

# GENOMES 4

**T.A. BROWN**

# GENOMES 4



**Taylor & Francis**

Taylor & Francis Group

<http://taylorandfrancis.com>

# GENOMES 4

**T. A. BROWN**

## Garland Science

Vice President: Denise Schanck  
Senior Editor: Elizabeth Owen  
Assistant Editor: David Borrowdale  
Senior Production Editor: Georgina Lucas  
Production Editor: Deepa Divakaran  
Illustrator: Matthew McClements, Blink Studio Ltd  
Layout: Georgina Lucas  
Cover Designer: Matthew McClements, Blink Studio Ltd  
Copyeditor: Heather Whirlow Cammarn  
Proofreader: Jo Clayton  
Indexer: Bill Johncocks

The cover image shows a circo plot showing the similarities between the genomes of four species: human, chimpanzee, mouse, and zebrafish. Courtesy of Martin Krzywinski, BC Cancer Research Centre.

© 2018 by Garland Science, Taylor & Francis Group, LLC

This book contains information obtained from authentic and highly regarded sources. Every effort has been made to trace copyright holders and to obtain their permission for the use of copyright material. Reprinted material is quoted with permission, and sources are indicated. A wide variety of references are listed. Reasonable efforts have been made to publish reliable data and information, but the author and the publisher cannot assume responsibility for the validity of all materials or for the consequences of their use.

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means—graphic, electronic, or mechanical, including photocopying, recording, taping, or information storage and retrieval systems—without permission of the copyright holder.

The publisher makes no representation, express or implied, that the drug doses in this book are correct. Readers must check up-to-date product information and clinical procedures with the manufacturers, current codes of conduct, and current safety regulations.

ISBN 9780815345084

## Library of Congress Cataloging-in-Publication Data

Names: Brown, T. A. (Terence A.), author.  
Title: Genomes 4 / Terry Brown.  
Other titles: Genomes | Genomes four  
Description: 4th. | New York, NY : Garland Science, [2017] | Preceded by: Genomes 3 / T. A. Brown. 3rd ed. New York : Garland Science, c2007. | Includes bibliographical references and index.  
Identifiers: LCCN 2017013507 | ISBN 9780815345084 (alk. paper)  
Subjects: | MESH: Genome  
Classification: LCC QH447 | NLM QU 470 | DDC 572.8/6--dc23  
LC record available at <https://lcn.loc.gov/2017013507>

Published by Garland Science, Taylor & Francis Group, LLC, an informa business,  
711 Third Avenue, New York, NY 10017, USA, and 2 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

 **Garland Science**  
Taylor & Francis Group

Visit our web site at <http://www.garlandscience.com>

## About the Author

I became fascinated with the natural world when I was very young. I began my research career studying the effects of metal pollution on microorganisms and the tolerance that some plants display to high concentrations of toxic metals. I then became excited by DNA and worked on mitochondrial genes in fungi in order to learn the new (in those days) techniques for gene cloning and DNA sequencing. I contributed to the discovery of mitochondrial introns and to work that described the base-paired structure of these introns. I then became interested in ancient DNA and was one of the first people to carry out DNA extractions with bones and preserved plant remains. This work has required close collaboration with archaeologists, and has led to my current interests in paleogenomics, the origins of agriculture, and the evolution of domesticated plants.

I obtained my PhD from University College London in 1977 and then worked in New York, Oxford, Colchester, and Manchester before beginning in 1984 as a Lecturer in Biotechnology at the University of Manchester Institute of Science and Technology (UMIST). I was appointed Professor of Biomolecular Archaeology in 2000 and was Head of Biomolecular Sciences at UMIST from 2002–2004. I was then Associate Dean in the Faculty of Life Sciences of the University of Manchester until 2006, before taking a break from administration in order to have more time to do research.

My other undergraduate textbooks include Introduction to Genetics, A Molecular Approach (Garland Science).

# PREFACE

There have been remarkable advances in our knowledge of genomes since the previous edition of this book was published ten years ago. Back in 2007, next-generation sequencing was in its infancy and high-throughput methods for transcriptomics and proteomics were only beginning to be exploited. The application of these methods over the last ten years has resulted in an exponential increase in the number of species for which genome sequences and annotations are now available, and has enabled multiple versions of the genome of a single species to be examined. The profusion of new sequences has had a particularly dramatic impact on bacterial genomics, with introduction of the pan-genome concept and the discovery of extensive lateral transfer of genes between species. Our knowledge of eukaryotic genomes has undergone equally dramatic change, with the discovery of new types of noncoding RNA, including the vast numbers of long RNAs that are transcribed from the supposedly intergenic regions of many genomes.

*Genomes 4* retains the overall structure of the previous editions, with the book divided in four parts, on genome sequencing and annotation, genome anatomies, genome expression, and genome replication and evolution. With some small changes, the order of chapters remains unchanged. However, the text throughout has been completely updated and, in many chapters, substantially revised. In particular, the development of transcriptomics and proteomics has reached the point where in *Genomes 4* it is possible to describe the processes of transcription and translation from a genomewide perspective, rather than simply through an examination of the expression of individual genes. This was my aim when I wrote the first edition of *Genomes* way back in 1999, but the information available at that time meant that these core chapters were fairly orthodox treatments of gene rather than genome expression. We are still some way from being able to describe the entire expression of a genome as a single integrated process, but we are getting there and I hope that in *Genomes 4* I have been able to convey to the reader at least some aspects of the joined-up nature of genome expression.

*Genomes 4* has been a long time in the making and I would like to thank Liz Owen of Garland Science for her continued enthusiasm for the book and her gentle reminders about approaching deadlines. I also wish to thank David Borrowdale and Georgina Lucas at Garland for managing the production of the book, and Matthew McClements for his splendid artwork. As with the previous editions, *Genomes 4* would not have been finished without the support of my wife, Keri. The acknowledgment in the first edition that “if you find this book useful then you should thank Keri, not me, because she is the one who ensured that it was written” is equally true for the fourth edition.

## A NOTE TO THE READER

I have tried to make the fourth edition of *Genomes* as user friendly as possible. The book therefore includes a number of devices intended to help the reader and to make the book an effective teaching and learning aid.

### Organization of the Book

*Genomes 4* is divided into four parts:

**Part I – Studying Genomes** begins with an orientation chapter that introduces the reader to genomes, transcriptomes, and proteomes, and then in [Chapter 2](#) moves on to the methods, centered on PCR and cloning, that were used in the pre-genome era to examine individual genes. The techniques that are used for constructing genetic and physical maps, which are still important in many genome projects, are then described in [Chapter 3](#), followed in [Chapter 4](#) by the methodology for obtaining DNA sequences and assembling reads into draft and finished genomes sequences. Two chapters are then devoted to analysis of genome sequences: [Chapter 5](#) on the annotation of a genome by identification of genes and other features, and [Chapter 6](#) on functional analysis of the genes that are discovered.

**Part II – Genome Anatomies** surveys the anatomies of the various types of genome that are found on our planet. [Chapter 7](#) covers eukaryotic nuclear genomes, with emphasis on the human genome, partly because of the importance of the human genome in so many areas of research, but also because our genome is the best studied of all those for which sequences are available. [Chapter 8](#) deals with the genomes of prokaryotes and of eukaryotic organelles, the latter included here because of their prokaryotic origins, and [Chapter 9](#) describes viral genomes and mobile genetic elements, these being grouped together because some types of mobile element are related to viral genomes.

**Part III – How Genomes are Expressed** describes how the biological information contained in a genome is utilized by the cell within which that genome resides. [Chapter 10](#) addresses the important issue of how the packaging of DNA into chromatin affects expression of different parts of the genome, and [Chapter 11](#) then describes the central role that DNA-binding proteins play in expressing those parts of the genome that are active at a particular time. [Chapter 12](#) moves on to the transcriptome, describing how transcriptomes are studied, their compositions, and how a cell's transcriptome is synthesized and maintained. [Chapter 13](#) gives an equivalent description of proteomics and the proteome, and [Chapter 14](#) concludes this part of the book by exploring how the genome acts within the context of the cell and organism, responding to extracellular signals and driving the biochemical changes that underlie differentiation and development.

