proliferating tumor cells. Enhanced protein synthesis often promotes increased expression of cell cycle regulators such as cyclin D1 and cyclin E, while constitutively active Akt contributes to the inactivation of cell cycle inhibitors p27 and p21. The role of mTORC-2 in tumorigenesis is still not very well defined; however, part of the mTORC-2 complex called Rictor is overexpressed in many gliomas. Higher expression promotes mTORC-2 complex assembly and activation, allowing cells to have higher proliferation and increased invasion. These events suggest that cancer could be considered a metabolic disorder.

Indeed, mTOR pathway deregulation has also been shown to contribute towards the development of various metabolic disorders such as **obesity, nonalcoholic fatty liver and diabetes type 2. For** example, in the hypothalamus, leptin relays signals via mTORC-1 for the reduction in food intake. Overactivation of mTORC-1 due to the high-fat diet may promote obesity by favoring resistance to the leptin-induced anorexic signals, thus promoting hyperphagia. Moreover, increased activation of mTORC-1 promotes adipogenesis and adipose tissue expansion, inhibits insulin signaling in skeletal muscles, liver and pancreas by promoting insulin resistance, and contributes to the induction of apoptosis in pancreatic β-cells by exhausting the homeostatic process of β-cell compensation.

Most human tumors are nonviral in origin and arise from spontaneous or induced mutations

Approximately 85% of human tumors arise as a result of point or deletion mutations. These mutations may be spontaneous or induced by carcinogens or radiation, resulting in overexpression or hyperactivity of the proto-oncogenes.

Indeed, *Ras*, which has been found to be mutated to a constitutively activated form in approximately 25% of all tumors, appears to exert many, if not all, of its effects ultimately by upregulating concentrations of cyclin D and, hence, by stimulating cell cycle progression. This upregulation of cyclin D expression results from stimulation of the MAPK cascade by *Ras*, and the induction of the transcription factor AP-1, which regulates induction of cyclin D expression.

Whole-exome/genome sequencing of individual patients, utilized to determine the specific mutational landscape within cancer subtypes, has enabled links to be established between seemingly diverse cancers that result from similar genetic mutations

Of note, B-Raf^{V600E} mutation was recently found to be present in all in **hairy cell leukemia** (HCL) patients assessed. B-Raf^{V600E} is oncogenic in a number of tumors, including melanoma, encoding active B-Raf kinase leading to the constitutive activation of the MEK/ERK signaling pathway. This mutation has a major impact on cell cycle. Interestingly, now that the B-Raf^{V600E} mutation has been linked as a causal/driver event of HCL, it presents the opportunity for targeted therapies, such as PLX-4720, a specific inhibitor of active B-Raf, which would otherwise have not been considered for the treatment of HCL. Identification of specific driver mutations in a particular cancer cell opens up the possibility of treating patients with a targeted therapy towards that mutation.

Karyotyping of tumor cells has also shown that chromosomal translocation can bring the oncogene under the control of an inappropriate promoter. For example, in Burkitt's lymphoma, overexpression of the *Myc* gene occurs through its translocation into the vicinity of one of the Ig loci. Because *Myc* normally acts as a nuclear proliferative signal, overexpression of *Myc* induces the cell to divide, even under conditions that would normally dictate growth arrest.

Tumor suppressor genes: subversion of the cell cycle

Mutations in tumor suppressor genes are recessive and thus, mutations in both copies of the gene are usually required for transformation. As it is very difficult to identify the loss of function of a single gene in a cell, much of the initial information relating to tumor suppressor genes was obtained by studying a range

of inherited cancer syndromes (Table 42.1).

Table 42.1

Selected inherited cancer syndromes

Syndrome	Cancer	Gene product
Li–Fraumeni	Sarcomas, adrenocortical, carcinomas of breast, lung, larynx. Colon and brain tumors and leukemias	p53: transcription factor, DNA damage and stress
Familial retinoblastoma	Retinoblastoma, osteosarcoma	Rb1: cell cycle and transcriptional regulation
Familial adenomatous polyposis (FAP)	Colorectal cancer; colorectal adenomas, duodenal and gastric tumors, jaw osteomas and desmoid tumors (Gardner syndrome), medulloblastoma (Turcot syndrome)	APC: regulation of β-catenin, microtubule binding
Wiedmann–Beckwith syndrome	Wilms' tumor, organomegaly, hemihypertrophy, hepatoblastoma, adrenocortical cancer	P57/KIP2: cell cycle regulator
PTEN hamartoma tumor syndrome	Benign tumors: breast, thyroid, colorectal, endometrial and kidney cancers	PTEN: protein and lipid phosphatase, regulating Akt kinase and cell cycle
Neurofibromatosis type 1 (NF1)	Neurofibrosarcoma, AML, brain tumors	GTP-ase activating protein (GAP) for Ras
Hereditary papillary renal cancer	Renal cancer	MET receptor for HGF
Familial melanoma	Melanoma, pancreatic cancer, dysplastic neviatypical moles	P16 (CDK): inhibitor of cyclin-dependent kinase (CDK4/6)

AML, acute myeloid leukemia; HGF, hepatocyte growth factor; KIP2, 57 kDa inhibitor of cyclin–CDK complexes; HGF, hepatocyte growth factor.

p53: guardian of the genome

The protein p53 plays a critical role in regulating the G1/S phase transition of the cell cycle and monitoring for DNA damage, and becomes activated upon sensing damage, stress and oncogenic signals to induce cell cycle arrest and/or death. Therefore it is perhaps unsurprising that p53 function is commonly disrupted in cancer cells. This can either be through inactivating mutations of p53 function directly, abrogating its transcriptional activity, or by dysregulating pathways responsible for p53 activation. The importance of p53 is highlighted in individuals who only have one functional copy of the *p53* gene. People with this syndrome, called **Li–Fraumeni syndrome,** are predisposed to develop a wide range of tumors, including sarcomas, carcinomas of the lung, breast, larynx and colon, brain tumors and leukemias. This syndrome is rare, and tumor cells in affected patients exhibit defects in both copies of *p53*. Deletion of *p53*, in addition to allowing uncontrolled cell cycle progression, also permits replication of damaged DNA, leading to further carcinogenic mutations or gene amplification.

Phosphatase and TENsin homologue (PTEN)

The tumor suppressor PTEN is one of the most commonly inactivated proteins in sporadic cancer

As described above, PI3K-mediated signaling pathways are activated in response to a multitude of growth factor stimuli, resulting in the promotion of cell growth, survival and proliferation. The main protein responsible for attenuating PI3K activity and downstream pathways is the dual-specificity protein and lipid phosphatase PTEN. It reverses the activity of PI3K by dephosphorylating PIP_3. The tumor suppressor PTEN is one of the most commonly inactivated proteins in sporadic cancer, resulting in sustained PI3K/Akt signaling and uncontrolled cell survival and proliferation. Mutations in PTEN have been identified in a wide range of cancers, including breast, thyroid, prostate and brain cancers. Interestingly, individuals with inherited germline mutations in PTEN, known as PTEN hamartoma tumor syndrome, develop benign tumors associated with breast, thyroid, colorectal, endometrial and kidney. However, these patients also carry an elevated lifetime risk of developing malignant cancers in these tissues. The increased susceptibility to cancer in cells carrying germline PTEN mutations underlines the importance of this protein as a tumor suppressor.

Clinical box Specific mutations that define a cancer therapy

In the majority of cases a number of mutations are required in to order to change a healthy cell into a cancer cell. One possible exception to this rule is **chronic myeloid leukemia** (CML), which in 95% of cases develops as a result of a translocation between chromosomes 9 and 22 (t(9;22)), to develop the Philadelphia chromosome, in the hematopoietic stem cell compartment. This type of translocation can also be found in 25–30% of **acute lymphoblastic leukemia** (ALL) cases and a minority of acute **myeloid leukemia** (AML) cases. The translocation forms a fusion product between the breakpoint cluster regions (*BCR*) and the Abl protein tyrosine kinase gene *ABL* Abl, resulting in the constitutive production of fusion protein BCR-Abl, which exhibits enhanced protein tyrosine kinase (PTK) activity. As Abl regulates a number of proteins involved with cell cycle, the result is that BCR-Abl

expression increases cell division, leading to an overproduction of myeloid lineage cells. Moreover, BCR-Abl inhibits DNA repair mechanisms, leading to genomic instability, which enables cells to accumulate several additional mutations that precipitate disease progression. This transforms the disease from the chronic phase that is stable for several years, towards an acute/blast crisis phase.

Comment.

BCR-Abl was one of the first proteins for which drugs were designed specifically to antagonize the signal transducer of interest, a so-called targeted therapy. Imatinib (Glivec) is a tyrosine kinase inhibitor (TKI) developed by Novartis to specifically inhibit the Abl kinase activity. While unable to completely eradicate the CML cells, it reduced their proliferation rate and delayed the onset of blast crisis. While imatinib is an important drug for the treatment of the majority of CML patients, a small percentage of patients are resistant, or become resistant to imatinib treatment, possibly due to acquired mutations within *BCR-Abl*, which has lead to the development of second- (e.g. dasatinib and nilotinib) and third-line therapies for CML. Of note, TKI therapies to date, have been unsuccessful at eradicating the CML stem cells, the root of the disease; therefore there is a risk of relapse in TKI-treated patients if the therapy is withdrawn or the patients develop chemoresistance.

Summary

■ Most proto-oncogenes and tumor suppressor genes have a function associated with signal transduction, mimicking the effects of persistent mitogenic stimulation, and thereby uncoupling cells from normal external controls.
■ These signaling pathways converge on the machinery that controls the transition of the cell through the G1 phase and prevent cell cycle exit.
■ Additional genes, many of which are targeted by cancer-specific chromosomal translocations and epigenetic regulation, lead to aberrant cell fate decisions that would normally induce apoptosis. The two tumor suppressor proteins, PTEN and p53, which have key roles in determining cell cycle progression and apoptosis, and the genes coding for these proteins, are most frequently disrupted in cancer cells.

Active learning

1. How is progression through the cell cycle regulated?
2. Describe, with reference to specific examples, how growth factor receptor ligation mediates cellular proliferation.
3. Contrast the distinct mechanisms by which a cell can undergo cell death.
4. Explain, with specific examples, how selected mutations within normal growth signals can give rise to a cancerous phenotype in human cells.

Further reading

Chiara Maiuri, M, Zalckvar, E, Kimchi, A, *et al.* Self-eating and self-killing: crosstalk between autophagy and apoptosis. *Nat Rev Mol Cell Biol*. 2007; 8:741–752.

Dick, JE. Stem cell concepts renew cancer research. *Blood*. 2008; 112:4793–4807.

Hollander, MC, Blumenthal, GM, Dennis, PA. PTEN loss in the continuum of common cancers, rare syndromes and mouse models. *Nat Rev Cancer*. 2011; 11:289–301.

Hotchkiss, RS, Strasser, A, McDunn, JE, *et al. Cell Death*. 2009; 361:157–183.

Laplante, M, Sabatini, DM. mTOR signaling in growth control and disease. *Cell*. 2012; 149:274–293.

Levine, AJ, Oren, M. The first 30 years of p53: growing ever more complex. *Nat Rev Cancer*. 2009; 9:749–758.

Malumbres, M, Barbacid, M. Cell cycle, CDKs and cancer: a changing paradigm. *Nat Rev Cancer*. 2009; 9:153–166.

Taylor, RC, Cullen, SP, Martin, SJ. Apoptosis: controlled demolition at the cellular level. *Nat Rev Mol Cell Biol*. 2008; 9:231–241.

Websites and downloads

Nature's Encyclopedia of Life Science. www.els.net/

Kimball's Biology Page. http://users.rcn.com/jkimball.ma.ultranet/BiologyPages/A/Apoptosis.html

KEGG – Human Cell Cycl. www.genome.jp/kegg/pathway/hsa/hsa04110.html

CHAPTER 43

Aging

John W. Baynes

Learning objectives

After reading this chapter you should be able to:

- Describe the relationship between aging and disease.
- Explain the Gompertz plot and how it describes the rate of aging of different species.
- Differentiate between biological and chemical theories of aging.
- Explain the precepts of the free radical theory of aging, including identification of characteristic oxidation products that accumulate in long-lived proteins with age.
- Outline the evidence that the rate of mutation of DNA is an important determinant of the rate of aging.
- Outline the evidence in support of the mitochondrial theory of aging and describe how this theory interfaces with the free radical theory of aging.
- Describe the etiology and pathology characteristic of several diseases of accelerated aging.
- Describe the effects of caloric restriction on the rate of aging of rodents.

Introduction

Aging may be defined as the time-dependent deterioration in function of an organism

While it has broad-based physiologic effects, aging is fundamentally the result of changes in cellular structure and function, biochemistry and metabolism (Table 43.1). The result of aging, even healthy aging, is increased susceptibility to disease and increased probability of death – the endpoint of aging. However, aging is not a disease. Diseases affect a fraction of the population; aging affects all of us, whether it is programmed or stochastic.

Table 43.1

Decline in biochemical and physiologic systems with age

Biochemical	Physiologic
Basal metabolic rate	Lung expansion volume
Protein turnover	Renal filtration capacity (glomerular)
Glucose tolerance	Renal concentration capacity (tubular)
Reproductive capacity	Cardiovascular performance
Telomere shortening	Musculoskeletal system
Oxidative phosphorylation	Nerve conduction velocity
	Endocrine and exocrine systems
	Immunological defenses
	Sensory systems (vision, audition)

With the aging of the population, gerontology and geriatric medicine are becoming increasingly important. This chapter presents an overview on biochemical and physiologic changes associated with aging, in general, and with the aging of specific organ systems. It includes a review of current theories on aging (there are several theories and, in general, the more theories there are, the less we really understand about something) and concludes with a discussion of the relationship between cancer and aging and an update on approaches to lifespan extension.

Aging of complex systems

Excluding genetic defects, childhood disease and accidents, humans survive until about age 50 with limited maintenance requirements or risk of death; then we become increasingly frail and our death rate increases with time, reaching a maximum at about age 76. Our lifespan is affected by our genetics and our

environmental exposure, and our death is usually attributable to failure of a critical organ system (cardiovascular, renal, pulmonary, etc.). The capacity of these interdependent physiologic systems usually declines as a linear function of age, leading to an exponential increase in our age-specific death rate (Fig. 43.1). Historically, improvements in health care and environment have resulted in **'rectangularization' of the survival curve** – our mean lifespan has increased but without a significant effect on our maximum lifespan (see Fig. 43.1A).