**Part IV – How Genomes Replicate and Evolve** links DNA replication, mutation, and recombination with the gradual evolution of genomes over time. In [Chapters 15–17](#) the molecular processes responsible for replication, mutation, repair, and recombination are described, and in [Chapter 18](#) the ways in which these processes are thought to have shaped the structures and genetic contents of genomes over evolutionary time are considered. [Chapter 18](#) then ends with a small number of case studies to illustrate how molecular phylogenomics and population genomics are being used in research and biotechnology.

## LEARNING AIDS

Each chapter has a set of Short Answer Questions and In-Depth Problems, as well as an annotated Further Reading list. At the end of the book there is an extensive Glossary.

**Short answer questions** require 50- to 500-word answers. The questions cover the entire content of each chapter in a fairly straightforward manner, and most can be marked simply by checking each answer against the relevant part of the text. A student can use the short answer questions to work systematically through a chapter, or can select individual ones in order to evaluate their ability to answer questions on specific topics. The short answer questions could also be used in closed-book tests.

**In-depth problems** require a more detailed answer. They vary in nature and in difficulty, the simplest requiring little more than a literature survey, the intention of these particular problems being that the student advances his or her learning a few stages from where *Genomes 4* leaves off. Other problems require that the student evaluates a statement or a hypothesis, based on their understanding of the material in the book, possibly supplemented by reading around the subject. These problems will, hopefully, engender a certain amount of thought and critical awareness. A few problems are difficult, in some cases to the extent that there is no solid answer to the question posed. These are designed to stimulate debate and speculation, which stretches the knowledge of each student and forces them to think carefully about their statements. The in-depth problems can be tackled by students working individually, or alternatively can form the starting point for a group discussion.

**Further Reading** lists at the end of each chapter include those research papers, reviews, and books that I look on as the most useful sources of additional material. My intention throughout *Genomes 4* has been that students should be able to use the reading lists to obtain further information when writing extended essays or dissertations on particular topics. Research papers are therefore included, but only if their content is likely to be understandable to the average reader of the book. Emphasis is also placed on accessible reviews, one strength of these general articles being the context and relevance that they provide to a piece of work. The reading lists are divided into sections reflecting the organization of information in the chapter, and in some cases I have appended a few words summarizing the particular value of each item to help the reader decide which ones he or she wishes to seek out. In some cases, Further Reading also includes URLs for databases and other online resources relevant to the material covered in a chapter.

The **Glossary** defines every term that is highlighted in bold in the text, along with a number of additional terms that the reader might come across when referring to books or articles in the reading lists. The glossary therefore provides a quick and convenient means by which the reader can remind themselves of the technical terms relevant to the study of genomes, and also acts as a revision aid to make sure those definitions are clearly understood during the minutes of uncertainty that many students experience immediately before an exam.



## INSTRUCTOR RESOURCES

The images from the book are available through [www.garlandscience.com](http://www.garlandscience.com) in two convenient formats: PowerPoint® and JPEG. They have been optimized for display on a computer. Figures are searchable by figure number, by figure name, or by keywords used in the figure legend from the book. Help on answering the In-Depth Problems, found at the end of each chapter, is also available.

## ACKNOWLEDGMENTS

The Author and Publisher of *Genomes 4* gratefully acknowledge the contribution of the following reviewers in the development of this edition.

David Baillie, Simon Fraser University; Linda Bonen, University of Ottawa; Hugh Cam, Boston College; Yuri Dubrova, University of Leicester; Bart Eggen, University of Groningen; Robert Fowler, San José State University; Sidney Fu, George Washington University; Adrian Hall, Sheffield Hallam University; Lee Hwei Huih, Universiti Tunku Abdul Rahman; Glyn Jenkins, Aberystwyth University; Julian M. Ketley, University of Leicester; Torsten Kristensen, University of Aarhus; Gerhard May, University of Dundee; Mike McPherson, University of Leeds; Isidoro Metón, Universitat de Barcelona; Gary Ogden, St. Mary's University; Paul Overvoorde, Macalester College; John Rafferty, University of Sheffield; Andrew Read, University of Manchester; Joaquin Cañizares Sales, Universitat Politècnica de València; Michael Schweizer, Heriot-Watt University; Eric Spana, Duke University; David Studholme, Exeter University; John Taylor, University of Newcastle; Gavin Thomas, University of York; Matthew Upton, Plymouth University; Guido van den Ackerveken, Utrecht University; Vassie Ware, Lehigh University; Wei Zhang, Illinois Institute of Technology.

---

# CONTENTS IN BRIEF

<b>CHAPTER 1</b>	GENOMES, TRANSCRIPTOMES, AND PROTEOMES	1
<b>CHAPTER 2</b>	STUDYING DNA	27
<b>CHAPTER 3</b>	MAPPING GENOMES	55
<b>CHAPTER 4</b>	SEQUENCING GENOMES	87
<b>CHAPTER 5</b>	GENOME ANNOTATION	119
<b>CHAPTER 6</b>	IDENTIFYING GENE FUNCTIONS	135
<b>CHAPTER 7</b>	EUKARYOTIC NUCLEAR GENOMES	155
<b>CHAPTER 8</b>	GENOMES OF PROKARYOTES AND EUKARYOTIC ORGANELLES	181
<b>CHAPTER 9</b>	VIRAL GENOMES AND MOBILE GENETIC ELEMENTS	203
<b>CHAPTER 10</b>	ACCESSING THE GENOME	219
<b>CHAPTER 11</b>	THE ROLE OF DNA-BINDING PROTEINS IN GENOME EXPRESSION	241
<b>CHAPTER 12</b>	TRANSCRIPTOMES	257
<b>CHAPTER 13</b>	PROTEOMES	293
<b>CHAPTER 14</b>	GENOME EXPRESSION IN THE CONTEXT OF CELL AND ORGANISM	329
<b>CHAPTER 15</b>	GENOME REPLICATION	357
<b>CHAPTER 16</b>	MUTATIONS AND DNA REPAIR	389
<b>CHAPTER 17</b>	RECOMBINATION AND TRANSPOSITION	411
<b>CHAPTER 18</b>	HOW GENOMES EVOLVE	429
<b>GLOSSARY</b>		463
<b>INDEX</b>		491

# CONTENTS

## CHAPTER 1 GENOMES, TRANSCRIPTOMES, AND PROTEOMES

### 1.1 DNA

Genes are made of DNA

DNA is a polymer of nucleotides

The double helix is stabilized by base pairing and base stacking

The double helix has structural flexibility

### 1.2 RNA AND THE TRANSCRIPTOME

RNA is a second type of polynucleotide

The RNA content of the cell

Many RNAs are synthesized as precursor molecules

There are different definitions of the transcriptome

### 1.3 PROTEINS AND THE PROTEOME

There are four hierarchical levels of protein structure

Amino acid diversity underlies protein diversity

The link between the transcriptome and the proteome

The genetic code is not universal

The link between the proteome and the biochemistry of the cell

### SUMMARY

### SHORT ANSWER QUESTIONS

### IN-DEPTH PROBLEMS

### FURTHER READING

## CHAPTER 2 STUDYING DNA

### 2.1 ENZYMES FOR DNA MANIPULATION

The mode of action of a template-dependent DNA polymerase

The types of DNA polymerase used in research

Restriction endonucleases enable DNA molecules to be cut at defined positions

Gel electrophoresis is used to examine the results of a restriction digest

Interesting DNA fragments can be identified by Southern hybridization

Ligases join DNA fragments together 37  
End-modification enzymes 38

## 1 2.2 THE POLYMERASE CHAIN REACTION 38

Carrying out a PCR 39  
The rate of product formation can be followed during a PCR 40  
PCR has many and diverse applications 41

## 8 2.3 DNA CLONING 41

Why is gene cloning important? 41  
The simplest cloning vectors are based on *E. coli* plasmids 43  
Bacteriophages can also be used as cloning vectors 44  
Vectors for longer pieces of DNA 47  
DNA can be cloned in organisms other than *E. coli* 48

### SUMMARY 50

### SHORT ANSWER QUESTIONS 51

### IN-DEPTH PROBLEMS 51

### FURTHER READING 52

## CHAPTER 3 MAPPING GENOMES 55

### 3.1 WHY A GENOME MAP IS IMPORTANT 55

Genome maps are needed in order to sequence the more complex genomes 55

Genome maps are not just sequencing aids 57

## 27 3.2 MARKERS FOR GENETIC MAPPING 58

Genes were the first markers to be used 58

RFLPs and SSLPs are examples of DNA markers 59

Single-nucleotide polymorphisms are the most useful type of DNA marker 61

### 3.3 THE BASIS TO GENETIC MAPPING 63

The principles of inheritance and the discovery of linkage 63

Partial linkage is explained by the behavior of chromosomes during meiosis 65

From partial linkage to genetic mapping 68

### 3.4 LINKAGE ANALYSIS WITH DIFFERENT TYPES OF ORGANISMS

Linkage analysis when planned breeding experiments are possible	69
Gene mapping by human pedigree analysis	71
Genetic mapping in bacteria	73
The limitations of linkage analysis	74

### 3.5 PHYSICAL MAPPING BY DIRECT EXAMINATION OF DNA MOLECULES

Conventional restriction mapping is applicable only to small DNA molecules	75
Optical mapping can locate restriction sites in longer DNA molecules	77
Optical mapping can be used to map other features in a DNA molecule	79

### 3.6 PHYSICAL MAPPING BY ASSIGNING MARKERS TO DNA FRAGMENTS

Any unique sequence can be used as an STS	81
DNA fragments for STS mapping can be obtained as radiation hybrids	82
A clone library can be used as the mapping reagent	83

<b>SUMMARY</b>	84
----------------	----

<b>SHORT ANSWER QUESTIONS</b>	85
-------------------------------	----

<b>IN-DEPTH PROBLEMS</b>	85
--------------------------	----

<b>FURTHER READING</b>	86
------------------------	----

## CHAPTER 4 SEQUENCING GENOMES 87

### 4.1 CHAIN-TERMINATION SEQUENCING

Chain-termination sequencing in outline	87
Not all DNA polymerases can be used for sequencing	89
Chain-termination sequencing with <i>Taq</i> polymerase	90
Strengths and limitations of chain-termination sequencing	91

### 4.2 NEXT-GENERATION SEQUENCING

Preparation of a sequencing library is the common feature of next-generation methods	92
Various next-generation sequencing methods have been devised	95
Third- and fourth-generation methods enable sequencing in real time	97

### 4.3 HOW TO SEQUENCE A GENOME

The potential of the shotgun method was proven by the <i>Haemophilus influenzae</i> sequence	98
Many prokaryotic genomes have been sequenced by the shotgun method	99
	100

Shotgun sequencing of eukaryotic genomes requires sophisticated assembly programs	102
More complex genomes can be sequenced by a hierarchical shotgun approach	104
What is a genome sequence and do we always need one?	107

### 4.4 A SURVEY OF EUKARYOTIC GENOME SEQUENCING PROJECTS

The Human Genome Project: genome sequencing in the heroic age	109
The Neanderthal genome: assembly of an extinct genome by use of the human sequence as a reference	110
The giant panda genome: shotgun sequencing based entirely on next-generation data	111
The barley genome: the concept of gene space	113

<b>SUMMARY</b>	115
----------------	-----

<b>SHORT ANSWER QUESTIONS</b>	115
-------------------------------	-----

<b>IN-DEPTH PROBLEMS</b>	116
--------------------------	-----

<b>FURTHER READING</b>	117
------------------------	-----

## CHAPTER 5 GENOME ANNOTATION 119

### 5.1 GENOME ANNOTATION BY COMPUTER ANALYSIS OF THE DNA SEQUENCE

The coding regions of genes are open reading frames	119
Simple ORF scans are less effective with genomes of higher eukaryotes	120
Locating genes for noncoding RNA	122
Homology searches and comparative genomics give an extra dimension to gene prediction	123

### 5.2 GENOME ANNOTATION BY ANALYSIS OF GENE TRANSCRIPTS

Hybridization tests can determine if a fragment contains transcribed sequences	125
Methods are available for precise mapping of the ends of transcripts	126
Exon–intron boundaries can also be located with precision	126

### 5.3 ANNOTATION BY GENOMEWIDE RNA MAPPING

Tiling arrays enable transcripts to be mapped onto chromosomes or entire genomes	128
Transcript sequences can be directly mapped onto a genome	129

<b>5.4 GENOME BROWSERS</b>	131
----------------------------	-----

**SUMMARY** 132  
**SHORT ANSWER QUESTIONS** 132  
**IN-DEPTH PROBLEMS** 133  
**FURTHER READING** 133

**CHAPTER 6**  
**IDENTIFYING GENE FUNCTIONS** 135

**6.1 COMPUTER ANALYSIS OF GENE FUNCTION** 135  
 Homology reflects evolutionary relationships 135  
 Homology analysis can provide information on the function of a gene 136  
 Identification of protein domains can help to assign function to an unknown gene 137  
 Annotation of gene function requires a common terminology 138

**6.2 ASSIGNING FUNCTION BY GENE INACTIVATION AND OVEREXPRESSION** 139  
 Functional analysis by gene inactivation 140  
 Individual genes can be inactivated by homologous recombination 140  
 Gene inactivation without homologous recombination 142  
 Gene overexpression can also be used to assess function 144  
 The phenotypic effect of gene inactivation or overexpression may be difficult to discern 145

**6.3 UNDERSTANDING GENE FUNCTION BY STUDIES OF EXPRESSION PATTERN AND PROTEIN PRODUCT** 146  
 Reporter genes and immunocytochemistry can be used to locate where and when genes are expressed 146  
 Directed mutagenesis can be used to probe gene function in detail 147

**6.4 USING CONVENTIONAL GENETIC ANALYSIS TO IDENTIFY GENE FUNCTION** 149  
 Identification of human genes responsible for inherited diseases 150  
 Genomewide association studies can also identify genes for diseases and other traits 151

**SUMMARY** 152  
**SHORT ANSWER QUESTIONS** 153  
**IN-DEPTH PROBLEMS** 153  
**FURTHER READING** 154

**CHAPTER 7**  
**EUKARYOTIC NUCLEAR GENOMES** 155

**7.1 NUCLEAR GENOMES ARE CONTAINED IN CHROMOSOMES** 155  
 Chromosomes are much shorter than the DNA molecules they contain 155  
 Special features of metaphase chromosomes 157  
 DNA–protein interactions in centromeres and telomeres 159

**7.2 HOW ARE THE GENES ARRANGED IN A NUCLEAR GENOME?** 161  
 Genes are not evenly distributed within a genome 161  
 A segment of the human genome 162  
 The yeast genome is very compact 164  
 Gene organization in other eukaryotes 165

**7.3 HOW MANY GENES ARE THERE AND WHAT ARE THEIR FUNCTIONS?** 167  
 Gene numbers can be misleading 168  
 Gene catalogs reveal the distinctive features of different organisms 169  
 Families of genes 172  
 Pseudogenes and other evolutionary relics 174

**7.4 THE REPETITIVE DNA CONTENT OF EUKARYOTIC NUCLEAR GENOMES** 176  
 Tandemly repeated DNA is found at centromeres and elsewhere in eukaryotic chromosomes 176  
 Minisatellites and microsatellites 176  
 Interspersed repeats 177

**SUMMARY** 178  
**SHORT ANSWER QUESTIONS** 178  
**IN-DEPTH PROBLEMS** 179  
**FURTHER READING** 179

**CHAPTER 8**  
**GENOMES OF PROKARYOTES AND EUKARYOTIC ORGANELLES** 181

**8.1 PHYSICAL FEATURES OF PROKARYOTIC GENOMES** 181  
 The traditional view of the prokaryotic chromosome 181  
 Some bacteria have linear or multipartite genomes 183