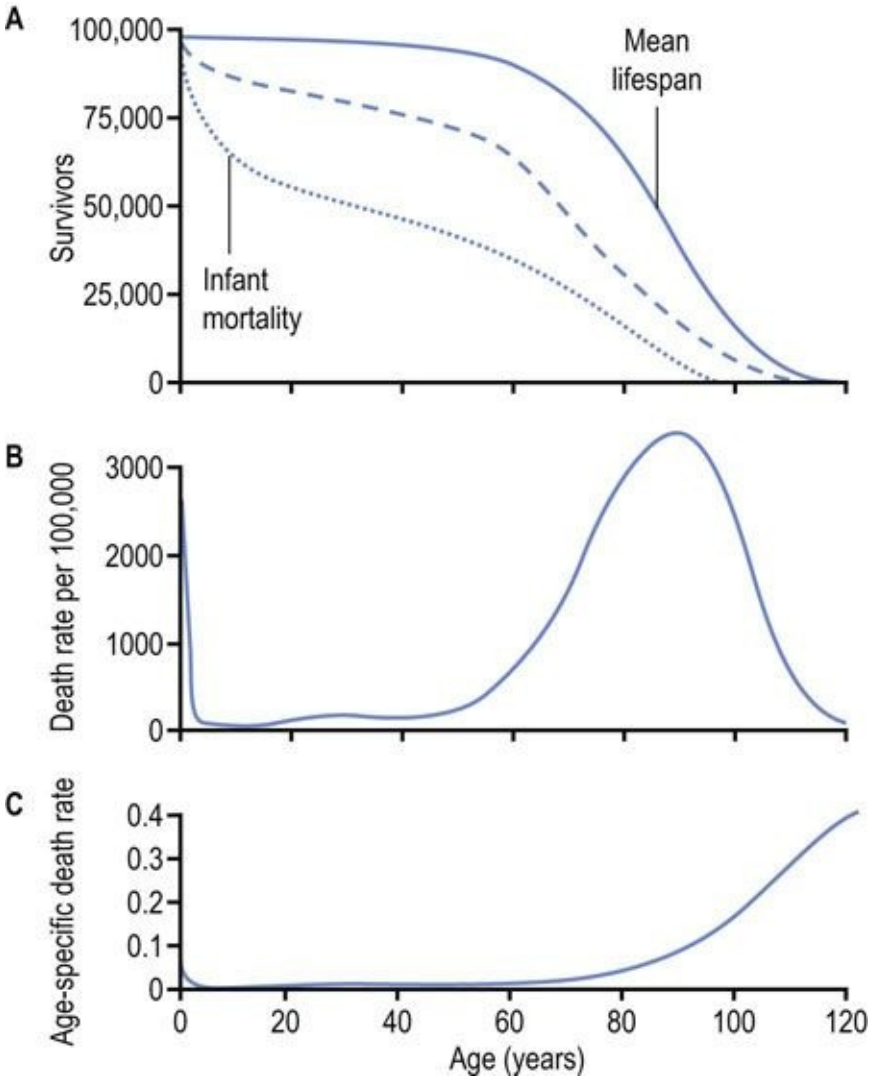


FIG. 43.1 Survival curve and death rate.
(A) Mean lifespan is defined as the age at which 50% of a population survives (or has died). The negative slope of the survival curve reaches a maximum at the mean lifespan of a species. The dotted line describes a survival curve in the Third World where infant mortality and disease significantly decrease mean lifespan. The dashed line describes

the survival curve for the United States in the early part of the 20th century. The solid line applies to 21st-century Europe. **(B)** The **death rate** reaches a maximum at the mean lifespan. **(C)** The **age-specific death rate**, defined as the number of deaths per time at a given age, *e.g.* deaths per 100,000 persons of a specific age per year, increases exponentially with age. The lifespan or maximum lifespan potential (MLSP) is defined as the maximum age attainable by a member of the population, which is about 120 years for humans.

The Hayflick limit – replicative senescence

The replicative capacity of cells decreases with age

Differentiated cells from animals undergo only a limited number of cell divisions (population doublings) in tissue culture, unless they become transformed to cancer cells by mutation or infection with certain viruses. The number of potential cell divisions is greater in longer-lived animals, suggesting a relationship between cell division potential and longevity. Human neonatal fibroblasts will divide about 60 times, then enter a nondividing state, while fibroblasts from mice and rats, which have shorter lifespans, undergo fewer cell divisions in vitro. Cells from younger donors have greater replicative capacity and a greater number of cell divisions in cell culture, but the number of dividing cells decreases with age. This limited doubling capacity, described by Dr Leonard Hayflick, is known as the **Hayflick limit**. The relevance of the Hayflick limit to human aging is still debated – certainly, human cells retain some replicative capacity, even at advanced age, and major tissues, such as muscle and nerve, are largely postmitotic, *i.e.* not actively dividing. However, changes in the metabolism of senescent cells, including decreased responsiveness to hormones and the decline in their synthetic and degradative capacities, *e.g.* in the immunologic and reticuloendothelial systems, may affect our adaptability and susceptibility to stress and age-related diseases, placing limits on our lifespan.

Mathematical models of aging

In poikilotherms, the rate of aging is correlated with temperature, physical activity and metabolic rate

In the early 19th century, Gompertz observed that the age-specific death rate of humans increased exponentially after 35 years of age, and that human survival

curves could be modeled by what is now known as the **Gompertz equation** (Fig. 43.2):

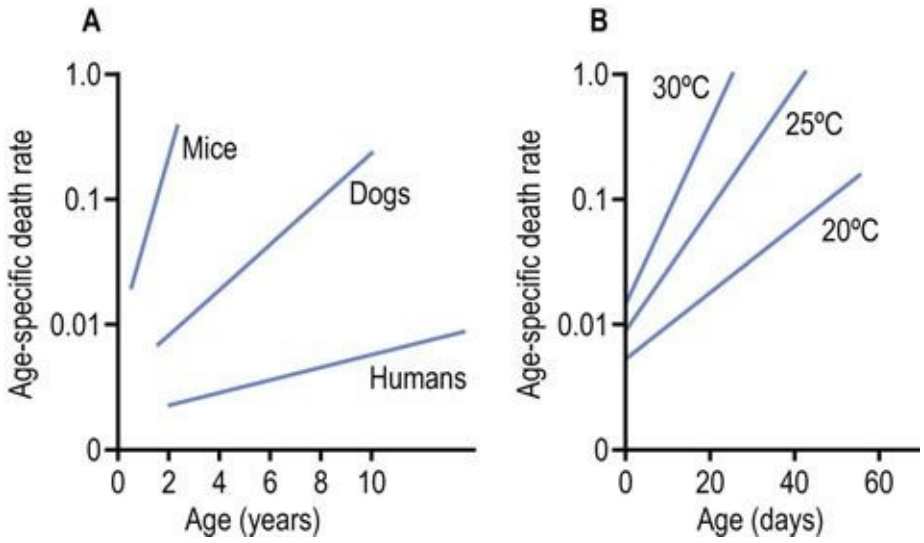


FIG. 43.2 Gompertz plots for humans and other species.
(A) Humans and other vertebrates. (B) Flies raised at various temperatures (adapted from the work of Professor RS Sohal).

$$m_t = Ae^{\alpha t}$$

The term m_t is the age-specific death rate at age t; α is the slope, the effect of time on the death rate; and A, the y-axis intercept, is the death rate at birth. The Gompertz–Makeham equation:

$$m(t) = Ae^{\alpha t} + B$$

adds a constant, B, to correct for the age-independent death rate, *e.g.* as a result of infant mortality or accidents, and provides a better fit to actuarial data.

The Gompertz plots in Figure 43.2 illustrate the time-dependent changes in death rate for three different species of vertebrates and for flies raised at different temperatures. Shorter-lived mammals have a greater age-adjusted rate of death (α = slope), while the death rate for poikilotherms varies with ambient

temperature – flies live longer when grown at lower temperatures. This observation has been interpreted as evidence for '**rate of living**' or '**wear and tear**' theories of aging. Flies, being more active at higher temperature, consume more energy and die more rapidly. Flies that are restrained, *e.g.* in a matchbox, rather than a large carboy, also live longer; wingless flies live longer; and male flies, segregated from females, also live longer. In each case, in small enclosures, without wings, and in the absence of the opposite sex, male flies are less active, have lower basal metabolic rates, and have longer mean and maximum lifespans. None of these strategies for lifespan extension is applicable to humans.

Theories of aging

Theories of aging can be divided into two general categories: biological and chemical

Biological theories treat aging as a genetically controlled event, determined by the programmed expression or repression of genetic information. Aging and death are seen as the orchestrated endstage of birth, growth, maturation and reproduction. Apoptosis (programmed cell death) and thymic involution are examples of genetically programmed events at the level of cells and organs, and the decline in the immunologic, neuroendocrine and reproductive systems may be seen, in a broader context, as evidence for action of a biological clock affecting the integrated functions of an organism. Biological theories attribute differences in lifespan to interspecies differences in genetics but also provide an explanation for the observation that there is a genetic component to longevity within a species, *e.g.* in families with a history of longevity. Differences in lifespan among species are also closely correlated with the efficiency of DNA repair mechanisms. Longer-lived species have more efficient DNA repair processes (Fig. 43.3). Numerous diseases of accelerated aging (progeria) also illustrate the importance of genetics and maintenance of the integrity of the genome during aging.

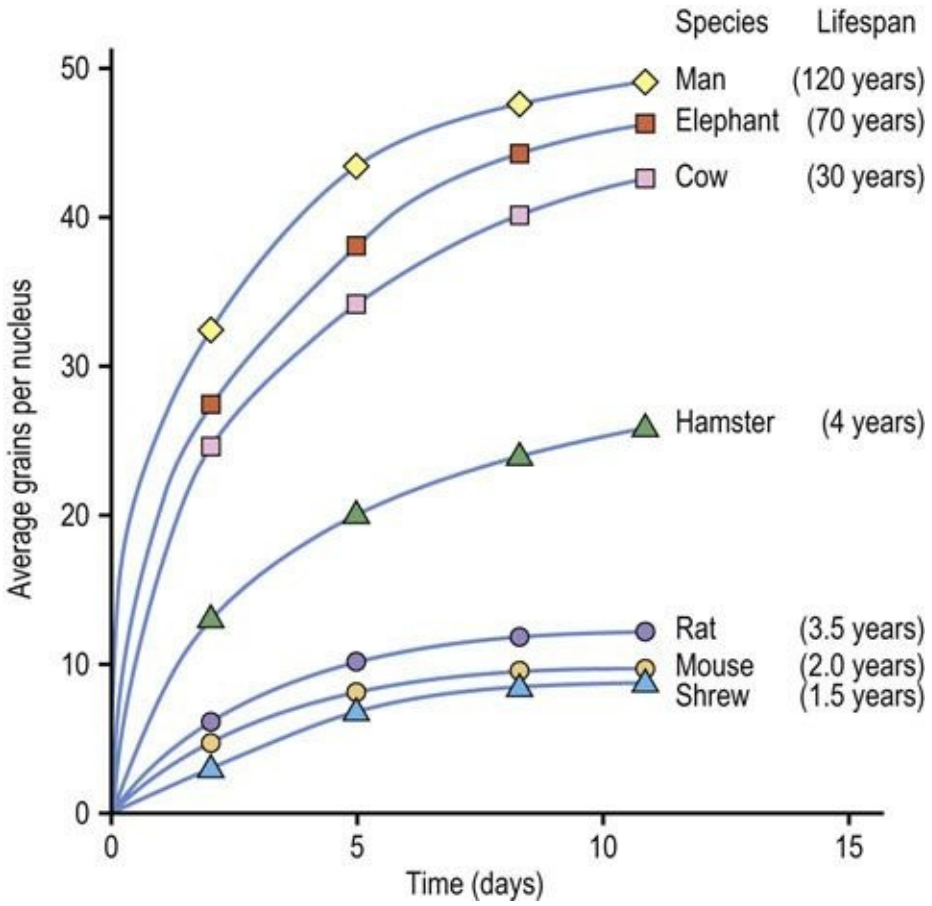


FIG. 43.3 Relationship between DNA repair activity and longevity. Fibroblasts from various species were irradiated briefly, forming thymine dimers and thymine glycol (Chapter 32). The oxidized bases are removed and replaced by excision repair. DNA repair was assessed by the rate of incorporation of a [^{3}H]thymidine tracer into DNA by autoradiography (adapted from Hart RW, Setlow RB: Correlation between deoxyribonucleic acid excision-repair and lifespan in a number of mammalian species. *Proc Natl Acad Sci USA* 71:2169–2173, 1974).

Chemical theories of aging treat it as a somatic process resulting from cumulative damage to biomolecules. At one extreme, the **error-catastrophe theory** proposes that aging is the result of cumulative errors in the machinery for replication, repair, transcription, and translation of genetic information. Eventually, errors in critical enzymes, such as DNA and RNA polymerases or enzymes involved in the synthesis and turnover of proteins, gradually affect the fidelity of expression of genetic information and permit the accumulation of altered proteins. The propagation of errors and resultant accumulation of dysfunctional macromolecules lead eventually to the collapse of the system. Consistent with this theory, increasing amounts of immunologically detectable,

but denatured or modified, functionally inactive enzymes accumulate in cells as a function of age.

More general chemical theories treat aging as the result of chronic, cumulative chemical (**nonenzymatic**) modification, insults or damage to all biomolecules (Table 43.2). Like rust or corrosion, the accumulation of damage with age gradually affects function. This damage is most apparent in long-lived tissue proteins, such as lens crystallins and extracellular collagens, which accumulate chemical modifications with age. These proteins gradually brown with age as a result of formation of a wide range of conjugated compounds with absorbance in the yellow-red region of the spectrum (Fig. 43.4); in the lens, they act as a filter, contributing to the loss of color vision with age. Highly modified crystallins, the major protein in the lens, gradually precipitate, leading to development of cataracts. Chemical damage to the integrity of the genome also occurs, but is more difficult to quantify because of the efficiency of repair processes that excise and repair modified nucleotides. As noted in Table 43.2, there are a number of silent consequences of DNA damage. This damage is primarily endogenous but is enhanced by xenobiotic and environmental agents.

Table 43.2

Age-dependent chemical changes in biomolecules

Protein modification	DNA modification and mutation	Other
Crosslinking	Oxidation	Lipofuscin
Oxidation	Depurination	Inactive enzymes
Deamidation	Substitutions	
D-aspartate	Insertions and deletions	
Protein carbonyls	Inversions and transpositions	
Glycoxidation		
Lipoxidation		

Long-lived proteins, such as lens crystallins and tissue collagens, accumulate damage with age. Modification and crosslinking of proteins occurs as a result of nonoxidative (deamidation, racemization) or oxidative (protein carbonyls) mechanisms or by reactions of proteins with products of carbohydrate or lipid peroxidation (glycoxidation, lipoxidation). Damage to DNA is often silent, *i.e.* modified forms of nucleotides may not accumulate, but the damage increases in the form of mutations resulting from errors in repair.

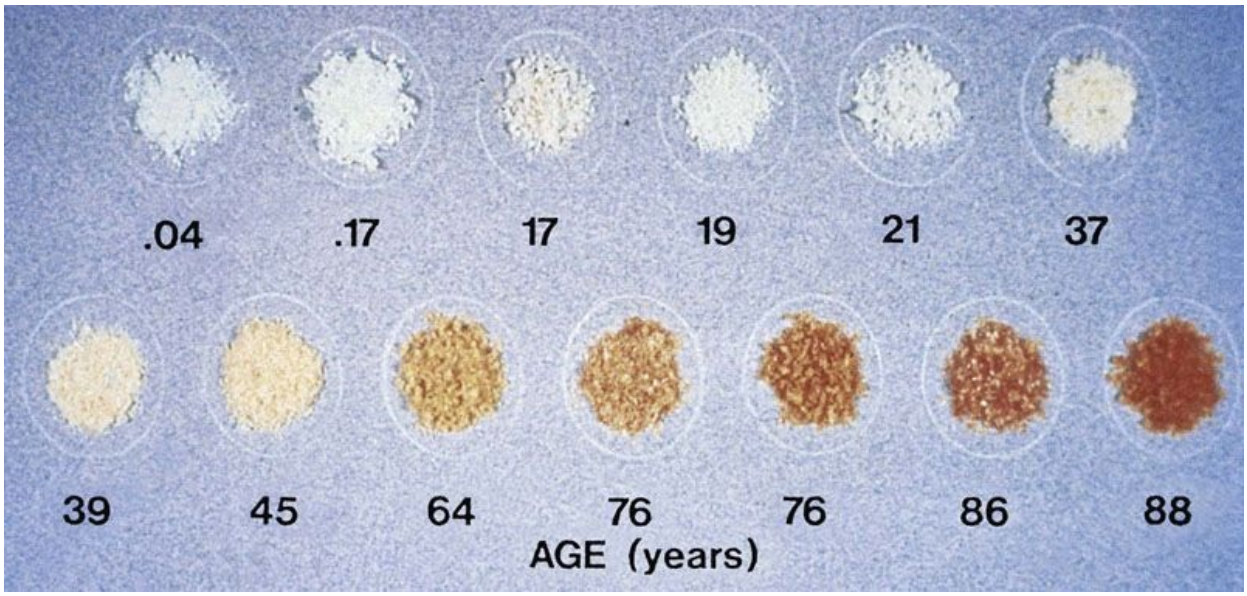


FIG. 43.4 Changes in costal cartilage with age.
Browning is a characteristic feature of the aging of proteins, not just in the lens, which is exposed to sunlight, but also in tissue collagens throughout the body. Crosslinking of proteins also increases with browning. Crosslinking contributes to the gradual insolubilization of lens protein with age. Crosslinking of articular and vascular collagens decreases the resilience of vertebral disks and compliance of the vascular wall with age. These changes in extracellular proteins are similar to changes induced by reaction of carbohydrates and lipids with protein during the cooking of foods, a process known as the **Maillard or browning reaction**. At one level, humans have been described as low-temperature ovens, operating at 37°C, with long cooking cycles (≈75 years). Many of the Maillard reaction products detected in the crust of bread and pretzels have been identified in human crystallins and collagens, and increase with age. (See also discussion of diabetic complications in Chapter 21.)