**8.2 GENETIC FEATURES OF PROKARYOTIC GENOMES** 186  
 Gene organization in the *E. coli* K12 genome 186

Operons are characteristic features of prokaryotic genomes	188		
Prokaryotic genome sizes and numbers of genes vary according to biological complexity	189		
Genome sizes and numbers of genes vary within individual species	190		
Distinctions between prokaryotic species are further blurred by lateral gene transfer	192		
Metagenomes describe the members of a community	194		
<b>8.3 EUKARYOTIC ORGANELLAR GENOMES</b>	195		
The endosymbiont theory explains the origin of organellar genomes	195		
Most organellar genomes are circular	196		
The gene catalogs of organellar genomes	197		
<b>SUMMARY</b>	198		
<b>SHORT ANSWER QUESTIONS</b>	200		
<b>IN-DEPTH PROBLEMS</b>	201		
<b>FURTHER READING</b>	201		
<b>CHAPTER 9</b>			
<b>VIRAL GENOMES AND MOBILE GENETIC ELEMENTS</b>	203		
<b>9.1 THE GENOMES OF BACTERIOPHAGES AND EUKARYOTIC VIRUSES</b>	203		
Bacteriophage genomes have diverse structures and organizations	203		
Replication strategies for bacteriophage genomes	205		
Structures and replication strategies for eukaryotic viral genomes	206		
Some retroviruses cause cancer	207		
Genomes at the edge of life	209		
<b>9.2 MOBILE GENETIC ELEMENTS</b>	210		
RNA transposons with long terminal repeats are related to viral retroelements	210		
Some RNA transposons lack long terminal repeats	212		
DNA transposons are common in prokaryotic genomes	213		
DNA transposons are less common in eukaryotic genomes	214		
<b>SUMMARY</b>	216		
<b>SHORT ANSWER QUESTIONS</b>	216		
<b>IN-DEPTH PROBLEMS</b>	217		
<b>FURTHER READING</b>	217		
<b>CHAPTER 10</b>			
<b>ACCESSING THE GENOME</b>	219		
<b>10.1 INSIDE THE NUCLEUS</b>	219		
The nucleus has an ordered internal structure	220		
The DNA content of a nondividing nucleus displays different degrees of packaging	221		
The nuclear matrix is thought to provide attachment points for chromosomal DNA	222		
Each chromosome has its own territory within the nucleus	223		
Each chromosome comprises a series of topologically associated domains	224		
Insulators mark the boundaries of topologically associated domains	226		
<b>10.2 NUCLEOSOME MODIFICATIONS AND GENOME EXPRESSION</b>	228		
Acetylation of histones influences many nuclear activities including genome expression	228		
Histone deacetylation represses active regions of the genome	229		
Acetylation is not the only type of histone modification	230		
Nucleosome repositioning also influences gene expression	231		
<b>10.3 DNA MODIFICATION AND GENOME EXPRESSION</b>	234		
Genome silencing by DNA methylation	234		
Methylation is involved in genomic imprinting and X inactivation	235		
<b>SUMMARY</b>	236		
<b>SHORT ANSWER QUESTIONS</b>	237		
<b>IN-DEPTH PROBLEMS</b>	238		
<b>FURTHER READING</b>	238		
<b>CHAPTER 11</b>			
<b>THE ROLE OF DNA-BINDING PROTEINS IN GENOME EXPRESSION</b>	241		
<b>11.1 METHODS FOR STUDYING DNA-BINDING PROTEINS AND THEIR ATTACHMENT SITES</b>	241		
X-ray crystallography provides structural data for any protein that can be crystallized	241		
NMR spectroscopy is used to study the structures of small proteins	243		
Gel retardation identifies DNA fragments that bind to proteins	244		

Protection assays pinpoint binding sites with greater accuracy	244	RNA silencing was first identified as a means of destroying invading viral RNA	276
Modification interference identifies nucleotides central to protein binding	246	MicroRNAs regulate genome expression by causing specific target mRNAs to be degraded	278
Genomewide scans for protein attachment sites	247		
<b>11.2 THE SPECIAL FEATURES OF DNA-BINDING PROTEINS</b>	249	<b>12.4 INFLUENCE OF RNA PROCESSING ON THE COMPOSITION OF A TRANSCRIPTOME</b>	278
The helix–turn–helix motif is present in prokaryotic and eukaryotic proteins	249	The splicing pathway for eukaryotic pre-mRNA introns	279
Zinc fingers are common in eukaryotic proteins	250	The splicing process must have a high degree of precision	280
Other nucleic acid-binding motifs	251	Enhancer and silencer elements specify alternative splicing pathways	282
<b>11.3 INTERACTION BETWEEN DNA AND ITS BINDING PROTEINS</b>	252	<b>12.5 TRANSCRIPTOMES IN RESEARCH</b>	284
Direct readout of the nucleotide sequence	252	Transcriptome analysis as an aid to genome annotation	284
The nucleotide sequence has a number of indirect effects on helix structure	253	Cancer transcriptomes	286
Contacts between DNA and proteins	253	Transcriptomes and the responses of plants to stress	287
<b>SUMMARY</b>	254	<b>SUMMARY</b>	289
<b>SHORT ANSWER QUESTIONS</b>	255	<b>SHORT ANSWER QUESTIONS</b>	289
<b>IN-DEPTH PROBLEMS</b>	256	<b>IN-DEPTH PROBLEMS</b>	290
<b>FURTHER READING</b>	256	<b>FURTHER READING</b>	290
<b>CHAPTER 12</b>	<b>257</b>	<b>CHAPTER 13</b>	<b>293</b>
<b>TRANSCRIPTOMES</b>		<b>PROTEOMES</b>	
<b>12.1 COMPONENTS OF THE TRANSCRIPTOME</b>	257	<b>13.1 STUDYING THE COMPOSITION OF A PROTEOME</b>	293
The mRNA fraction of a transcriptome is small but complex	257	The separation stage of a protein profiling project	294
Short noncoding RNAs have diverse functions	259	The identification stage of a protein profiling project	297
Long noncoding RNAs are enigmatic transcripts	260	Comparing the compositions of two proteomes	299
Microarray analysis and RNA sequencing are used to study the contents of transcriptomes	262	Analytical protein arrays offer an alternative approach to protein profiling	300
<b>12.2 SYNTHESIS OF THE COMPONENTS OF THE TRANSCRIPTOME</b>	263	<b>13.2 IDENTIFYING PROTEINS THAT INTERACT WITH ONE ANOTHER</b>	301
RNA polymerases are molecular machines for making RNA	264	Identifying pairs of interacting proteins	301
Transcription start points are indicated by promoter sequences	266	Identifying the components of multiprotein complexes	304
Synthesis of bacterial RNA is regulated by repressor and activator proteins	268	Identifying proteins with functional interactions	305
Synthesis of bacterial RNA is also regulated by control over transcription termination	271	Protein interaction maps display the interactions within a proteome	306
Synthesis of eukaryotic RNA is regulated primarily by activator proteins	272	<b>13.3 SYNTHESIS AND DEGRADATION OF THE COMPONENTS OF THE PROTEOME</b>	308
<b>12.3 DEGRADATION OF THE COMPONENTS OF THE TRANSCRIPTOME</b>	275	Ribosomes are molecular machines for making proteins	308
Several processes are known for nonspecific RNA turnover	275		

During stress, bacteria inactivate their ribosomes in order to downsize the proteome	311	Yeast mating types are determined by gene conversion events	338
Initiation factors mediate large-scale remodeling of eukaryotic proteomes	312	Genome rearrangements are responsible for immunoglobulin and T-cell receptor diversity	339
The translation of individual mRNAs can also be regulated	313		
Degradation of the components of the proteome	314		
<b>13.4 INFLUENCE OF PROTEIN PROCESSING ON THE COMPOSITION OF THE PROTEOME</b>		<b>14.3 CHANGES IN GENOME ACTIVITY UNDERLYING DEVELOPMENT</b>	341
The amino acid sequence contains instructions for protein folding	315	Bacteriophage $\lambda$ : a genetic switch enables a choice to be made between alternative developmental pathways	342
Some proteins are activated by proteolytic cleavage	315	<i>Bacillus</i> sporulation: coordination of activities in two distinct cell types	343
Important changes in protein activity can be brought about by chemical modification	318	<i>Caenorhabditis elegans</i> : the genetic basis of positional information and the determination of cell fate	346
	320	Fruit flies: conversion of positional information into a segmented body plan	348
	322	Homeotic selector genes are universal features of higher eukaryotic development	350
	322	Homeotic genes also underlie plant development	352
	323		
<b>SUMMARY</b>	326	<b>SUMMARY</b>	352
<b>SHORT ANSWER QUESTIONS</b>	326	<b>SHORT ANSWER QUESTIONS</b>	353
<b>IN-DEPTH PROBLEMS</b>	327	<b>IN-DEPTH PROBLEMS</b>	354
<b>FURTHER READING</b>	327	<b>FURTHER READING</b>	354
<b>CHAPTER 14</b>		<b>CHAPTER 15</b>	
<b>GENOME EXPRESSION</b>		<b>GENOME REPLICATION</b>	<b>357</b>
<b>IN THE CONTEXT OF CELL AND ORGANISM</b>	<b>329</b>		
<b>14.1 THE RESPONSE OF THE GENOME TO EXTERNAL SIGNALS</b>		<b>15.1 THE TOPOLOGY OF GENOME REPLICATION</b>	357
Signal transmission by import of the extracellular signaling compound	330	The double-helical structure complicates the replication process	358
Receptor proteins transmit signals across cell membranes	330	The Meselson–Stahl experiment proved that replication is semiconservative	359
Some signal transduction pathways have few steps between receptor and genome	332	DNA topoisomerases provide a solution to the topological problem	361
Some signal transduction pathways have many steps between receptor and genome	333	Variations on the semiconservative theme	363
Some signal transduction pathways operate via second messengers	334		
	336	<b>15.2 THE INITIATION PHASE OF GENOME REPLICATION</b>	364
		Initiation at the <i>E. coli</i> origin of replication	364
		Origins of replication have been clearly defined in yeast	365
		Origins in higher eukaryotes have been less easy to identify	366
<b>14.2 CHANGES IN GENOME ACTIVITY RESULTING IN CELLULAR DIFFERENTIATION</b>	336	<b>15.3 EVENTS AT THE REPLICATION FORK</b>	367
Some differentiation processes involve changes to chromatin structure	336	DNA polymerases are molecular machines for making (and degrading) DNA	367
		DNA polymerases have limitations that complicate genome replication	369



Okazaki fragments must be joined together to complete lagging-strand replication	370	Defects in DNA repair underlie human diseases, including cancers	406
<b>15.4 TERMINATION OF GENOME REPLICATION</b>	372	<b>SUMMARY</b>	406
Replication of the <i>E. coli</i> genome terminates within a defined region	373	<b>SHORT ANSWER QUESTIONS</b>	407
Little is known about termination of replication in eukaryotes	374	<b>IN-DEPTH PROBLEMS</b>	407
Telomerase completes replication of chromosomal DNA molecules, at least in some cells	375	<b>FURTHER READING</b>	408
Telomere length is implicated in cell senescence and cancer	378	<b>CHAPTER 17</b>	
<i>Drosophila</i> has a unique solution to the end-shortening problem	379	<b>RECOMBINATION AND TRANSPOSITION</b>	<b>411</b>
<b>15.5 REGULATION OF EUKARYOTIC GENOME REPLICATION</b>	380	<b>17.1 HOMOLOGOUS RECOMBINATION</b>	412
Genome replication must be synchronized with the cell cycle	380	The Holliday and Meselson–Radding models for homologous recombination	412
Origin licensing is the prerequisite for passing the G1–S checkpoint	380	The double-strand break model for homologous recombination	414
Replication origins do not all fire at the same time	382	RecBCD is the most important pathway for homologous recombination in bacteria	415
The cell has various options if the genome is damaged	383	<i>E. coli</i> can also carry out homologous recombination by the RecFOR pathway	417
<b>SUMMARY</b>	384	Homologous recombination pathways in eukaryotes	417
<b>SHORT ANSWER QUESTIONS</b>	385	The primary role of homologous recombination is thought to be DNA repair	418
<b>IN-DEPTH PROBLEMS</b>	385	<b>17.2 SITE-SPECIFIC RECOMBINATION</b>	419
<b>FURTHER READING</b>	386	Bacteriophage $\lambda$ uses site-specific recombination during the lysogenic infection cycle	419
<b>CHAPTER 16</b>		Site-specific recombination is an aid in construction of genetically modified plants	421
<b>MUTATIONS AND DNA REPAIR</b>	<b>389</b>	<b>17.3 TRANSPOSITION</b>	421
<b>16.1 THE CAUSES OF MUTATIONS</b>	389	Replicative and conservative transposition of DNA transposons	422
Errors in replication are a source of point mutations	390	Retroelements transpose replicatively via an RNA intermediate	423
Replication errors can also lead to insertion and deletion mutations	391	<b>SUMMARY</b>	425
Mutations are also caused by chemical and physical mutagens	394	<b>SHORT ANSWER QUESTIONS</b>	426
<b>16.2 REPAIR OF MUTATIONS AND OTHER TYPES OF DNA DAMAGE</b>	398	<b>IN-DEPTH PROBLEMS</b>	427
Direct repair systems fill in nicks and correct some types of nucleotide modification	398	<b>FURTHER READING</b>	427
Base excision repairs many types of damaged nucleotide	399	<b>CHAPTER 18</b>	
Nucleotide excision repair is used to correct more extensive types of damage	401	<b>HOW GENOMES EVOLVE</b>	<b>429</b>
Mismatch repair corrects replication errors	402	<b>18.1 GENOMES: THE FIRST 10 BILLION YEARS</b>	429
Single- and double-strand breaks can be repaired	403	The first biochemical systems were centered on RNA	429
If necessary, DNA damage can be bypassed during genome replication	405	The first DNA genomes	432
		How unique is life?	433

<b>18.2 EVOLUTION OF INCREASINGLY COMPLEX GENOMES</b>	434
Genome sequences provide extensive evidence of past gene duplications	434
A variety of processes could result in gene duplication	438
Whole-genome duplication is also possible	439
Smaller duplications can also be identified in the human genome and other genomes	442
Both prokaryotes and eukaryotes acquire genes from other species	444
Genome evolution also involves rearrangement of existing genes	445
There are competing hypotheses for the origins of introns	448
The evolution of the epigenome	449
<b>18.3 GENOMES: THE LAST 6 MILLION YEARS</b>	450
The human genome is very similar to that of the chimpanzee	451
Paleogenomics is helping us understand the recent evolution of the human genome	452
<b>18.4 GENOMES TODAY: DIVERSITY IN POPULATIONS</b>	453
The origins of HIV/AIDS	454
The first migrations of humans out of Africa	455
The diversity of plant genomes is an aid in crop breeding	457
<b>SUMMARY</b>	458
<b>SHORT ANSWER QUESTIONS</b>	459
<b>IN-DEPTH PROBLEMS</b>	460
<b>FURTHER READING</b>	460
<b>GLOSSARY</b>	463
<b>INDEX</b>	491



**Taylor & Francis**

Taylor & Francis Group

<http://taylorandfrancis.com>

# PART I

## STUDYING GENOMES

### GENOMES, TRANSCRIPTOMES, AND PROTEOMES

#### CHAPTER

# 1

#### 1.1 DNA

#### 1.2 RNA AND THE TRANSCRIPTOME

#### 1.3 PROTEINS AND THE PROTEOME

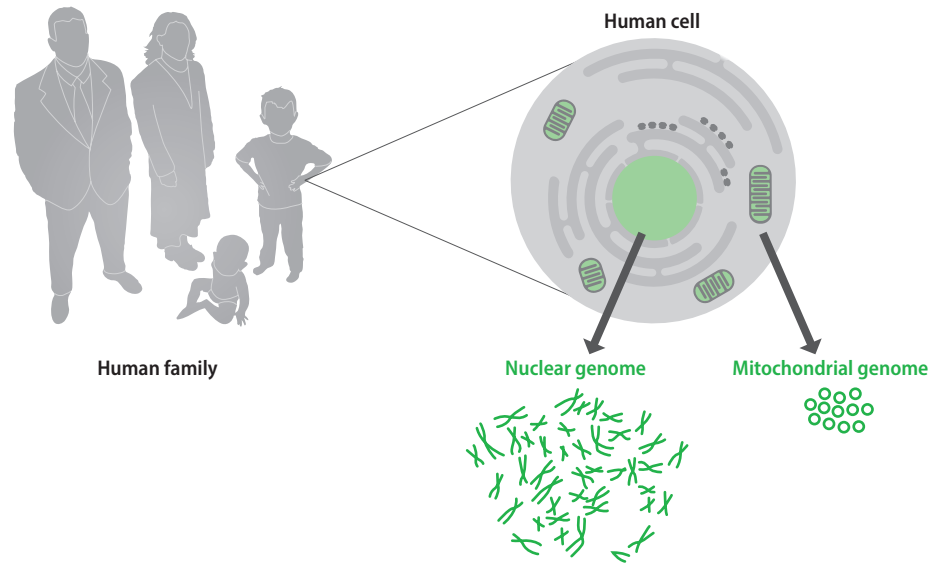
Life as we know it is specified by the **genomes** of the myriad organisms with which we share the planet. Every organism possesses a genome that contains the **biological information** needed to construct and maintain a living example of that organism. Most genomes, including the human genome and those of all other cellular life forms, are made of **DNA** (deoxyribonucleic acid), but a few viruses have **RNA** (ribonucleic acid) genomes. DNA and RNA are **polymeric** molecules made up of chains of monomeric subunits called **nucleotides**. Each molecule of DNA comprises two **polynucleotides** wound around one another to form the famous **double helix**, in which the two strands are held together by chemical bonds that link adjacent nucleotides into structures called **base pairs**.