Organ system theories of aging incorporate various aspects of the above theories. These theories attribute aging to the failure of integrative systems, such as the immunologic, neurologic, endocrine or circulatory system. While they do not assign a specific cause, these theories integrate biological and chemical theories, acknowledging both genetic and environmental contributions to aging.

The free radical theory of aging

The free radical theory of aging is the most widely accepted theory of aging

The free radical theory of aging (FRTA) treats aging as the result of cumulative oxidative damage to biomolecules: DNA, RNA, protein, lipids, and glycoconjugates. From the viewpoint of the FRTA, longer-lived organisms have

lower rates of production of **reactive oxygen species** (**ROS**; Chapter 37), better antioxidant defenses, and more efficient repair or turnover processes. While it is a chemical theory, the FRTA does not ignore the importance of genetics and biology in limiting the production of ROS, and the role of antioxidant and repair mechanisms. It also interfaces with other theories of aging, such as the rate of living theory (because the rate of generation of ROS is a function of the overall rate and/or extent of oxygen consumption) and the crosslinkage theory (because some products of ROS damage crosslink protein). Finally, as a chemical hypothesis, the FRTA does not exclude cumulative chemical damage, independent of ROS, such as racemization and deamidation of amino acids, but focuses on ROS as the primary source of damage and the fundamental cause of aging.

The FRTA is supported by the inverse correlation between basal metabolic rate (rate of oxygen consumption per unit weight) and maximum lifespan of mammals, and by evidence of increased oxidative damage to proteins with age. Protein carbonyl groups, such as glutamic and aminoadipic acid semialdehyde, formed by oxidative deamination of arginine and lysine, respectively, are formed in proteins exposed to ROS. The steady-state level of **protein carbonyls** in intracellular proteins increases logarithmically with age and at a rate inversely proportional to the lifespan of species. Protein carbonyls are also much higher in fibroblasts from patients with **progeria** (accelerated aging), *e.g.* Werner's or Hutchinson–Gilford syndromes, compared to age-matched subjects. Similar concentrations of protein carbonyls are also present in tissues of old rats and elderly humans, arguing that similar changes occur at old age in a range of organisms, regardless of the difference in their lifespans.

Figure 43.5 illustrates the accumulation of two relatively stable amino acid oxidation products in human skin collagen: **methionine sulfoxide and ortho-tyrosine**. These compounds are formed by different mechanisms involving different ROS (Chapter 37) and are present at significantly different concentrations in skin collagen, but increase in concert with age. Other amino acid modifications that accumulate in skin collagen with age include **advanced glycoxidation and lipoxidation end products (AGE/ALEs)**, such as N^{ε}-(carboxymethyl)lysine (CML) and pentosidine (Fig. 43.6 and see Fig. 21.19), and D-aspartate.

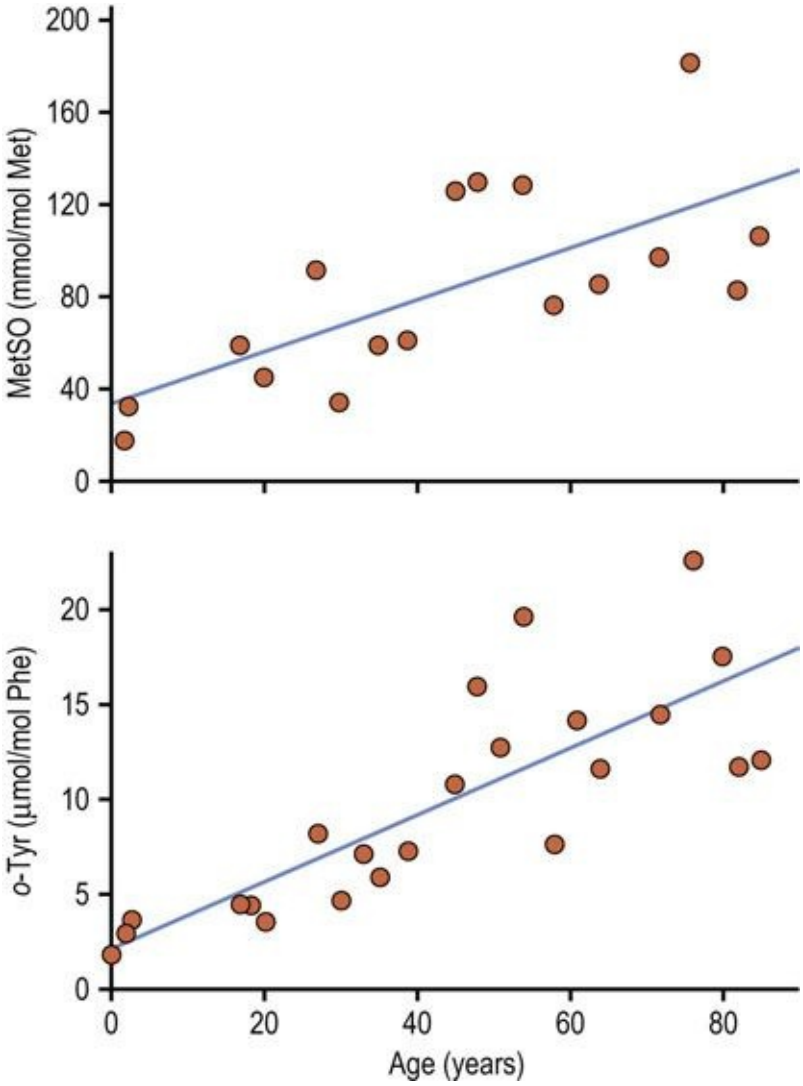


FIG. 43.5 Accumulation of amino acid oxidation products in human skin collagen with age.
Methionine is oxidized to methionine sulfoxide (MetSO) by HOCl or H_2O_2; *ortho*-tyrosine is a product of hydroxyl radical addition to phenylalanine (Phe). Despite a 100-fold difference in their rate of accumulation in collagen, levels of MetSO and *o*-tyrosine correlate strongly with one another, indicating that multiple ROS contribute to oxidative damage to proteins (adapted from Wells-Knecht MC et al: Age-dependent accumulation of *ortho*-tyrosine and methionine sulfoxide in human skin collagen is not increased in diabetes: evidence against a generalized increase in oxidative stress in diabetes. *J Clin Invest* 100:839–846, 1997).

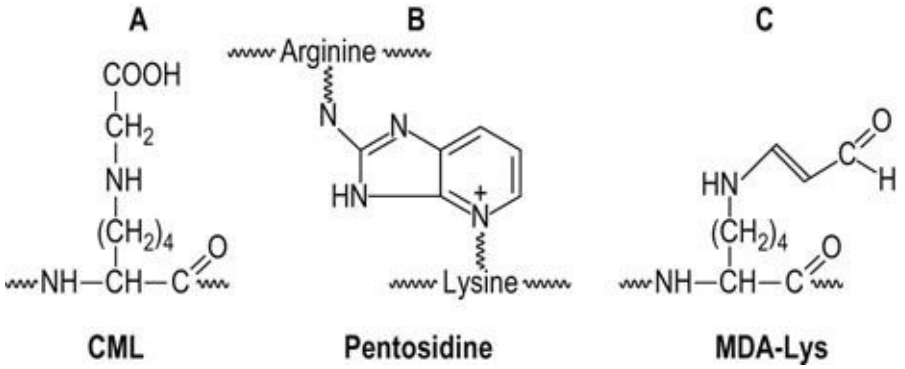


FIG. 43.6 Structure of major advanced glycoxidation and lipoxidation endproducts (AGE/ALEs).

(A) The AGE/ALE, N^{ε}-(carboxymethyl)lysine (CML), which is formed during both carbohydrate and lipid peroxidation reactions. (B) The AGE pentosidine, a fluorescent crosslink in proteins. (C) The ALE, malondialdehyde-lysine (MDA-Lys), a reactive ALE that may proceed to form aminoenimine (RNHCH=CHCH=NR) crosslinks in proteins.

Clinical box Progerias: accelerated aging resulting from defects in DNA repair

Certain genetic diseases are considered models of accelerated aging (progeria). These monogenic diseases display many, but never all, of the features of normal aging; few progeric patients develop dementia or age-related pathologies, such as Alzheimer's disease. The progerias are sometimes described as caricatures of aging, but are useful models for understanding the aging process.

Werner's and Bloom's syndromes are autosomal recessive diseases caused by mutation of distinct DNA helicase genes which have specific roles in repair of damaged DNA. Patients with Werner's syndrome appear normal during childhood, but stop growing in their teens. They gradually show many symptoms of premature aging, including graying and loss of hair, thinning of skin, development of early cataracts, impaired glucose tolerance and diabetes, atherosclerosis and osteoporosis, and increased rates of cancer. Death usually occurs in their mid-40s from cardiovascular disease. Fibroblasts from Werner's patients divide only about 20 times in cell culture, compared to 60 for normal

cells, and have higher levels of protein-bound carbonyl groups, an indicator of increased oxidative stress.

Bloom's syndrome is characterized by increased frequency of chromosomal breaks, dwarfism, photosensitivity and increased frequency of cancer and leukemia; death occurs typically in the mid-20s. Ataxia-telangiectasia, or fragile chromosome syndrome, is associated with increased loss of telomeres with cell division and deficiency in repair of double-strand DNA breaks. It is caused by a defect in a protein kinase involved in signal transduction, cell cycle control and DNA repair.

Hutchinson–Gilford syndrome is a severe, pediatric form of progeria. Patients have many of the symptoms of Werner's syndrome, but the symptoms appear at an earlier age and death usually occurs by the mid-20s. This syndrome is caused by a defect in a gene for lamin, a component of the nuclear envelope. Hutchinson–Gilford is one of several distinct syndromes associated with lamin mutations, which cause an increase in nuclear fragility and aberrant mRNA splicing; as in Werner's syndrome, cultured fibroblasts become prematurely senescent. These progeric diseases illustrate the importance of efficient repair of DNA for normal growth and aging.

D-**aspartate** is a nonoxidative modification of protein that is formed by spontaneous, age-dependent racemization of L-aspartate, the natural form of the amino acid in protein. The more rapid turnover of skin, compared to articular, collagen yields a lower rate of accumulation of D-aspartate in skin collagen with age and also explains the lower rates of accumulation of AGE/ALEs in skin versus articular collagen. AGE/ALEs are even higher in lens crystallins, which have the slowest rate of turnover among proteins in the body. **Deamidation** of asparagine and glutamine is another nonoxidative chemical modification that increases with age in proteins; it has been described primarily in intracellular proteins.

The rate of accumulation of these modifications depends on the rate of turnover of the collagens (Fig. 43.7) and is accelerated by hyperglycemia and

hyperlipidemia in diabetes and atherosclerosis. The increase in AGE/ALEs and oxidative crosslinking of collagen is thought to impair the turnover and contribute to the thickening of basement membranes with age. Increased age-adjusted levels of AGE/ALEs in collagen are implicated in the pathogenesis of complications of diabetes and atherosclerosis. These products are also increased together in the brain in various neurodegenerative diseases, including Alzheimer's and Parkinson's disease and Creutzfeld–Jakob (prion) disease.

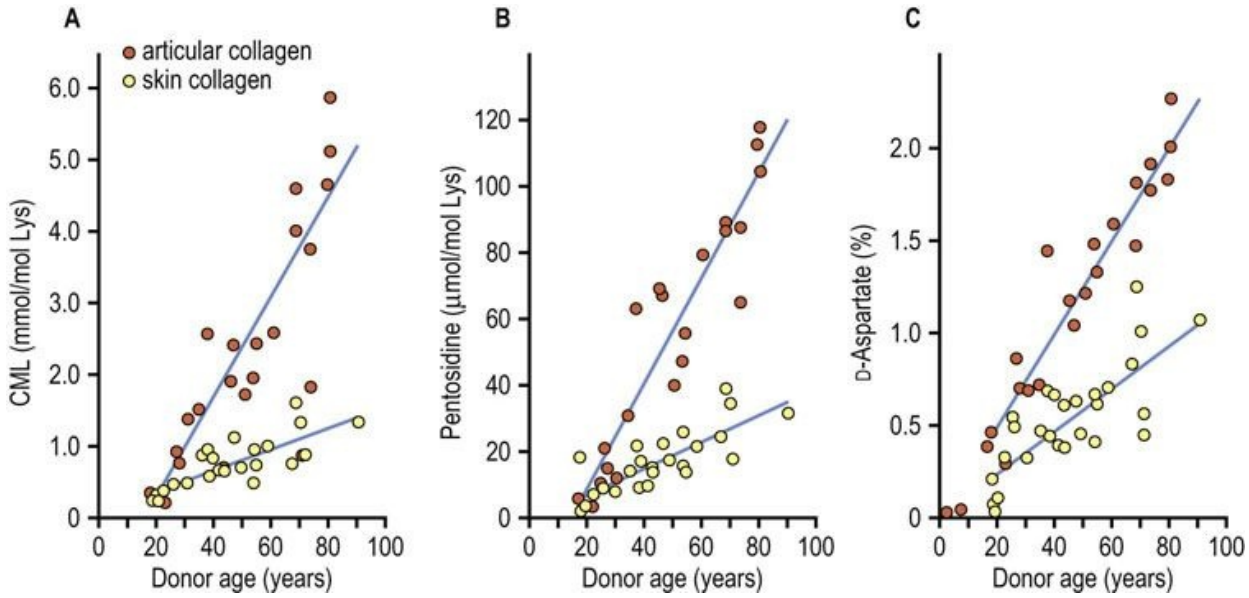


FIG. 43.7 Accumulation of advanced glycoxidation and lipoxidation end products (AGE/ALEs) and D-Aspartate in articular and skin collagens with age.
N^{ε}-(carboxymethyl)lysine (CML) is formed by oxidative mechanisms from glycated proteins or reaction of glucose, ascorbate or lipid peroxidation products with protein. The fluorescent crosslink pentosidine is formed by oxidative reaction of glucose or ascorbate with proteins. D-Aspartate is formed nonoxidatively by racemization of L-aspartate residues in protein. Tissue levels of oxidative and nonoxidative biomarkers correlate with one another, and differences in their rates of accumulation in articular and skin collagen result from differences in rates of turnover of these collagens (adapted from Verzijl N et al: Effect of collagen turnover on the accumulation of advanced glycation end products. *J Biol Chem* 275:39027–39031, 2000).

The age pigment **lipofuscin** is a less well-characterized but characteristic biomarker of aging. It accumulates in the form of fluorescent granules, derived from lysosomes, in the cytoplasm of postmitotic cells at a rate that is inversely related to species lifespan. It is considered the accumulated, indigestible debris of reactions between lipid peroxides and proteins. Lipofuscin may account for

10–15% of the volume of cardiac muscle and neuronal cells at advanced age, and its rate of deposition in cardiac myocytes in cell culture is accelerated by growth under hyperoxic conditions. In flies, the rate of accumulation of lipofuscin varies directly with ambient temperature and activity, and inversely with lifespan, consistent with the effects of these variables on lifespan (see Fig. 43.2B).

In summary, there is a wide range of chemical modifications, both oxidative and nonoxidative, that accumulate in proteins with age. While attention is often focused on modification of protein, the real damage from free radicals and oxidative stress is at the level of the genome; if the DNA is not repaired correctly, the cell will die, its capacity may be impaired or the damage will be propagated. Damage to DNA accumulates not in the form of modified nucleic acids but as chemically 'silent' errors in repair – insertions, deletions, substitutions, transpositions and inversions of DNA sequences – that affect the expression and structure of proteins. Because repair is fairly efficient in humans compared to other animals, and the composition of DNA does not change on repair, mutations in DNA are not detectable in tissues by conventional analytic techniques. However, the presence of oxidized pyrimidines and purines in urine (see Figure 37.6) provides evidence of chronic oxidative damage to the genome.