The human genome, which is typical of the genomes of all multicellular animals, consists of two distinct parts (**Figure 1.1**):

- The **nuclear genome** comprises approximately 3,235,000,000 base pairs of DNA, divided into 24 linear molecules, the shortest 48,000,000 base pairs in length and the longest 250,000,000 base pairs, each contained in a different **chromosome**. These 24 chromosomes consist of 22 **autosomes** and the two **sex chromosomes**, X and Y. Altogether, some 45,500 **genes** are present in the human nuclear genome.
- The **mitochondrial genome** is a circular DNA molecule of 16,569 base pairs, up to 10 copies of which are present in each of the energy-generating organelles called mitochondria. The human mitochondrial genome contains just 37 genes.

Each of the approximately  $10^{13}$  cells in the adult human body has its own copy or copies of the nuclear genome, the only exceptions being those few cell types, such as red blood cells, that lack a **nucleus** in their fully differentiated state. The vast majority of cells are **diploid** and so have two copies of each autosome, plus two sex chromosomes, XX for females or XY for males—46 chromosomes in all. These are called **somatic cells**, in contrast to **sex cells**, or **gametes**, which are **haploid** and have just 23 chromosomes, one of each autosome and one sex chromosome. Each cell also has multiple copies of the mitochondrial genome: 2000–7000 copies in somatic cells, such as those in the liver and heart tissue, and over 100,000 copies in each female **oocyte**.

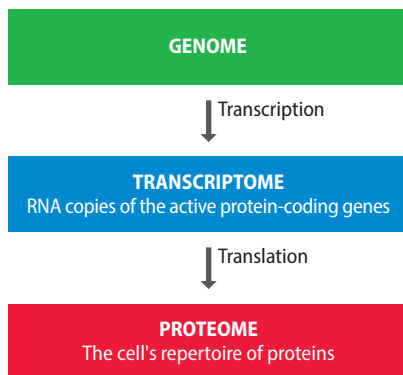
**Figure 1.1** Nuclear and mitochondrial components of the human genome.



The genome is a store of biological information, but on its own it is unable to release that information to the cell. Utilization of the biological information contained in the genome requires the coordinated activity of enzymes and other proteins, which participate in a complex series of biochemical reactions referred to as **genome expression** (Figure 1.2). The initial product of genome expression is the **transcriptome**, a collection of RNA molecules derived from those genes that are active in the cell at a particular time. The transcriptome is maintained by the process called **transcription**, in which individual genes are copied into RNA molecules. The second product of genome expression is the **proteome**, the cell's repertoire of **proteins**, which specifies the nature of the biochemical reactions that the cell is able to carry out. The proteins that make up the proteome are synthesized by **translation** of some of the individual RNA molecules present in the transcriptome.

This book is about genomes and genome expression. It explains how genomes are studied (Part I), how they are organized (Part II), how they function (Part III), and how they replicate and evolve (Part IV). It was not possible to write this book until quite recently. Since the 1950s, molecular biologists have studied individual genes or small groups of genes, and from these studies they have built up a wealth of knowledge about how genes work. But only during the last few years have techniques been available that make it possible to examine entire genomes. Individual genes are still intensively studied, but information about individual genes is now interpreted within the context of the genome as a whole. This new, broader emphasis applies not just to genomes but to all of biochemistry and cell biology. No longer is it sufficient simply to understand individual biochemical pathways or subcellular processes. The challenge now is provided by **systems biology**, which attempts to link together these pathways and processes into networks that describe the overall functioning of living cells and living organisms.

This book will lead you through our knowledge of genomes and show you how this exciting area of research is underpinning our developing understanding of biological systems. First, however, we must pay attention to the basic principles of molecular biology by reviewing the key features of the three types of biological molecule involved in genomes and genome expression: DNA, RNA, and protein.



**Figure 1.2** Genome expression. The genome specifies the transcriptome, and the transcriptome specifies the proteome.

## 1.1 DNA

DNA was discovered in 1869 by Friedrich Miescher, a Swiss biochemist working in Tübingen, Germany. The first extracts that Miescher made from human white blood cells were crude mixtures of DNA and chromosomal proteins, but the following year he moved to Basel, Switzerland (where the research institute

named after him is now located), and prepared a pure sample of **nucleic acid** from salmon sperm. Miescher's chemical tests showed that DNA is acidic and rich in phosphorus and also suggested that the individual molecules are very large, although it was not until the 1930s, when biophysical techniques were applied to DNA, that the huge lengths of the polymeric chains were fully appreciated.

## Genes are made of DNA

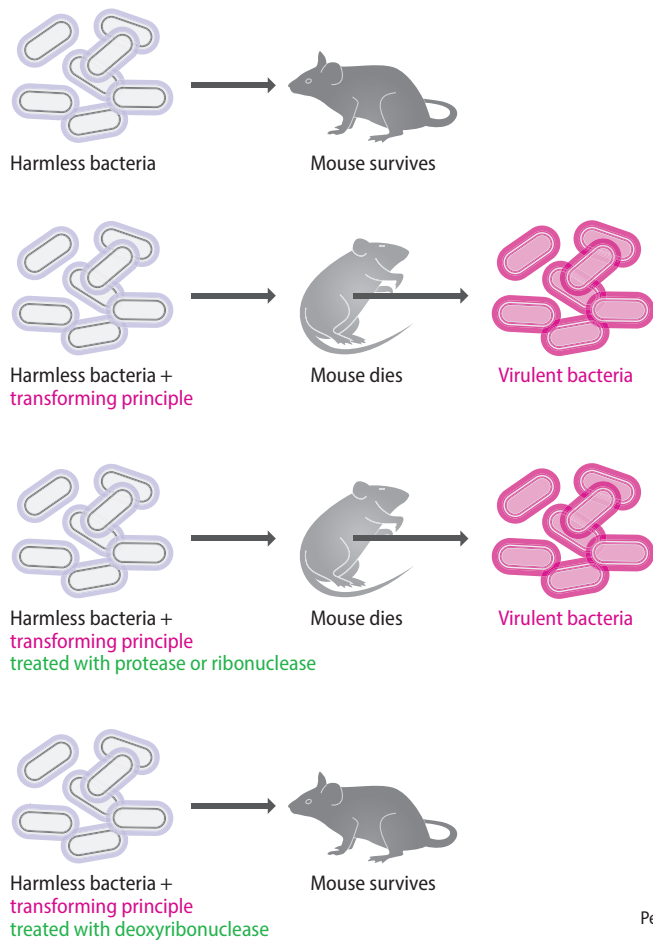
The fact that genes are made of DNA is so well known today that it can be difficult to appreciate that for the first 75 years after its discovery the true role of DNA was unsuspected. As early as 1903, W. S. Sutton had realized that the inheritance patterns of genes parallel the behavior of chromosomes during cell division, an observation that led to the **chromosome theory**, the proposal that genes are located in chromosomes. Examination of cells by **cytochemistry**, which makes use of stains that bind specifically to just one type of biochemical, showed that chromosomes are made of DNA and protein, in roughly equal amounts. Biologists at that time recognized that billions of different genes must exist and the genetic material must therefore be able to take many different forms. But this requirement appeared not to be satisfied by DNA, because in the early part of the twentieth century it was thought that all DNA molecules were the same. On the other hand, it was known, correctly, that proteins are highly variable, polymeric molecules, each one made up of a different combination of 20 chemically distinct amino acid monomers (Section 1.3). Genes simply had to be made of protein, not DNA.