Mitochondrial theories of aging

Clinical box Biomarkers of oxidative stress and aging

Advanced glycoxidation and lipoxidation end products (AGE/ALEs) are formed by reaction of proteins with products of oxidation of carbohydrates and lipids (see Fig. 43.6). Some compounds, such as N^{ε}-(carboxymethyl)lysine (CML), may be formed from either carbohydrates or lipids; others, such as pentosidine, are formed only from carbohydrates, and others, such as the malondialdehyde adduct to lysine, are formed exclusively from lipids. Carbohydrate sources of AGEs include glucose, ascorbate, and glycolytic intermediates; ALEs are derived from oxidation of polyunsaturated fatty acids in phospholipids. Lysine, histidine and cysteine residues are the major sites of AGE/ALE formation in protein. Over 30 different AGE/ALEs have been

detected in tissue proteins, and many of these are known to increase with age. AGE/ALEs are useful biomarkers of the aging of proteins and their exposure to oxidative stress.

Clinical box Alzheimer's disease: oxidative stress in neurodegenerative disease

Alzheimer's disease (AD) is the most common form of progressive cognitive deterioration in the elderly. It is characterized microscopically by the appearance of **neurofibrillary tangles** and senile plaques in cortical regions of the brain. The tangles are localized inside neurons, and are rich in τ (tau) protein, which is derived from microtubules; it is hyperphosphorylated and polyubiquitinated. Plaques are extracellular aggregates, localized around amyloid deposits, formed from insoluble peptides derived from a family of **amyloid precursor proteins**. AD affects primarily cholinergic neurons, and drugs that inhibit the degradation of acetylcholine within synapses are a mainstay of therapy. A similar approach is used for preservation of dopamine in dopaminergic neurons in Parkinson's disease, *i.e.* by inhibiting the degradative enzyme monoamine oxidase.

Several studies have shown that both AGEs and ALEs are increased in tangles and plaques in the brain of AD patients, compared with age-matched controls. Other indicators of generalized oxidative stress in the AD brain include increased levels of protein carbonyls, nitrotyrosine and 8-OH-deoxyguanosine, all detected by immunohistochemical methods. The amyloid protein is toxic to neurons in cell culture and catalyzes oxidative stress and inflammatory responses in glial cells. Significant quantities of decompartmentalized, redox-active iron, a catalyst of Fenton reactions (Chapter 27) are also detectable histologically in the AD brain and can be removed reversibly (in

vitro) by treatment with chelators, such as desferrioxamine. Based on these data, oxidative stress is strongly implicated in the development and/or progression of AD, and chelators are being evaluated clinically for treatment of AD. Epidemiologic studies also indicate that long-term treatment with nonsteroidal anti-inflammatory drugs, such as ibuprofen and acetaminophen, may reduce the risk of AD and delay its onset or slow its progression.

Advanced concept box Aging of the circulatory system

The **extracellular matrix** of the aorta and major arteries becomes thicker and more highly crosslinked with age, contributing to both the decrease in elasticity and the ability of the endothelium to dilate blood vessels in response to physical and chemical stimuli. These changes occur naturally with age, independent of pathology, but may account for the increase in cardiovascular risk in the elderly. AGEs and ALEs are implicated in the crosslinking of the vascular extracellular matrix, explaining the age-adjusted increase in arterial crosslinking in diabetes and dyslipidemia. Increases in AGE/ALEs and protein crosslinking are also implicated in the altered filtration properties of the renal glomerular basement membrane in diabetes.

Mitochondrial DNA is particularly susceptible to oxidative damage

Mitochondrial theories of aging are a blend of biological and chemical theories, treating aging as the result of chemical damage to mitochondrial DNA (mtDNA). Mitochondria contain proteins specified by both nuclear and mitochondrial DNA but only 13 mitochondrial proteins are encoded by mitochondrial DNA. While this may seem trivial, these include essential

subunits of the three proton pumps and ATP synthase. MtDNA is especially sensitive to mutations: mitochondria are the major site of ROS production in the cell (see Fig. 37.4), mtDNA is not protected by a sheath of histones, and mitochondria have limited capacity for DNA repair.

Mitochondrial diseases commonly involve defects of energy metabolism, including the pyruvate dehydrogenase complex, pyruvate carboxylase, electron transport complexes, ATP synthase, and enzymes of ubiquinone biosynthesis. These defects can be caused by mutations in both nuclear and mitochondrial DNA, but mtDNA suffers many more mutations than nuclear DNA. Such defects often result in the accumulation of lactic acid because of impaired oxidative phosphorylation, and may cause cell death, especially in skeletal (myopathies) and cardiac muscles (cardiomyopathies) and nerve (encephalopathies), all of which are heavily dependent on oxidative metabolism. The number of mitochondria and multiple copies of the mitochondrial genome in the cell may provide some protection against mitochondrial dysfunction as a result of mutation, but loss of fully functional mitochondria, and sometimes the number of mitochondria, is a characteristic feature of aging.

Genetic models of increased lifespan

Advanced concept box Telomeres: a clock of aging

Telomeres are the repetitive sequences at the ends of chromosomal DNA, typically thousands of copies of short, highly redundant, repetitive DNA, TTAGGG in humans (Chapter 32). DNA polymerase requires a double-stranded template for replication; RNA primers at the 5′-end of the template serve to initiate DNA synthesis. However, at the extreme ends of the chromosomes, DNA synthesis is restricted, because there are no sequences further upstream for DNA primase engagement. Therefore, each round of chromosome replication results in chromosome shortening. The enzyme telomerase is a reverse transcriptase containing an RNA with a sequence complementary to the telomere DNA. It functions to maintain the length of telomeres at the 3′-end of chromosomes. Telomerase is found in fetal tissues, adult germ cells and in tumor cells. but the somatic cells of multicellular organisms lack telomerase activity. This has led to the hypothesis that shortening of the telomere may contribute to the Hayflick limit and is involved in aging of multicellular organisms. Increased expression of telomerase in human cells results in elongated telomeres and an increase in the longevity of those cells by at least 20 cell doublings. Cells from individuals with premature aging diseases (progeria) also have short telomeres. In contrast, cancer cells, which are immortal, express an active telomerase activity. All of these observations suggest that the decrease in telomere length is associated with cellular senescence and aging. Knockout mice, in which the telomerase gene has been deleted, have chromosomes lacking detectable telomeres. These mice have high frequencies of aneuploidy and chromosomal abnormalities. The disease, autosomal dyskeratosis congenita, features a mutation in the telomerase locus, with inability of somatic cells to reconstitute

their telomeres, and hence loss of epidermis and hematopoietic marrow. This disease has many of the characteristics of accelerated aging.

Advanced concept box Aging of muscle: damage to mitochondrial DNA

Old age is characterized by a general decrease in skeletal muscle mass (**sarcopenia**) and strength, as a result of a decrease in both the number of motoneurons and the number and size of myofibers. The fiber loss is accompanied by an increase in interstitial, fibrous connective tissue, and a reduction in capillary density, which limits the blood supply. The decrease in muscle mass and strength contributes to frailty and increased risk of mortality. The loss in skeletal muscle mass may also contribute to glucose intolerance in the elderly as a result of the decreasing mass of tissue available to take up glucose from blood.

One of the major changes in muscle biochemistry with age is an increase in the number of muscle cells with mitochondria deficient in cytochrome oxidase, which limits the muscle's ability to do work. As mitochondria become less efficient in oxidizing NADH, they become more reduced, and the accumulation of partially reduced ubiquinone (semiquinone) promotes the reduction of molecular oxygen, leading to increased superoxide production in older mitochondria (see Chapter 37). Under these conditions, when oxidative phosphorylation is impaired, cells appear to generate ATP primarily by glycolysis. NADH is also oxidized extramitochondrially, primarily by NADH oxidases in the plasma membrane, which produce hydrogen peroxide, but no ATP.

$$\text{NADH oxidase}: \quad \text{NADH} + \text{H}^+ + \text{O}_2 \rightarrow \text{NAD}^+ + \text{H}_2\text{O}_2$$

These changes are observed in both cardiac and skeletal muscle

and appear to result from major, random deletions in mitochondrial DNA (25–75% of total mtDNA), which are then amplified by clonal expansion, leading to fiber atrophy and breakage. The muscle fiber is only as strong as its weakest link, so that small regions of fiber loss affect overall muscle capacity. Fortunately, sarcopenia can be delayed and partially reversed by resistance exercise, thus the emphasis on regular exercise among the elderly.

The effect of genetics on longevity is readily apparent in animal models

Different strains of mice vary by more than twofold in lifespan, and there are also significant differences in the lifespan of male and female mice of the same strain raised under identical conditions. Deficiencies in some hormones or defects in their receptors or postreceptor signaling pathways have a significant effect on mouse lifespan. Profound effects are observed in Ames and Snell dwarf mice. These mice have different pituitary defects, both resulting in negligible secretion of growth hormone (GH; stimulates IGF-1 secretion by liver), thyroid-stimulating hormone and prolactin (see Chapter 39). Their body weights are decreased as young adults by about 35% and their maximum lifespan increased by about 45% compared to littermates, but oddly they become obese with age. Similar effects on weight and lifespan are observed in mice with defects in GH or IGF-1 receptors or signal transduction. Many of these strains are fragile: Ames and Snell dwarfs are hypothyroid, hypoglycemic and hypoinsulinemic, and have low body temperature; they have impaired reproductive capacity, are more susceptible to infection, and require special housing conditions to maintain body temperature – but they live longer! Treatment of hypothyroidism in Snell dwarfs resulted in a restoration of a near normal lifespan, while hypophysectomy of young rats increases their maximum lifespan by 15–20%. Thus, three hormones that have a profound effect on metabolism and growth (**growth hormone, IGF-1 (and insulin) and thyroxine**) also have profound effects on lifespan.

In humans, the most significant genetic determinant of lifespan is sex: women live longer than men. Genetics accounts for an estimated 20–50% of the remaining variance in lifespan, the other 50–80% being attributed to

environment and random developmental variations. It was estimated (in 2008) that there are at least 30 genes that have a significant effect on human lifespan. The cross-breeding of human populations and the many allelic combinations of these genes may obscure effects seen in inbred strains of worms or rodents. However, there is a recent report that Ashkenazi Jews who live past age 95 have a higher frequency of mutations in the gene for the IGF-1 receptor (IGF-1R). There are other genes or gene products that are associated with increased longevity in humans, *e.g.* variants in ApoE, ApoC3 and CETP. However, these genes seem to increase mean lifespan, probably by modulating the cardiovascular effects of dietary cholesterol, rather than cause an increase in maximum lifespan.

Anti-aging interventions – what works and what doesn't

Antioxidant supplements

Antioxidant supplements may improve health, but do not increase lifespan

Based on the FRTA, it seems reasonable to speculate that antioxidant supplementation should have an effect on longevity. In fact, however, there is no rigorous, reproducible experimental evidence that antioxidant supplements have any effect on maximum lifespan of humans or other vertebrates. At the same time, antioxidant supplements, most of which include vitamins, may improve health, particularly in persons with vitamin deficiencies. Thus, effects of antioxidant therapy on mean (and healthy) lifespan are not unexpected. Failure to affect maximum lifespan may result from the fact that there are so many mechanisms for production and control of free radicals and inhibiting or reversing damage to biomolecules. Many of these processes depend on the activity of enzymes that detoxify ROS or regenerate endogenous antioxidants. These enzymes, such as superoxide dismutase and glutathione peroxidase (Chapter 37), are induced in response to oxidative stress and may also be repressed during times of low oxidative stress. Thus, the body may respond to maintain a homeostatic balance between pro-oxidant and antioxidant forces (see Fig. 37.2), countering efforts to enhance antioxidant defenses. This response may be essential, for example, to maintain effective bactericidal activity during the respiratory burst accompanying phagocytosis.

Calorie restriction

Caloric restriction is the only regimen known to increase lifespan in animals

Calorie restriction (CR) is the only intervention that consistently extends maximum lifespan in a variety of species, including mammals, fish, flies, worms, and yeast. Reduction in total caloric intake is the essential feature of this

intervention, *i.e.* the beneficial, life-extending effects are observed whenever CR is applied and regardless of dietary composition, although early and prolonged intervention has more impressive effects. As shown in Figure 43.8, CR leads to a significant increase in both the mean and maximum lifespan of laboratory rats, equivalent to extending human lifespan to about 180 years. Calorie-restricted rats have fewer muscle fibers lacking in cytochrome oxidase and decreased levels of deletions in muscle mitochondrial DNA. CR mice also have lower levels of inducible genes for hepatic detoxification, DNA repair and response to oxidative stress (heat shock proteins), suggesting a lower rate of oxidative stress and damage to proteins and DNA.

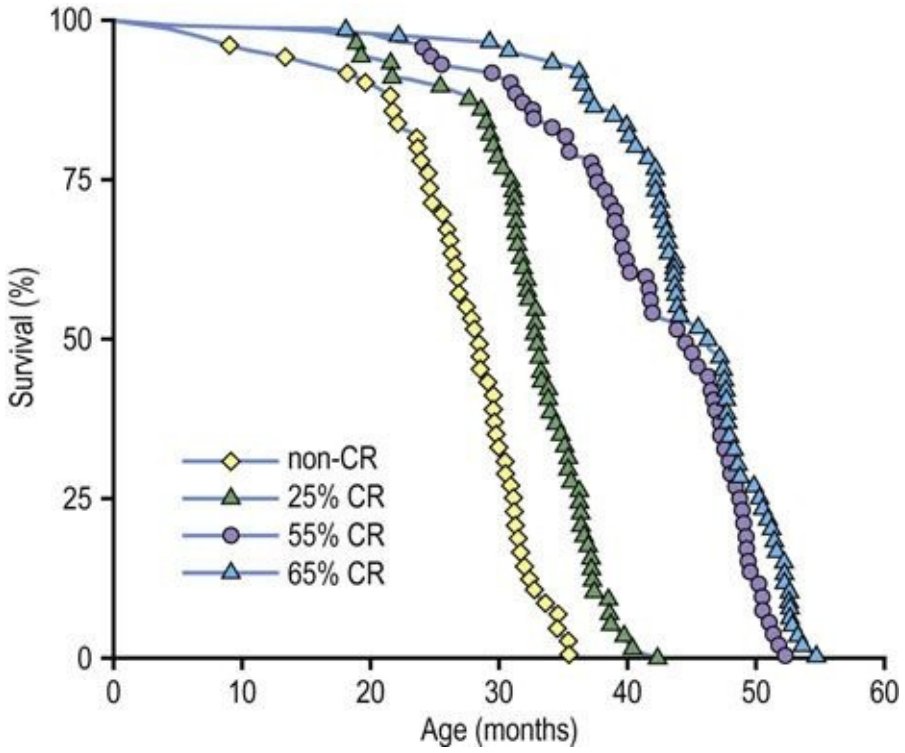


FIG. 43.8 Calorie restriction (CR) extends longevity of mice.
The non-CR group received food ad libitum. Other groups were restricted by 25%, 55% and 65% of the ad lib diet, starting at 1 month of age (adapted from Weindruch R et al: Retardation of aging in mice by dietary restriction. *J Nutr* 116:651–654, 1986).

Caloric restriction delays the onset of age-related diseases, including cancer (Fig. 43.9)

CR is the most potent, broad-acting cancer prevention regimen in rodents. It is argued that the extension of maximum lifespan by CR is achieved by delaying the onset of cancer. Long-lived animals are more efficient in protecting their genome and thereby delaying the onset of cancer, but CR may limit damage even more, preserving the integrity of the genome and thereby leading to a longer lifespan. While long-term CR has not been tested in humans, obesity, at the opposite end of the weight spectrum, is a pro-inflammatory state and is a risk factor for cancer in humans.

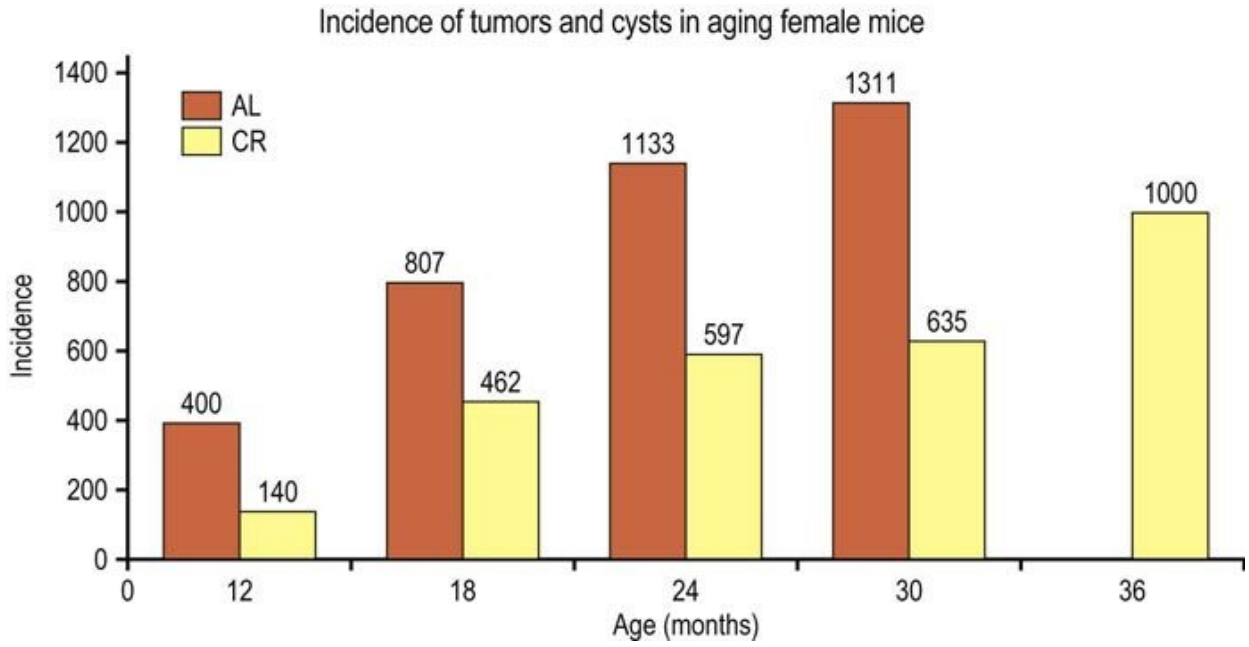


FIG. 43.9 Effect of calorie restriction (CR) on development of tumors in mice. Over 1000 mice, 50:50 male and female, from four different genotypes were divided into two groups, one fed ad libitum (AL), the other receiving 60% of the caloric intake of the control group (CR), but with comparable intake of vitamins, minerals and micronutrients. Cohorts of animals were sacrificed at specified times and total lesions (tumors plus cysts) were measured. At 24 months, 51% AL and 13% CR mice had tumors. Note the absence of tumors in control mice at 36 months – all of the control mice are dead! CR extended the mean and maximum lifespan of the mice and also delayed the onset of cancer (adapted from Bronson RT, Lipman RD: Reduction in rate of occurrence of age related lesions in dietary restricted laboratory mice. *Growth Dev Aging* 55:169–184, 1991).

In CR experiments, it has been difficult to differentiate between the effects of dietary restriction on energy expenditure (rate of living) versus the reduction in body weight or adipose tissue mass that accompanies dietary restriction. FIRKO

(adipose tissue (fat) insulin receptor knockout) mice have a 15–25% decrease in body mass, largely because of a 50% decrease in fat mass. However, these mice consume identical amounts of food per day as control littermates, actually more than the control animals when normalized to their body weight. They also have a 20% increase in lifespan, suggesting that the decrease in body or fat mass is more important than caloric intake in determining maximum lifespan potential. In another study, overexpression of the gluconeogenic enzyme PEPCK in skeletal muscle produced a leaner mouse, with 50% of body weight and 10% of fat mass, compared to controls. These mice were seven times as active and ate 60% more than control mice, but lived longer and had longer reproductive life. Overall, the decrease in body weight or adiposity during CR, rather than the decrease in food consumption, appears to have the greater effect on lifespan extension. One general outcome from these dietary and genetic experiments is that mitochondrial efficiency, measured as lower rates of ROS/ATP production, appears to be an important determinant of longevity.

Studies on CR in longer-lived, primate species have been under way since the 1980s. From these studies, there is clear evidence that monkeys on CR are more active and younger in appearance, have better insulin sensitivity and plasma lipid profiles and decreased risk for diabetes, have better overall cardiovascular and renal health, experience less age-related sarcopenia and brain atrophy, and have decreased risk for cancer, compared to age-matched normally fed animals. However, the evidence of lifespan extension is weak and still controversial (Fig. 43.10). Even if CR can be shown to extend lifespan in monkeys, it is unlikely that humans will be able to adopt the strict dietary control required for this regimen. However, similar improvements in health have been observed in shorter-term studies with humans, *i.e.* improvement in fasting glucose, insulin sensitivity and plasma lipid profile, and decreased blood pressure, compared to a matched control group. Understanding the biological mechanisms of the effects of CR may lead to alternative strategies that mimic CR and possibly extend, at least, the healthy lifespan of humans.

FIG. 43.10 Effects of calorie restriction in primates.
The animal on the left was raised with calorie restriction, while that on the right was fed a normal diet. A study from the University of Wisconsin in 2009 concluded that calorie restriction extended the lifespan of Rhesus monkeys, while another by the US National Institutes of Health in 2012, also with Rhesus monkeys, concluded there was no effect. Studies with other primate species are still in progress, so the issue is not fully resolved; however, it is clear in all studies that calorie restriction produces a healthier phenotype, with fewer age-related chronic diseases and decreased risk for cancer (with permission from the National Institutes of Health).

Summary

■ Aging is characterized by a gradual decline in the capacity of physiologic systems, leading eventually to failure of a critical system, then death.
■ At the biochemical level, aging is considered the result of chronic chemical modification of all classes of biomolecules.
■ According to the free radical theory of aging, ROS are the primary culprits, causing alterations in the sequence of DNA (mutations) and structure of proteins. Longevity is achieved by developing efficient systems to limit and/or repair chemical damage.
■ Caloric restriction is, at present, the only widely applicable mechanism for delaying aging and extending the mean, healthy, and maximum lifespan of species.
■ CR appears to work, in part, by inhibiting the production of ROS and limiting damage to biomolecules, delaying many of the characteristic features of aging, including cancer.

Active learning

1. Discuss the nature of protein carbonyls and lipofuscin and their relevance to aging.
2. Discuss the relative importance of chemical damage to protein and DNA during aging.
3. Review recent literature on mouse genetic models of mammalian aging and discuss the relationship between growth rate, obesity, calorie restriction and aging in the mouse.
4. Nearly a dozen genes have been identified that, when mutated, extend the lifespan of animals. Why are the wild-type genes preserved in the gene pool?
5. Discuss the evidence that caloric restriction increases the mean, healthy and maximum lifespan of primates.

Further reading

Baur, JA, Ungvari, Z, Minor, RK, *et al.* Are sirtuins viable targets for improving healthspan and lifespan?

Nat Reviews. 2012; 11:443–461.

Cefalu, CA. Theories and mechanisms of aging. *Clin Geriatr Med*. 2011; 27:491–506.

Dai, DF, Rabinovitch, PS, Ungvari, Z. Mitochondria and cardiovascular aging. *Circ Res*. 2012; 110:1109–1124.

Ford, ES, Zhao, G, Tsai, J, *et al*. Low-risk lifestyle behaviors and all-cause mortality: finding from the National Health and Nutrition Examination Survey (NHANES) III Mortality Study. *Am J Public Health*. 2011; 101:1922–1929.

Jenny, NS. Inflammation in aging: cause, effect, or both? *Discov Med*. 2012; 13:451–460.

Kemnitz, JW. Calorie restriction and aging in nonhuman primates. *ILAR J*. 2011; 52:66–77.

Libert, S, Guarente, L. Metabolic and neuropsychiatric effects of calorie restriction and sirtuins. *Annu Rev Physiol*. 2013; 75:669–684.

Semba, RD, Nicklett, EJ, Ferrucci, L. Does accumulation of advanced glycation end products contribute to the aging phenotype? *J Gerontol A Biol Sci Med Sci*. 2010; 65:963–975.

Websites

Caloric restriction. www.crsociety.org.

Lecture notes on gerontology:

www.pathguy.com/lectures/aging.htm

www.benbest.com/lifeext/aging.html

Progeria. www.progeriaresearch.org/index.html.

Telomeres. www.telomeres.net/.

APPENDIX 1:

Selected Clinical Laboratory Reference Ranges

Yee Ping Teoh and Marek H. Dominiczak

Reference ranges

Reference ranges are values of a given substance (analyte) obtained in a reference population, which is usually a group of healthy individuals

These values represent the physiologic quantities of a substance to be expected in healthy persons. The term ‘reference range’ is preferred over ‘normal range’ because while the reference population can be clearly defined, no clear definition exists for what is ‘normal’ in a clinical sense.

Deviation above or below the reference range may be associated with a disease process, and the severity of the disease process may be associated with the magnitude of the deviation. The ideal reference population is one that is appropriate for the individual's age, sex and ethnicity. The reference range may also vary with the instrument and assay procedure used for the measurement.

Distribution of values within the reference population

When data from a large cohort of healthy subjects fit a Gaussian distribution, the reference limits are defined as two standard deviations above and below the mean

This constitutes the central 95% interval of the distribution. However, many analyte distributions are non-Gaussian and these values are usually mathematically transformed (e.g. logarithmic, reciprocal, exponential transformation) to yield a Gaussian distribution.

Interpretation of laboratory results in individual persons

The interpretation of results of laboratory tests is based on comparison with reference values

If a value falls outside the reference interval, it signifies that – with the probability of 95% – the result is different from the reference population. Note that this does not necessarily mean that it is abnormal: by definition, 5% of individuals in a reference population (1 in 20 subjects) will have results outside the reference range. However, the further the result is from the reference range, the greater is the probability that it is associated with a pathology.

Clinical decision limits

In some cases instead of reference values, clinical decision limits are the basis for interpretation

Clinical decision limits are used in the interpretation of tests such as plasma glucose, lipids, and cardiac troponin measurements. These limits, or cut-off points, are usually derived from epidemiologic studies linking the levels of an analyte with the risk of a particular condition.

The reference values given here are taken from the UK Pathology Harmonization Reference Ranges and from the NHS Greater Glasgow and Clyde Clinical Biochemistry Service (UK) test menu.

In the tables below we have only included the basic tests that will be helpful in the interpretation of information presented in the Clinical Boxes throughout this book. The reader is referred to Further Reading for comprehensive lists of tests offered in clinical laboratories.

The reference intervals are given in SI and conventional units, wherever possible, with the factor for converting from SI to conventional units.

- To convert from an SI unit to a conventional unit, multiply by the conversion factor.
- To convert from a conventional unit to an SI unit, divide by the conversion factor.
- Unless indicated, the ranges given are for serum/plasma concentrations.

These values are given **for guidance only**. Remember that reference ranges may be different in different laboratories. Before interpreting laboratory tests in a clinical situation, it is essential to verify these with the local lab. Also, precisely as the Clinical Boxes aim to illustrate, **laboratory tests should always be interpreted in the context of medical history and physical examination**.

Table 1

Blood gases

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Arterial pH		Negative logarithm of the H^+ ion activity	7.35–7.45
H^+ ion activity	35–45 nmol/L		
Arterial oxygen partial pressure (PaO_2)	10.5–13.5 kPa	7.5	79–101 mmHg
Arterial carbon dioxide partial pressure ($PaCO_2$)	4.6–6.0 kPa	7.5	34–45 mmHg

Table 2

Serum electrolytes and markers of renal function

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Sodium	133–146 mmol/L	1.0	133–146 mEq/L
Potassium	3.5–5.3 mmol/L	1.0	3.5–5.3 mEq/L
Chloride	95–108 mmol/L	1.0	95–108 mEq/L
Bicarbonate	22–29 mmol/L	1.0	22–29 mEq/L
Anion gap $[(Na^+ + K^+) - (HCO_3^- + Cl^-)]$	12–16 mmol/l	1.0	12–16 mEq/L
Urea	2.5–7.8 mmol/L	6.02	15.2–47 mg/dL
Creatinine	44–80 μmol/L	0.0113	0.50–0.90 mg/dL
Calcium (adjusted for serum albumin)	2.20–2.60 mmol/L	4.0	8.8–10.4 mg/dL
Phosphate	0.8–1.5 mmol/L	3.1	2.5–4.7 mg/dL
Magnesium	0.7–1.0 mmol/L	2.43	1.7–2.4 mg/dL
Serum osmolality	275–295 mmol/kg	1.0	275–295 mOsm/kg

Table 3

Serum proteins and liver function tests

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Serum proteins			
Total protein	60–80 g/L	0.1	6–8 g/dL
Albumin	35–50 g/L	0.1	3.5–5.0 g/dL
Globulins	20–35 g/L Calculated value (albumin) [Globulins]= [Total Protein]-[Albumin]	0.1	2.0–3.5 g/dL
C-reactive protein (CRP)	<10 mg/L		
Liver function tests			
Bilirubin	<0.21 μmol/L	0.058	<1.2 mg/dL
Alkaline phosphatase adults	*		50–140 U/L
Alanine aminotransferase (ALT)	*		Men 10–40 U/L Women 7–35 U/L
Aspartate aminotransferase (AST)	*		Men 15–40 U/L Women 13–35 U/L
γ-Glutamyl transferase (GGT):	*		Men < 90 U/L Women < 50 U/L

1 U = 1/60 micro katal = 16.67 nano katal.

However, katal is not currently used in clinical practice. Therefore, in accordance with current practice, enzyme activity has been expressed in units per liter (U/L) throughout this book.

*The SI unit of enzyme activity is katal. One katal is the amount of enzyme that converst 1 mole of substrate per second (mol/s). The conversion factor from U/L to katal is:

Table 4

Selected hormones

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Thyroid stimulating hormone (TSH)	0.4–4.0 mU/L		
Free T4	9–25 pmol/L	0.08	0.7–2.0 ng/dL
Free T3	3.5–6.5 pmol/L	65	228–422 pg/dL
Cortisol (plasma):			
at 08:00 h	240–600 nmol/L	0.036	8.6-21.6 µg/dL
at 24:00 h	<50 nmol/L		<1.8 µg/dL
Urinary free cortisol	<250 nmol/24 h		<9 µg/dL
Follicle stimulating hormone (FSH):			
follicular phase	3–13 U/L	1.0	3–13 mIU/mL
midcycle phase	9–18 U/L		9–18 mIU/mL
luteal phase	1–10 U/L		1–10 mIU/mL
postmenopausal	18–150 U/L		18–150 mIU/mL
Luteinizing hormone:			
follicular phase	2–15 U/L	1.0	2–15 mIU/mL
midcycle phase	22–90 U/L		22–90 mIU/mL
luteal phase	1–19 U/L		1–19 mIU/mL
postmenopausal	16–64 U/L		16–64 mIU/mL
Progesterone (midluteal phase)	<20 nmol/L	0.33	6.4 ng/mL
Parathyroid hormone	1.1–6.9 pmol/L	9.16	11–69 pg/mL

Table 5

Cholesterol and other lipids

Analyte/test interpretation	SI units	Conversion factor (SI to conventional units)	Conventional units
Total cholesterol			
Desirable	<5.18 mmol/L	38.6	<200 mg/dL
Borderline-high	5.18–6.18		200–239 mg/dL
High	>6.21		>240 mg/dL
LDL-cholesterol		38.6	
Optimal	<2.6 mmol/L		<100 mg/dL
Near optimal	2.6–3.3		100–129 mg/dL
Borderline high	3.4– 4.1		130–159 mg/dL
High	4.2–4.9		160–189 mg/dL
Very high	>5.0		≥190 mg/dL
HDL-cholesterol			
high	<1.0 mmol/L	38.6	<40 mg/dL
low	>1.6 mmol/L		>60 mg/dL
Triglycerides (desirable)	<1.7 mmol/L	88.4	<150 mg/dL

The interpretation of lipid levels is based on clinical decision limits developed by ATP3, and not on reference ranges

ATP3: National Cholesterol Education Program (USA) Adult Treatment Panel 3

For more details on lipid values and goals of treatment see www.nhlbi.nih.gov/guidelines/cholesterol/atp3xsum.pdf (US recommendations) and www.bcs.com/download/651/JBS2final.pdf (British recommendations)

Note also that updated (ATP4) criteria might have been released while this book was in print.