The errors in understanding DNA structure lingered on, but by the late 1930s it had become accepted that DNA, like protein, has immense variability. The notion that protein was the genetic material initially remained strong but was eventually overturned by the results of two important experiments:

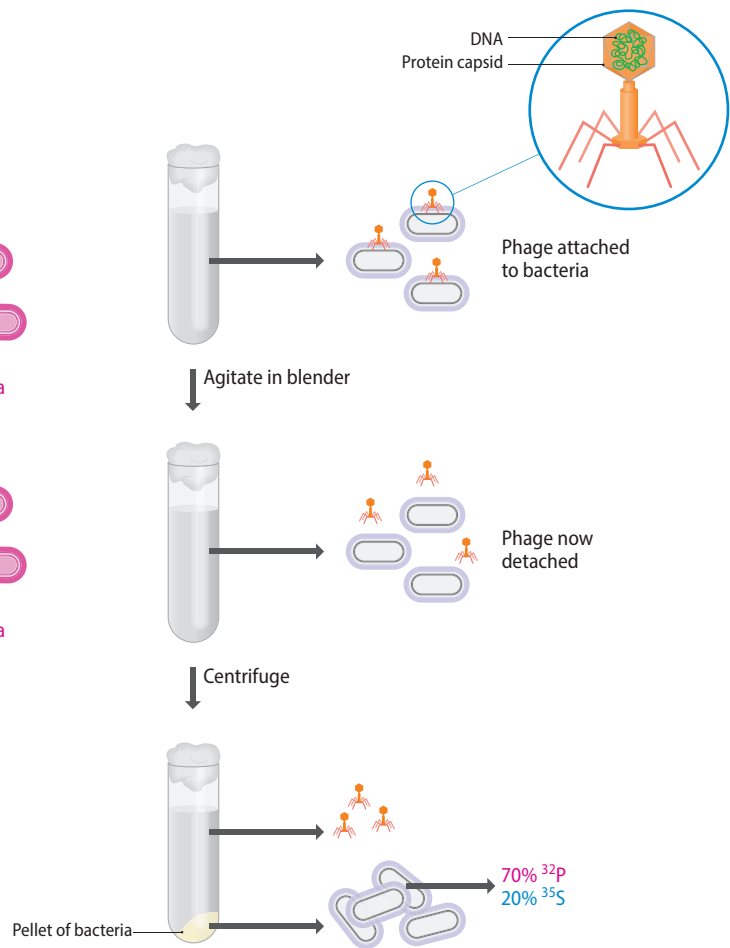
- Oswald Avery, Colin MacLeod, and Maclyn McCarty showed that DNA is the active component of the **transforming principle**, a bacterial cell extract that, when mixed with a harmless strain of *Streptococcus pneumoniae*, converts these bacteria into a virulent form capable of causing pneumonia when injected into mice (Figure 1.3A). In 1944, when the results of this experiment were published, only a few microbiologists appreciated that transformation involves transfer of genes from the cell extract into the living bacteria. However, once this point had been accepted, the true meaning of the Avery experiment became clear: bacterial genes must be made of DNA.
- Alfred Hershey and Martha Chase used **radiolabeling** to show that when a bacterial culture is infected with **bacteriophages** (also called **phages**, a type of virus), DNA is the major component of the bacteriophages that enters the cells (Figure 1.3B). This was a vital observation because it was known that, during the infection cycle, the genes of the infecting bacteriophages are used to direct synthesis of new bacteriophages, and this synthesis occurs within the bacteria. If only the DNA of the infecting bacteriophages enters the cells, then it follows that the genes of these bacteriophages must be made of DNA.

Although from our perspective these two experiments provide the key results that tell us that genes are made of DNA, biologists at the time were not so easily convinced. Both experiments have limitations that leave room for skeptics to argue that protein could still be the genetic material. For example, there were worries about the specificity of the **deoxyribonuclease** enzyme that Avery and colleagues used to inactivate the transforming principle. This result, a central part of the evidence for the transforming principle being DNA, would be invalid if, as seemed possible, the enzyme contained trace amounts of a contaminating **protease** and hence was also able to degrade protein. Neither is the bacteriophage experiment conclusive, as Hershey and Chase stressed when they published their results: "Our experiments show clearly that a physical separation of phage T2 into genetic and nongenetic parts is possible ... The chemical identification of the

## (A) The transforming principle



## (B) The Hershey–Chase experiment



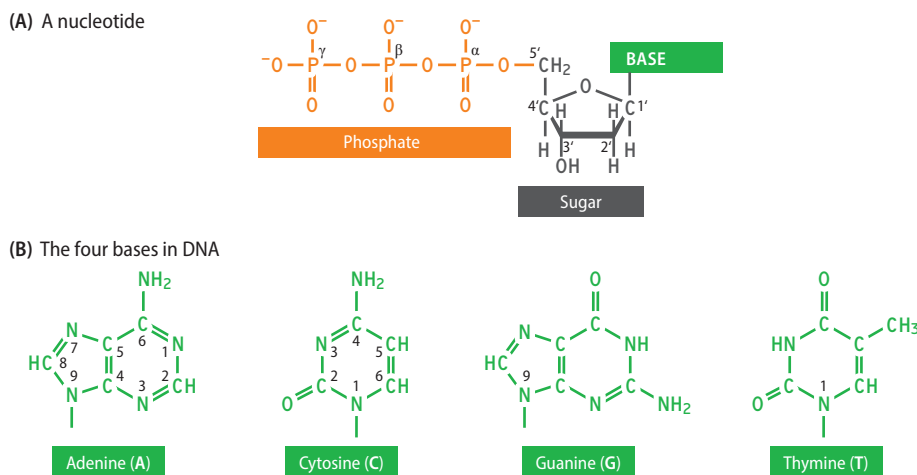
**Figure 1.3** The two experiments that suggested that genes are made of DNA. (A) Avery and colleagues showed that the transforming principle is made of DNA. The top two panels show what happens when mice are injected with harmless *Streptococcus pneumoniae* bacteria with or without addition of the transforming principle, a cell extract obtained from a virulent strain of *S. pneumoniae*. When the transforming principle is present, the mouse dies, because the genes in the transforming principle convert the harmless bacteria into the virulent form; these virulent bacteria subsequently were recovered from the lungs of the dead mouse. The lower two panels show that treatment with protease or ribonuclease has no effect on the transforming principle but that the transforming principle is inactivated by deoxyribonuclease. (B) The Hershey–Chase experiment used T2 bacteriophages, each of which comprises a DNA molecule contained in a protein capsid attached

to a body and legs that enable the bacteriophage to attach to the surface of a bacterium and inject its genes into the cell. The DNA of the bacteriophages was labeled with  $^{32}\text{P}$ , and the protein was labeled with  $^{35}\text{S}$ . A few minutes after infection, the culture was agitated to detach the empty phage particles from the cell surface. The culture was then centrifuged, which collects the bacteria plus phage genes as a pellet at the bottom of the tube but leaves the lighter phage particles in suspension. Hershey and Chase found that the bacterial pellet contained 70% of the  $^{32}\text{P}$ -labeled component of the phages (the DNA) but only 20% of the  $^{35}\text{S}$ -labeled material (the phage protein). In a second experiment, not depicted here, Hershey and Chase showed that new phages produced at the end of the infection cycle contained less than 1% of the protein from the parent phages. For more details of the bacteriophage infection cycle, see [Figure 2.27](#).

genetic part must wait, however, until some questions ... have been answered.” In retrospect, these two experiments are important not because of what they tell us but because they alerted biologists to the fact that DNA might be the genetic material and was therefore worth studying. This is what influenced Watson and Crick to work on DNA, and as we will see, it was their discovery of the double-helix structure, which solved the puzzling question of how genes can replicate, that really convinced the scientific world that genes are made of DNA.

### DNA is a polymer of nucleotides

The names of James Watson and Francis Crick are so closely linked with DNA that it is easy to forget that when they began their collaboration in October 1951, the detailed structure of the DNA polymer was already known. Their contribution was



**Figure 1.4 Structure of a nucleotide.** (A) General structure of a deoxyribonucleotide, which is the type of nucleotide found in DNA. (B) The four bases that occur in deoxyribonucleotides.

not to determine the structure of DNA per se but to show that in living cells two DNA chains are intertwined to form the double helix. First, therefore, we should examine what Watson and Crick knew before they began their work.