Table 6

Analyte/test interpretation	SI units	Conversion factor SI to conventional units	Conventional units
Glucose: normal concentrations and states of glucose intolerance			
Normal range (plasma)	4–6 mmol/L	18	72–109 mg/dL
Fasting glucose	<6.1		<110
Impaired fasting glucose	≥6.1 but <7.0 mmol/L		≥110 but <126 mg/dL
Impaired glucose tolerance (IGT)			
Fasting	<6.1 mmol/L		<110 mg/dL
2h Post glucose load	≥7.8 < 11.1		≥140 but <200
Prediabetes (ADA category)			
Fasting	≥5.6 <7.0 mmol/L		≥101 but <126 mg/dL
OGTT 2h post load	≥7.8 < 11.11		≥140 but <200
Glucose: Diabetes mellitus			
Fasting	≥7.0 mmol/L		≥126 mg/dL
OGTT 2h post glucose load	>11.1		>200
Overt diabetes in pregnancy			
Fasting	≥7.0 mmol/L		≥126 mg/dL
OGTT 2h post glucose load	≥11.1		≥200
Gestational diabetes			
Fasting	≥5.1 mmol/L		≥92 mg/dL
OGTT 1h post glucose load	≥10.1		≥182
OGTT 2h post glucose load	≥8.5		≥153
Glycated hemoglobin			
Normal range	20–38 mmol/mol	*	4.0–5.6%
Prediabetes (ADA category)	≥39 but <43		≥5.7 but <6.1
Diabetes mellitus	≥48		≥6.5
Overt diabetes in pregnancy	≥48		≥6.5

HbA1c (mmol/mol)= [HbA1c(%) −2.15] x 10.929

http://www.diabetes.org.uk/Professionals/Publications-reports-and-resources/Changes-to-HbA1c-values/

Criteria for the diagnosis of diabetes mellitus and glucose intolerance have been developed by several national and international bodies. Here we quote the widely accepted ones, developed by the American Diabetes Association (ADA), the World Health Organization (WHO), and the International Association of Diabetes in Pregnancy Study Groups (IADPSG).

Note the following:

1. The ADA and the WHO criteria for the diagnosis of diabetes in nonpregnant persons are the same. They differ slightly, however, in case of definitions of glucose intolerance. The ADA uses the term ‘prediabetes’, whereas the WHO distinguishes IFG and IGT conditions.

2. Criteria for the diagnosis of overt diabetes in pregnancy are set lower than in nonpregnant persons.
3. Interprtetation of OGTT in pregnancy involves fasting glucose concentration and both 1h and 2h post glucose load values, whereas in nonpregnant persons interpretation is based on the fasting and 2h post load values only.

For further details see Nolan et al in Further Reading.

*Formula to convert the conventional (DCCT) HbA1c values to IFCC (SI) units:

Table 7

Miscellaneous tests

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Amylase	*		0–100 U/L
Urate:			
men	0.2–0.5 mmol/L	16.8	5.0–8.0 mg/dL
women	0.1–0.4 mmol/L		2.5–6.2 mg/dL
Lactate	0.7–1.8 mmol/L	9.0	6–16 mg/dL

Table 8

Urine analysis

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Microalbumin in urine	<20 mg/L men		
Albumin/creatinine ratio: (ACR) Men	*		<2.5 mg/mmol creatinine
Women	*		<3.5 mg/mmol creatinine
Microalbumin excretion rate (AER)	*		<20 µg/min
Osmolality	50–1200 mmol/kg	1.0	50–1200 mOsm/kg

Table 9

Hematology tests

Analyte/test	SI units	Conventional units
Hemoglobin:		

men	130-180 g/L	13.0–18.0 g/dL
women	120-160 g/L	12.0–16.0 g/dL
Hematocrit	41–46%	41–46 mL/dL
Erythrocytesw cell count:		
men	4.4–5.9 × 10^{12}/L	4.4–5.9 × 10^{6}/mm^3
women	3.8–5.2 × 10^{12}/L	3.8–5.2 × 10^{6}/mm^3
Mean corpuscular volume (MCV)	80–96 fL	80–96 μm^3
Leukocytes, total	4.0–11.0 × 10^{9}/L	4000–11 000/mm^3
Leukocytes, differential count:		
neutrophils	2–7.5 × 10^{9}/L	45–74%
lymphocytes	1.3–4 × 10^{9}/L	16–45%
monocytes	0.2–0.8 × 10^{9}/L	4.0–10%
eosinophils	0.04–0.4 × 10^{9}/L	0.0–7.0%
basophils	0.01–0.1 × 10^{9}/L	0.0–2.0%
Platelets	150–400 × 10^{9}/L	150 000–400 000/mm^3
Reticulocytes	25–75 × 10^{9}/L	0.5–1.5% of erythrocytes
Erythrocyte sedimentation rate (ESR)	2–10 mm/h	
Activated partial thromboplastin time (APTT)	30–40 sec	
Prothrombin time (PT)	10–15 sec	
Thrombin clotting time (TCT)	10–15 sec	
Skin bleeding time	2.0–9.0 min	
d-Dimer	<0.25 g/L	

Table 10

Tests related to iron metabolism and investigation of anemia

Analyte	SI units	Conversion factor (SI to conventional units)	Conventional units
Ferritin (serum)	31–450 pmol/L	0.445	14–200 ng/mL
Vitamin B_{12} (serum)	138–780 pmol/L	1.36	187–1060 pg/mL
Folate (serum)	12–33 nmol/L	0.442	5.3–14.6 ng/mL

Further reading

Bakerman, S. *Bakerman's ABC of interpretive laboratory data*, ed 4. AZ: Scottsdale; 2002.

Burtis, CA, Ashwood, ER, Bruns, DE. *Tietz textbook of clinical chemistry and molecular diagnostics*, ed 5. Philadelphia: Saunders; 2011.

Dominiczak MH, ed. Seminars in clinical biochemistry. Glasgow: University of Glasgow, 1997.

Stone, NJ, Robinson, J, Lichtenstein, AH, et al, ACC/AHA Prevention Guideline: 2013 ACC/AHA Guideline on the Treatment of Blood Cholesterol to Reduce Atherosclerotic Cardiovascular Risk in Adults: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *Circulation* 2013; 0. http://circ.ahajournals.org/content/early/2013/11/11/01.cir.0000437738.63853.7a.citation

This most recent guideline (released in November 2013) provides updated recommendations on the use of statins to lower the risk of atherosclerotic

cardiovascular events. It recommends that persons with clinical atherosclerotic disease should be treated irrespective of the plasma LDL-cholesterol concentration. Otherwise, initiation of treatment is based on the pre-treatment LDL-cholesterol: therapy should be initiated when LDL-cholesterol equals or exceeds 4.4 mmol/L (190 mg/dl). Persons with diabetes mellitus and with an overall 10-year risk of event equal or above 7.5% should be treated at LDL levels between 1.8 mmol/L (70 mg/dl) and 4.3 mmol/L (169 mg/dL).

The major change compared to the previous guidelines is that this one no longer recommends treatment to a specific LDL-cholesterol 'target level'. Instead, it suggests 2 statin regimens, instituted depending on the risk of event and the risk of side effects: a high-intensity therapy regimen that lowers LDL-cholesterol by 50% or more, and a moderate intensity treatment, lowering the LDL-cholesterol by 30–50%. See the reference above for details.

Websites and downloads

American Diabetes Association. Standards of Medical Care in Diabetes – 2012. http://care.diabetesjournals.org/content/35/Supplement_1/S11.full.

Diabetes: update on ADA guidelines 2012. www.aca.org/conferences/winter2012/WorkshopPresentations/A-1D%20Diabetes%20An%20Update%20on%20American%20Diabetes%20Association%20Guidelines.pd

Lab Tests Online. www.labtestsonline.org.uk/.

Lab Tests Online. www.labtestsonline.org.uk/understanding/features/ref-ranges/.

National Cholesterol Education Program ATP III Guidelines At-A-Glance Quick Desk Reference. www.nhlbi.nih.gov/guidelines/cholesterol/atglance.pdf.

Nolan, CJ, Damm, P, Prentki, M. Type 2 diabetes across generations: from pathophysioplogy to prevention and treatment. *Lancet*. 2011; 378:169–181.

SI values conversion calculator. www.amamanualofstyle.com/oso/public/jama/si_conversion_table.html.

APPENDIX 2:

The Fundamentals of Recombinant DNA Technology: Molecular hybridization and DNA cloning

W. Stephen Kistler

Introduction

Our current ability to analyze and manipulate genomes began with reports in the 1970s of methods to cleave DNA at specific sites, to insert new DNA fragments into bacterial plasmids, and to sequence regions of DNA more than just a few nucleotides long. Prior to these developments genes in humans were known almost exclusively by their effects, that is from phenotypes and disease – genes were concepts, rather than structures. Gradually, it became possible to see exactly what a gene was and to to determine if genes were normal or mutated. A major step in the process was recognition that single strands of nucleic acids will form double stranded (ds) pairs with each other only if the sequences are highly similar or identical. Just as antibodies can detect single proteins in the midst of thousands of others, nucleic acids sequences will bind only to their complement in the presence of millions of non-matching sequences. A second major advance was the discovery of restriction enzymes, which convert chromosomal DNA into discrete fragments of useful length. Following separation by size, these smaller fragments could be detected by nucleic acid probes, and also sequenced. While many of the procedures used for analysis of DNA in the 20th century are mostly of historical interest today, some knowledge of them is necessary both for reading the older literature and for understanding modern recombinant DNA technology. The ability to splice and recombine fragments of DNA into viruses, bacterial plasmids, and even chromosomes has revolutionized the production of many clinically important human proteins as well as vaccines.

This Appendix provides an overview of some of the general techniques for forming so-called recombinant DNA and for cloning of this DNA.

The principles of molecular hybridization

Hybridization is based on the annealing properties of DNA

Hybridization is a process by which a piece of DNA or RNA of known nucleotide sequence, which can range in size from as little as 15 base pairs (bp) to several hundred kilobases, is used to identify a region or fragment of DNA containing complementary sequences. The first piece of DNA or RNA is called a probe. Probe DNA will form a complementary base pair with another strand of DNA, often termed the target, if the two strands are complementary and a sufficient number of hydrogen bonds are formed.

In molecular hybridization, it is essential that the probe and target are initially single-stranded

Probes can vary in both their size and their nature (DNA, RNA or oligonucleotide). However, one essential feature of any hybridization reaction is that both the probe and the target must be free to base pair with one another. For DNA hybridization, the two strands of DNA must first be separated by thermal or chemical treatment, a process called DNA **denaturation** or **melting**. Once both probe and target DNA are single-stranded, mixing of the two under conditions that favor the formation of a double-stranded helix will allow complementary bases to recombine. This process is called DNA **annealing** or **reassociation**, and when a probe strand reacts with a target strand, the complex is termed a **heteroduplex**.

Formation of probe–target heteroduplexes is the key to the usefulness of molecular hybridization

The conditions under which DNA hybridization occurs and the reliability and specificity, or stringency, of hybridization are affected by several factors:

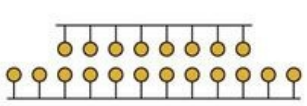
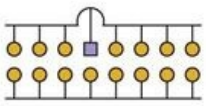
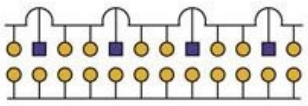
- **base composition:** GC pairs have three hydrogen bonds compared with the two in an AT pair. Double-stranded DNA with a high GC content is therefore more stable and has a higher meeting temperature (Tm).
- **strand length:** the longer a strand of DNA, the greater the number of hydrogen bonds between the two strands. Longer strands require higher temperatures or stronger alkali treatment to denature them; stability varies

dramatically with length for very short probes but above a few hundred base pairs, stability is relatively insensitive to length and is determined primarily by base composition.

■ **reaction conditions:** high cation concentration (typically Na^+) favors double-stranded DNA (because the negative charges on the sugar-phosphate backbone are shielded from each other), while high concentrations of urea or formamide favor single-stranded DNA (because these reagents reduce base-stacking and can compete for hydrogen bond formation). Hybridizations are said to be carried out at **low stringency** when conditions strongly favor duplex formation, permitting some mismatch in the DNA duplex, and at **high stringency** when only matched, complementary duplexes are formed.

Thus by appropriate selection of conditions (high stringency), a small 30–50 bp probe can require a perfect match to form a stable hybrid with its target. On the other hand, under low stringency a longer probe, *e.g.* 500 bp, might react with targets that contain multiple nucleotide mismatches or mutations (Fig. A2.1).

A Hybridization characteristics using a large conventional probe (>200 bases)

Match	Perfect	Single base mismatch	Multiple mismatch
Stringency	High	Intermediate	Low
Example	Human target + human probe 	Human target + human probe with mutation 	Human target + mouse probe 
Stability	Stable	Stable	Stable

B Hybridization characteristics using a small oligonucleotide probe

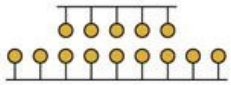
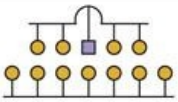
Match	Perfect	Single base mismatch	
Stringency	High	High	
Example			
Stability	Stable	Unstable	

FIG. A2.1 Probe–template hybridization.
(A) Large probes, *e.g.* 200 bases or more, can form stable heteroduplexes with the target DNA even if there are a significant number of non-complementary bases in conditions of low stringency. **(B)** Oligonucleotide probes, in contrast, may discriminate between targets that differ by a single base under stringent conditions.

The stability of a nucleic acid duplex can be assessed by determining its melting temperature (T_m)

The melting temperature (T_m) is the temperature, in vitro, at which 50% of a double-stranded duplex has dissociated into single-strand form. For relatively long DNA probes, T_m is determined primarily by base composition, with AT-rich DNA melting at a lower temperature than GC-rich DNA. For humans and other mammals, the average GC content is about 40% and the melting temperature in moderate salt is about 87°C. For short oligonucleotides, such as the primers used in polymerase chain reactions (PCR), effects of length, composition and even the various dinucleotide sequences must be taken into consideration. This is because double-stranded DNA is stabilized by the degree

of overlap by the stacked bases in successive nucleotides, and this varies depending on the specific nucleotide neighbors of a particular base. Computer programs are widely available to predict T_m values.

Probes must have a label to be identified

Implicit in the use of probes to identify pieces of complementary DNA is the notion that if hybridization occurs, the heteroduplex can be specifically detected. Thus, the probe is labeled so that the probe–target duplex can be identified. The labels generally fall into two categories, either isotopic, *i.e.* involving radioactive atoms, or nonisotopic, *e.g.* end-labeling probes with fluorescent tags or small ligand molecules. Many techniques involving probe hybridization and labeling still use radioisotopes such as ^{32}P, ^{35}S or ^{3}H and, as such, require a method for detecting and localizing the radioactivity. The most common method involves the process of **autoradiography**. Autoradiography allows information from a solid phase, *e.g.* a gel or fixed-tissue sample, to be detected and saved in two-dimensional form as an exposed photographic image.

Southern blots are the prototype for methods that use specific hybridization probes to identify sequences in DNA or RNA

One of the fundamental steps in the evolution of molecular biology was the discovery that DNA could be transferred from a semisolid gel onto a nitrocellulose membrane in such a way that the membrane could act as a record of the DNA information in the gel and could be used for multiple-probe experiments. The process whereby the DNA is transferred to the membrane was first described by Edward Southern, but subsequent techniques based on the transfer of RNA and proteins have also adopted the direction theme and are called Northern (RNA target) and Western (protein target) blots, respectively.

Restriction Enzymes

Use of restriction enzymes to analyze genomic DNA

Restriction enzymes cleave DNA at specific nucleotide sequences

Restriction endonucleases cleave double-stranded DNA. These enzymes are sequence specific and each enzyme acts at a limited number of sites in DNA called ‘recognition’ or ‘cutting’ sites. Restriction endonucleases are part of the bacterial ‘immune system’. Bacteria methylate their own DNA, protecting it from their own restriction enzymes, but cleave unmethylated infecting viral or bacteriophage DNAs at specific sites, thereby inactivating the virus and restricting viral infection.

If DNA is digested by a restriction enzyme, the DNA will be reduced to fragments of varying sizes depending on how many cutting sites for that restriction enzyme are present in the DNA. The cutting sites are frequently **palindromic sequences**, sites at which the base sequence reads the same, backward and forward. It is important to note that each enzyme will cut DNA into a unique set of fragments (Fig. A2.2). Many restriction enzymes recognize sites that are typically four (e.g. HaeIII), six (e.g. EcoR I) or eight nucleotides (e.g. NotI) in length (Table A2.1). Variation in just one nucleotide within the recognition sequence makes a sequence completely resistant to a particular enzyme.

Table A2.1

Restriction endonucleases in common use

Endonuclease	Restriction site	Ends
HaeIII	GG*CC	Flush
	CC*GG	
MspI	C*CGG	Sticky
	GGC*C	
EcoRV	GAT*ATC	Flush
	CTA*TAG	
EcoRI	G*AATTC	Sticky
	CTTAA*G	
NotI	GC*GGCCGC	Sticky
	CG CCGG*CG	

Enzymes can cleave DNA to produce ‘flush ends’ where the DNA is cut ‘vertically’ leaving two ends that do not have any overhanging nucleotides. However, if the DNA is cleaved ‘obliquely’, the DNA will have short single-stranded overhangs. Such ends are called ‘sticky’ because they will selectively rejoin (hybridize) to matching overhangs. The sites of cleavage of DNA by restriction enzymes are often described as palindromic because of their inverted repeat symmetry – they have identical sequences in opposite directions on the complementary strands.

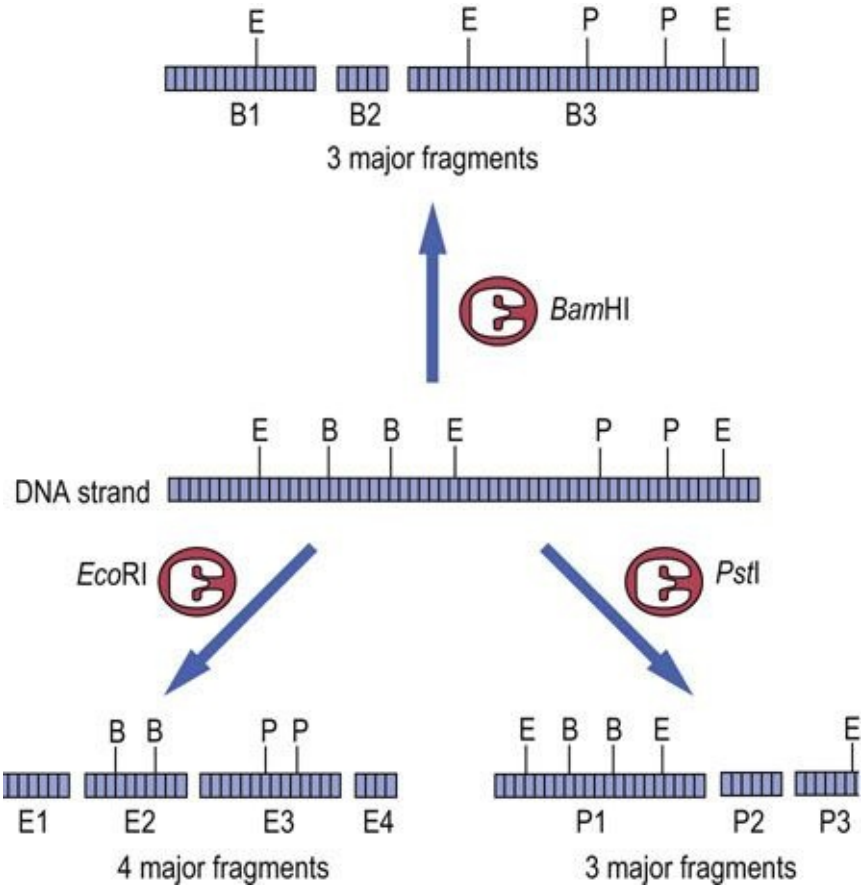


FIG. A2.2 Restriction enzyme digestion of DNA.
Digestion of a DNA molecule by several different restriction enzymes may result in many different fragments, even though the apparent size of the fragments is similar. For example, fragments E1 and P3 are of similar size but are clearly different pieces of DNA. E, *Eco*RI site; B, *Bam*HI site; P, *Pst*I site.

The frequency of the cutting sites for various enzymes varies with the length of the recognition site. Cut sites for an enzyme with a 4-base recognition site, such as HaeIII, would occur by chance once per 256 base pair sequence. Cut sites for an enzyme with an 8-base recognition sites, such as NotI, wold occur only once in about 656,000 base pairs. Thus, frequent cutters typically generate

many small fragments, while rare cutters generate fewer and larger fragments. These differences can be exploited in the analysis of gene structure and chromosomal location.

DNA fragments, blotted onto a solid gel phase, are used as a template for exposure to a range of molecular probes

If DNA is digested by a restriction enzyme, the resulting digest can be separated on the basis of size by gel electrophoresis. Agarose gel electrophoresis is commonly used to separate fragments ranging in size from 100 bases to approximately 20 kb in length (above 40 kb, resolution is minimal). Following electrophoresis, the gels are soaked in a strong alkali solution to denature the DNA. These single-stranded fragments can then be transferred to a nitrocellulose or nylon membrane to which they bind readily and, if preserved properly, permanently. The process of transfer involves the passage of solute through the gel, passively carrying the DNA and producing an image of the gel on the membrane (Fig. A2.3). The membrane may then be probed with an oligonucleotide or DNA fragment (Southern blot), *e.g.* for genotyping, paternity testing or identification of cells incorporating a gene during a cloning experiment (below).

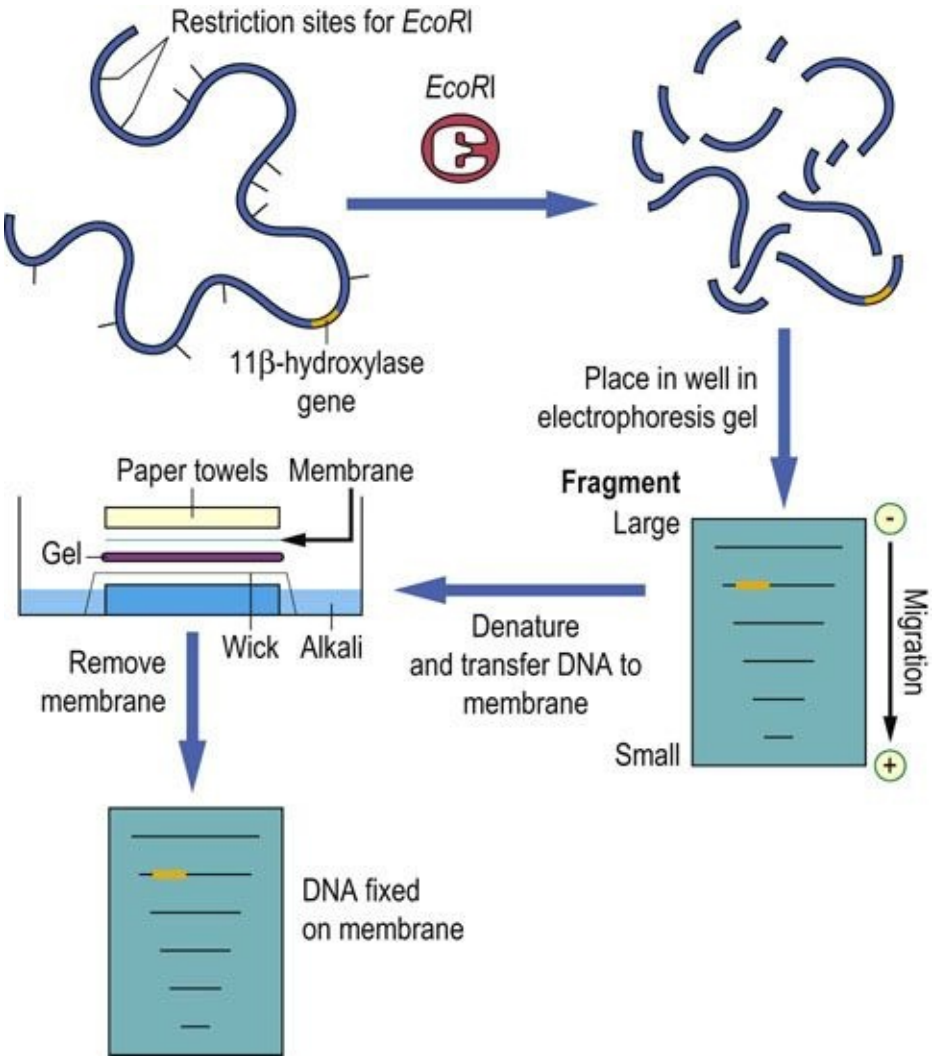


FIG. A2.3 Southern blotting of DNA.
DNA digested with a restriction enzyme is size-fractionated by agarose electrophoresis. The agarose gel is then placed in alkali to denature the DNA. The now single-stranded DNA can pass from the gel to the membrane (typically nylon or nitrocellulose) as buffer solution flows upward by capillary action, forming a permanent record of the digested DNA.

Restriction fragment length polymorphisms (RFLPs)

Analysis of restriction fragment length may be used to detect a mutation or polymorphism in a gene

If a change in DNA sequence creates or destroys a recognition site that yields a fragment detected by a probe, then the altered length of that fragment can be detected by Southern blotting. If a cleavage site is created, the fragment becomes smaller; if the cleavage site is eliminated, the fragment becomes larger. The different patterns generated as a result of a mutation or gene variant are known as restriction fragment length polymorphisms (RFLPs) (Fig. A2.4). Such RFLPs can be used either to identify disease-causing mutations, because of a single point mutation creating or abolishing a restriction site, or to study variation in noncoding DNA for use in the study of genetic linkage.

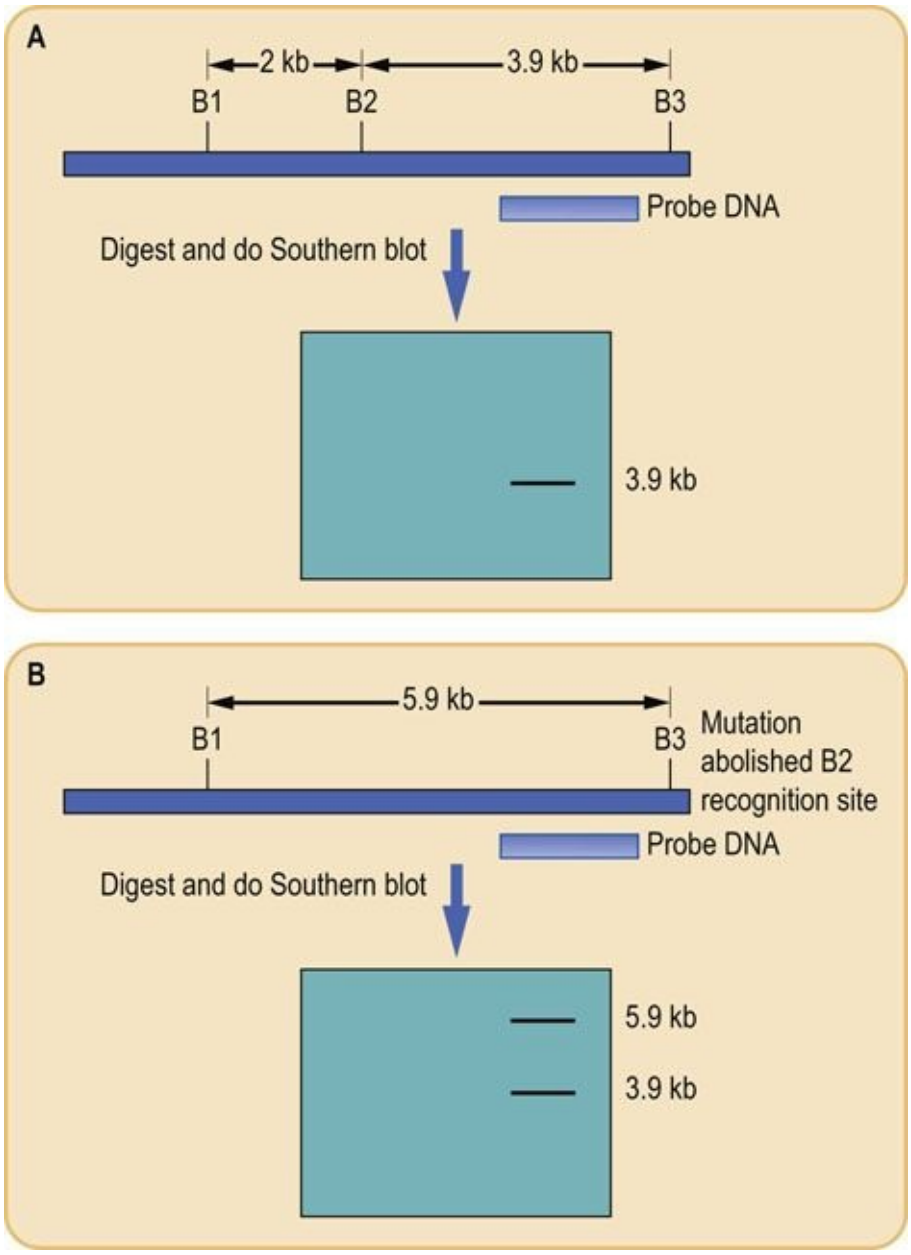


FIG. A2.4 Restriction fragment length polymorphisms (RFLP).
Variations in the nucleotide sequence of DNA, either due to natural variation in individuals or as a result of a DNA mutation, can abolish the recognition sites for restriction enzymes. This means that when DNA is digested with the enzyme whose site is abolished, the size of the resulting fragments is altered. Southern blotting and probe hybridization can be used to detect this change. Results are shown for a representative gene from **(A)** homozygous normal and **(B)** heterozygous mutant individuals. B, *Bam*HI restriction site.

RFLP analysis can also detect larger pathologic changes in the DNA sequence, either deletions or duplications. Large deletions of a gene may abolish

restriction sites; this leads to the disappearance of a fragment on a Southern blot in homozygous individuals. Alternatively, if a DNA duplication event occurs, a new gene may be formed, which has a different pattern of restriction sites that allow detection of the new gene. This type of hybridization is performed using large probes (0.5–5.0 kb) and is performed under moderate stringency, *i.e.* it is sufficiently rigid to allow hybridization of probe and target but also to tolerate minor differences, *e.g.* in noncoding DNA.

Low stringency hybridization of a probe to a Southern blot of digested DNA may allow genes related to, but not identical to, the starting gene to be identified. Many genes exist in families, or have nonfunctional, nearly identical copies elsewhere in the genome (pseudogenes), and thus hybridization of a probe may identify one or more restriction fragments, corresponding to related genes. Similarly, related genes in different species may be identified by using a single probe that can hybridize at low stringency to complementary sequences in blots of DNA from mouse, rat or other species.

Clinical box Use of RFLPs for detection of the sickle cell gene

A 24-year-old Afro-Caribbean woman was referred for prenatal counseling. Her younger brother had sickle-cell anemia, and she had become pregnant. Her partner was known to be a carrier of the sickle-cell mutation (sickle-cell trait) and she wanted to know if her child would develop sickle-cell anemia.

Since the patient is at risk of being a carrier, she opted to have chorionic villus sampling (CVS) performed to detect the presence or absence of the sickle-cell mutation in her child. Analysis of her own DNA revealed that she was a carrier, and the CVS showed that the child was also a carrier and would not develop sickle-cell anemia.

Comment.

Occasionally, a mutation will directly abolish or create a restriction site and thus allow the use of a restriction-based method to demonstrate the presence or absence of the mutant allele. One widely examined mutation is the A > T substitution at codon 6 in the sequence for the β-globin gene responsible for sickle-cell disease (see Chapter 5). This results in a glutamine-valine (Glu-

Val) mutation in the amino acid sequence β-globin and also abolishes a recognition site for MstII (CCTN(A > T)GG) in the β-globin gene. Digestion of normal human DNA with MstII and probing the Southern blot with a probe specific for the promoter of the β-globin gene yields a single band of 1.2 kb, as the nearest MstII site is 1.2 kb upstream in the 5'region of the gene. The abolition of the codon 6 restriction site means that the fragment size seen when probing MstII digested DNA is now 1.4 kb, as the next MstII site is located 200 bases downstream in the intron after exon 1. Thus, patients with sickle-cell anemia will show only one band, 1.4 kb, while carriers will have 2 bands, one 1.4 kb and another, 1.2 kb, and unaffected individuals will have a single 1.2 kb band.

Cloning of DNA

Cell-based cloning

Bacterial plasmids are bioengineered to optimize their use as vectors

Cell-based cloning relies on the ability of replicating cells, *e.g.* bacteria, to permit replication of so-called recombinant DNA within them. Recombinant DNA refers to any DNA molecule that is artificially constructed from two pieces of DNA not normally found together. One piece of DNA will be the target DNA that is to be amplified and the other will be the replicating **vector** or replicon, a molecule capable of initiating DNA replication in a suitable host cell.

Today, the majority of cell-based cloning is performed using bacterial cells. In addition to the bacterial chromosome, bacteria may contain extrachromosomal double-stranded DNA that can undergo replication. One such example is the bacterial plasmid. **Plasmids** are circular, double-stranded DNA molecules that undergo intracellular replication and are passed vertically from the parent cell to each daughter cell. However, unlike the bacterial chromosome, plasmids used in these techniques are copied many times during each cell division. Thus, plasmids are ideal carriers for the amplification of target DNA, and thus the encoded protein. Methods involving the use of plasmids are widespread throughout molecular biology.

Target DNA is introduced into a plasmid by using restriction enzymes to cut target and plasmid DNA so that the target DNA and the linearized vector DNA will have complementary sticky ends (Fig. A2.5). DNA ligase then covalently joins the target to the ends of the vector to form a closed circular recombinant plasmid. Once the target DNA is incorporated into the plasmid vector, the next step is to introduce the plasmid into a host cell to allow replication to occur. The cell membrane of bacteria is selectively permeable and prevents the free passage of large molecules such as DNA in and out of the cell. However, the permeability of cells can be altered temporarily by factors such as electric currents (electroporation) or high-solute concentration (osmotic stress), so that the membrane becomes temporarily permeable and DNA can enter the cell. Such a process renders the cells competent, *i.e.* they can take up foreign DNA from the extracellular fluid, a process known as **transformation**. This process is

generally inefficient, so that only a small fraction of cells may take up plasmid DNA, and often only a single plasmid per bacterium is introduced during transformation. However, it is this process of cellular uptake of plasmid DNA that forms a critical step in cell-based cloning. Individual recombinant DNAs are easily resolved from one another because they are taken up by separate cells that can be isolated simply by spreading them on an agar surface.

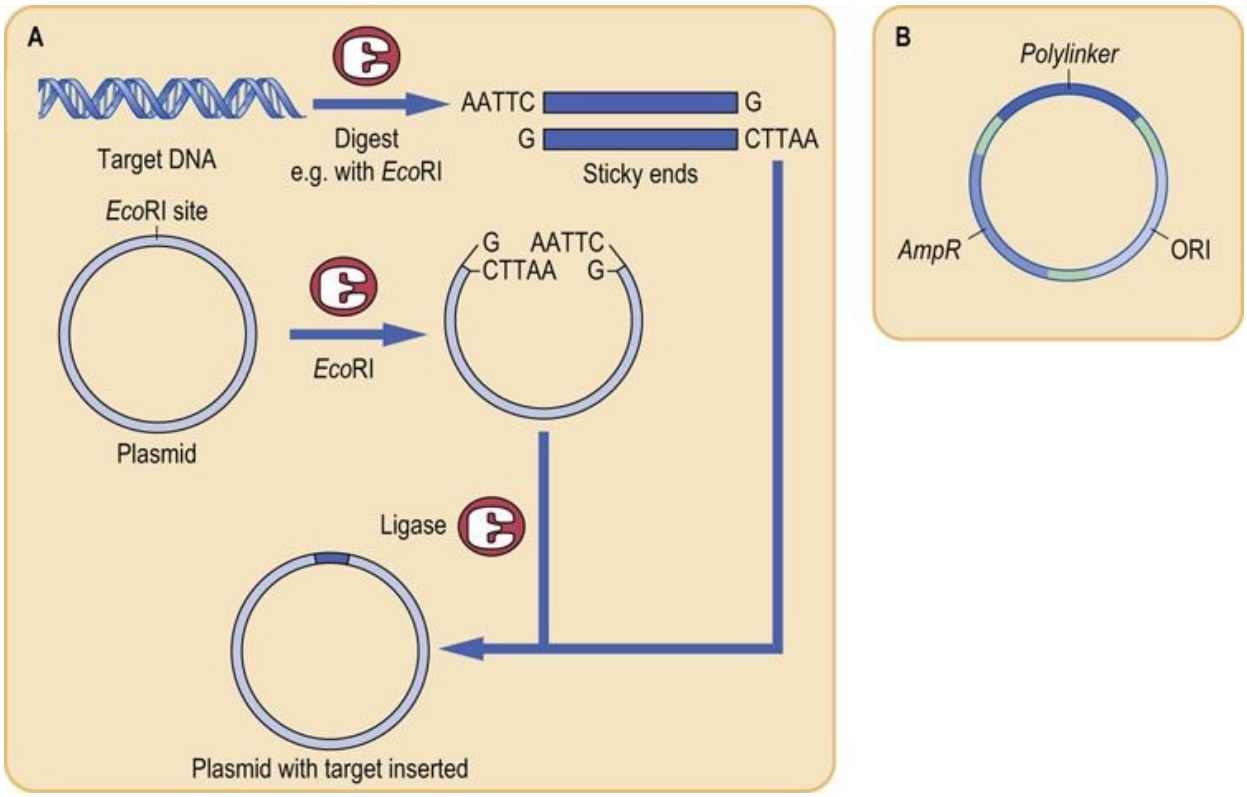


FIG. A2.5 Formation of a plasmid containing a target gene for cloning.
(A) DNA containing the target gene is digested with a restriction enzyme that will produce 'sticky ends', *e.g. Eco*RI. The plasmid also has a restriction site for EcoRI, so that when digested with *Eco*RI it becomes a linear DNA strand with 'sticky ends' complementary to the target. Upon ligation, the target and vector form a recombinant molecule. **(B)** Structure of a typical plasmid. The plasmid contains a gene conferring resistance to the antibiotic ampicillin (AmpR), and a polylinker region containing approximately 10 restriction enzyme recognition sites, which serve as sites for insertion of target DNA. The plasmid also contains ORI, the site for origin of DNA replication.

Following transformation, the cells are allowed to replicate, usually on a standard agar plate containing a suitable antibiotic (see Fig. A2.5B) to kill cells that do not harbor the plasmid containing the antibiotic resistance gene. This

selection or screening process, based on antibiotic resistance, is an important step because of the low efficiency of the uptake of plasmid DNA into bacteria. Colonies (clones of single cells) are then ‘picked’ and transferred to tubes for growth in liquid culture and a second phase of exponential increase in cell number. This work is done automatically in microplate systems, so that, from a single cell and a single molecule of DNA, an extremely large number of cells containing multiple, identical recombinant plasmids can be generated in a relatively short time (Fig. A2.6). Recovery of the plasmid DNA is easy because it is a small, covalently intact circle, readily separated from the bacterial chromosomal DNA by a variety of techniques.

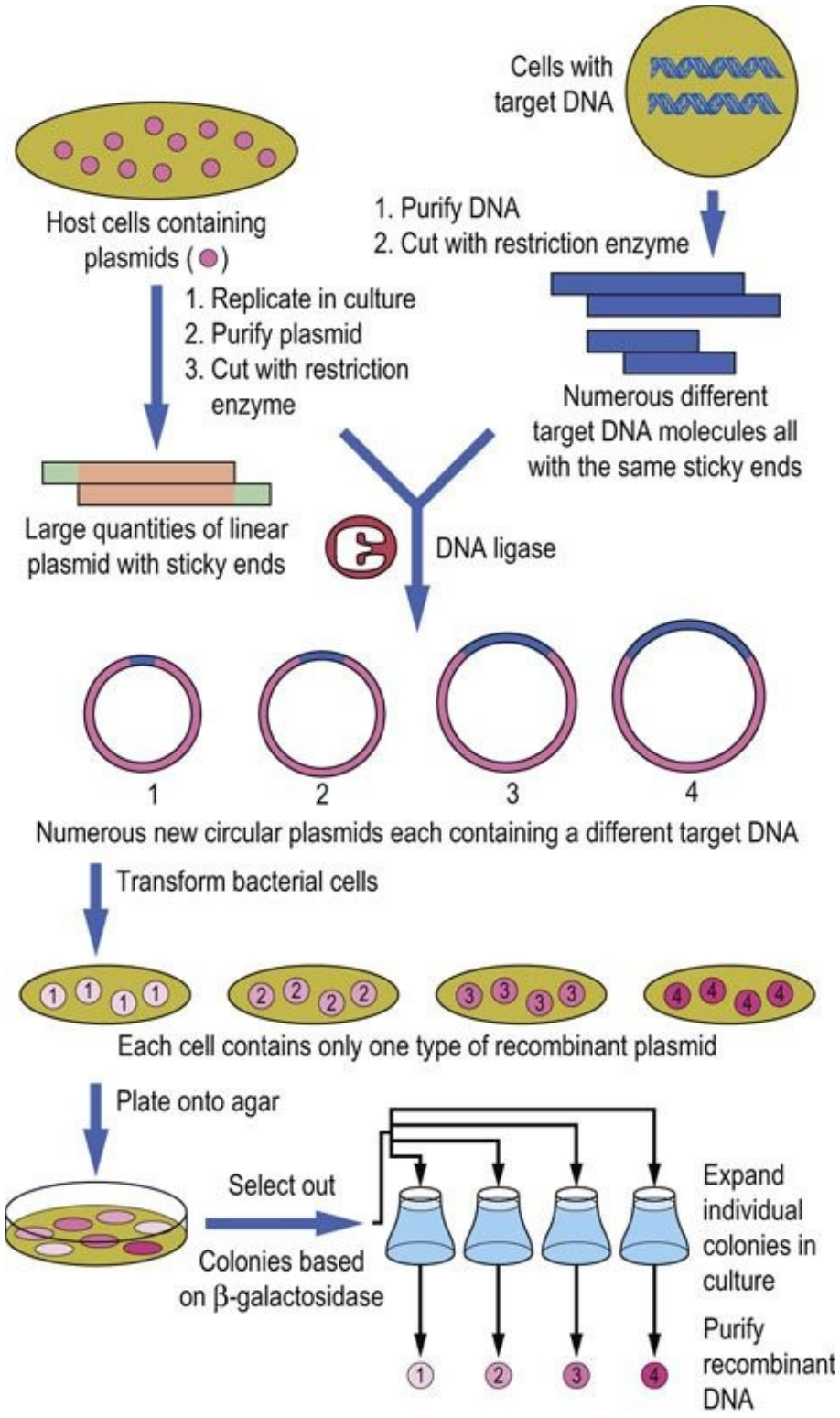


FIG. A2.6 Cell-based DNA cloning.
An example of cloning genomic DNA using bacterial cells. In general, each transformed bacterium will take up only a single plasmid molecule. Therefore individual bacterial colonies will contain many identical copies of just one particular recombinant DNA.

The bacterial cell is then grown in culture and the target protein is recovered following lysis of the cells. Alternatively, the plasmid might encode a protein with a signal sequence so that it is secreted into the medium; the signal sequence would be removed afterwards.

The technology for producing protein pharmaceuticals is a complex, multi-step process, often protected by trade secrets. For the bacterial synthesis of recombinant insulin (see Box on p. 615), for example, there is no gene or mRNA for the sequence of insulin – insulin is synthesized as preproinsulin, which is processed in pancreatic β-cells to yield the secreted hormone (see Chap. 21). The synthesis of human insulin in a bacterial system might involve the incorporation of the proinsulin gene into a plasmid, transformation into a bacterial host, synthesis of proinsulin, spontaneous disulfide crosslinking, processing by endopeptidases to remove the C-peptide, then folding to produce the active insulin molecule. The proteolytic processing might happen in the cell or following isolation of the proinsulin; intracellular processing would require encoding the protease in the bacterial plasmid. Other strategies might be imagined, such as synthesis of the A and B chains in separate bacterial hosts, then extracellular association into the active hormone. It might also be possible to design a protein product that would be secreted from the bacterial cell, then processed ex vivo to remove the secretory sequence and C-peptide.

Future Directions

Cloning of DNA is a rapidly evolving field in biomedical research and modern medicine. It is the basic methodology for production of genetically modified organisms (GMO), including agricultural products and transgenic and knock-out animals. More sophisticated eukaryotic expression systems, including human tumor cells, hen's eggs and plant cells, are now commonly used for production of protein pharmaceuticals. In some cases, these cells are engineered to contain specific processing enzymes, *e.g.* glycosyltransferases, for posttranslational modification of the protein. β-Glucosidase used for enzyme replacement therapy in Gaucher's disease (see Chap. 28) is produced in bioengineered Chinese hamster ovary (CHO) cells. The enzyme secreted from these cells contains a mannose-6-phosphate signal, so that it is taken up into lysosomes following intravenous injection. *Humanized* proteins may also be synthesized in murine cells, then processed by glycosidases and/or glycosyl transferases to yield a

protein with the proper posttranslational modifications for use in human plasma or cells. In the not-to-distant future, it may be possible to by-pass all of these steps by gene therapy, *i.e.* by incorporating the gene of interest directly into the relevant human cells using viral vectors. Despite all these advances, the fundamental idea of a foreign piece of DNA being inserted into a particular restriction site of a plasmid or viral vector remains completely relevant today.

Clinical box Production of recombinant proteins: insulin

A 13-year-old girl was admitted with dehydration, vomiting and weight loss. Her blood glucose level was 19.1 mmol/L(344 mg/dL) and she had ketonuria. A diagnosis of type 1 diabetes mellitus was made. She was started on recombinant human insulin, was rehydrated (Chapter 21), and made a prompt recovery.

Comment.

Prior to the advent of recombinant DNA technology, insulin therapy involved the use of animal insulins, most commonly pork or beef, which were chemically similar, but not identical, to human insulin. As a result of these differences, animal insulins often led to the development of antibodies, which reduced the efficacy of the insulin and could lead to treatment failures.

Insulin was the first clinically important human molecule to be produced by means of recombinant DNA technology. Following the cloning of the human insulin gene, large-scale production of pure human insulin was possible by inserting the cloned gene into a cell-based amplification system. Large amounts of insulin gene copies were produced, which were then expressed in either bacteria or yeast, and the resulting purified insulin was made available for use in treatment of diabetic patients. By this means, human recombinant insulin has largely replaced animal insulin in the treatment of diabetes. Other important recombinant human peptides used clinically include growth hormone, erythropoietin and parathyroid hormone.

Advanced concept box Vector systems for cloning large DNA fragments

One critical factor in recombinant DNA manipulations is the size of the target DNA. Conventional bacterial plasmids, although convenient to work with, are limited in the size of insert they can accept; 1–2 kb is the common size of the insert, with an upper limit of 5-10 kb. Some modified plasmid vectors called **cosmids** can accept larger fragments up to 20 kb. Another commonly used vector that has the ability to accept larger DNA fragments is the **bacteriophage lambda** (λ). This viral particle contains a double-stranded DNA genome packaged within a protein coat. The λ-phage can infect E. coli cells with high efficiency and introduce its DNA into the bacterium. Infection leads to the replication of viral DNA and the synthesis of new viral particles, which can then lyse the host cell and infect neighboring cells to repeat the process. The viral DNA is then re-isolated to obtain the recombinant DNA.

Larger inserts can be cloned by using modified chromosomes from either bacteria **(bacterial artificial chromosomes, BACs)**, or yeast **(yeast artificial chromosomes, YACs)**. Such vectors can accommodate DNA fragments up to 1–2 Mb. BACs have been particularly important in putting together the sequence of the human genome.

Index

Page numbers ending in ‘b’, ‘f’ and ‘t’ refer to Boxes, Figures and Tables respectively

B

C

D

F

G

H

J

K

L

M

N

O

P

R

S

T

U

V

W

X

Y

Z

Instructions for online access

Thank you for your purchase. Please note that your purchase of this Elsevier eBook also includes access to an online version. Please click here (or go to http://ebooks.elsevier.com) to request an activation code and registration instructions in order to gain access to the web version.