DNA is a linear, unbranched polymer in which the monomeric subunits are four chemically distinct nucleotides that can be linked together in any order in chains that are hundreds, thousands, or even millions of units in length. Each nucleotide in a DNA polymer is made up of three components (**Figure 1.4**):

- **2'-Deoxyribose**, which is a **pentose**, a type of sugar composed of five carbon atoms. These five carbons are numbered 1' (spoken as one-prime), 2', and so on. The name 2'-deoxyribose indicates that this particular sugar is a derivative of ribose, in which the hydroxyl (-OH) group attached to the 2'-carbon of ribose has been replaced by a hydrogen (-H) group.
- A **nitrogenous base**, one of **cytosine** or **thymine** (single-ring **pyrimidines**) or **adenine** or **guanine** (double-ring **purines**). The base is attached to the 1'-carbon of the sugar by a  **$\beta$ -N-glycosidic bond** attached to nitrogen number one of the pyrimidine or number nine of the purine.
- A **phosphate group**, comprising one, two, or three linked phosphate units attached to the 5'-carbon of the sugar. The phosphates are designated  $\alpha$ ,  $\beta$ , and  $\gamma$ , with the  $\alpha$ -phosphate being the one directly attached to the sugar.

A molecule made up of just the sugar and base is called a **nucleoside**; addition of the phosphates converts this to a nucleotide. Although cells contain nucleotides with one, two, or three phosphate groups, only the nucleoside triphosphates act as substrates for DNA synthesis. The full chemical names of the four nucleotides that polymerize to make DNA are

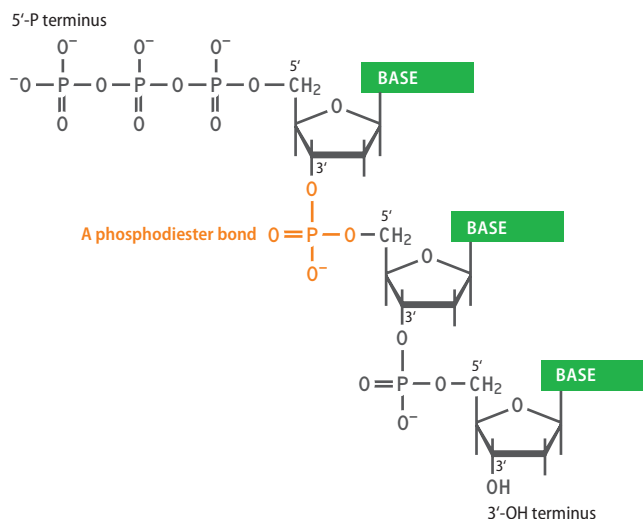
- 2'-deoxyadenosine 5'-triphosphate
- 2'-deoxycytidine 5'-triphosphate
- 2'-deoxyguanosine 5'-triphosphate
- 2'-deoxythymidine 5'-triphosphate

The abbreviations of these four nucleotides are dATP, dCTP, dGTP, and dTTP, respectively, or when referring to a DNA sequence, A, C, G, and T, respectively.

In a polynucleotide, individual nucleotides are linked together by **phosphodiester bonds** between their 5'- and 3'-carbons (**Figure 1.5**). From the structure



**Figure 1.5** A short DNA polynucleotide showing the structure of the phosphodiester bond. Note that the two ends of the polynucleotide are chemically distinct.

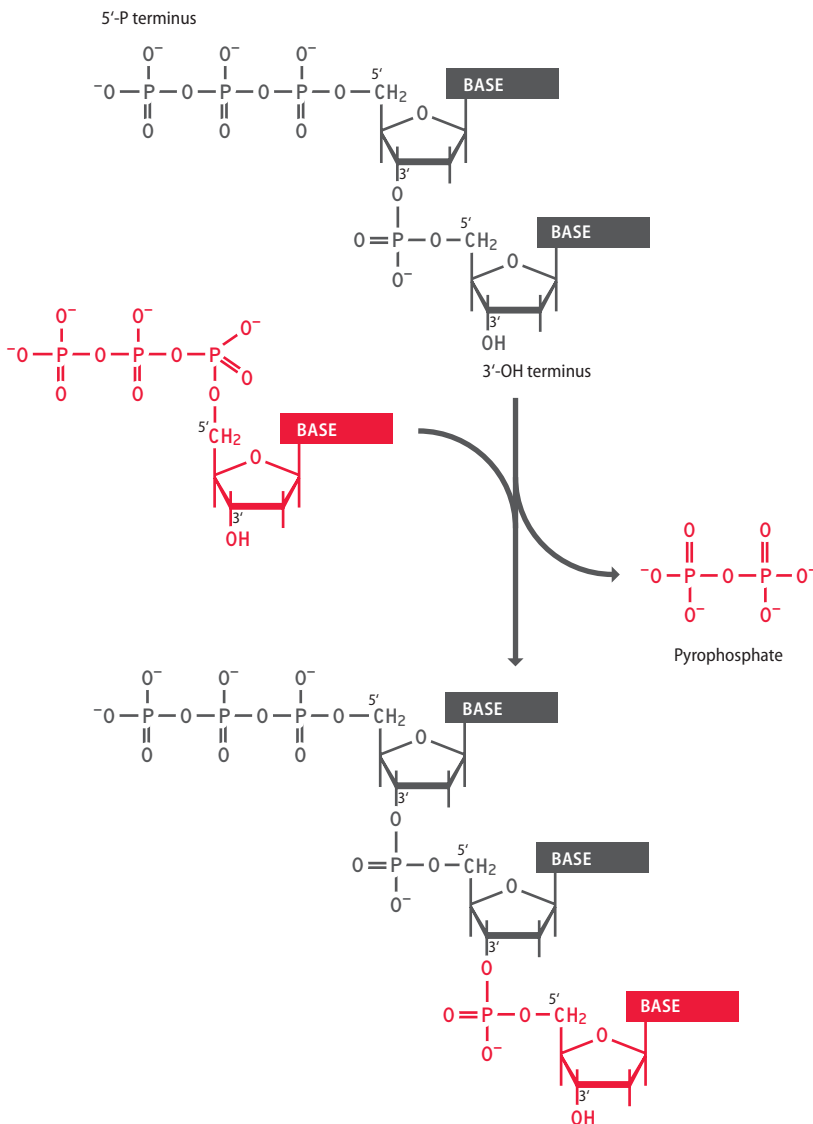


of this linkage, we can see that the polymerization reaction (**Figure 1.6**) involves removal of the two outer phosphates (the  $\beta$ - and  $\gamma$ -phosphates) from one nucleotide and replacement of the hydroxyl group attached to the 3'-carbon of the second nucleotide. Note that the two ends of the polynucleotide are chemically distinct, one having an unreacted triphosphate group attached to the 5'-carbon (the **5'**- or **5'-P terminus**) and the other having an unreacted hydroxyl attached to the 3'-carbon (the **3'**- or **3'-OH terminus**). This means that the polynucleotide has a chemical direction, expressed as either  $5' \rightarrow 3'$  (down in **Figure 1.5**) or  $3' \rightarrow 5'$  (up in **Figure 1.5**). An important consequence of the polarity of the phosphodiester bond is that the chemical reaction needed to extend a DNA polymer in the  $5' \rightarrow 3'$  direction is different from that needed to make a  $3' \rightarrow 5'$  extension. The **DNA polymerase** enzymes present in living organisms are only able to carry out  $5' \rightarrow 3'$  synthesis, which adds significant complications to the process by which double-stranded DNA is replicated (**Section 15.3**).

In the years before 1950, various lines of evidence had shown that cellular DNA molecules are composed of two or more polynucleotides assembled together in some way. The possibility that unraveling the nature of this assembly might provide insights into how genes work prompted Watson and Crick, among others, to try to solve the structure. According to Watson in his book *The Double Helix*, their work was a desperate race against the famous American biochemist Linus Pauling, who initially proposed an incorrect triple-helix model, giving Watson and Crick the time they needed to complete the double-helix structure. It is now difficult to separate fact from fiction, especially regarding the part played by Rosalind Franklin, whose **X-ray diffraction studies** provided the bulk of the experimental data in support of the double helix and who was herself very close to solving the structure. The one thing that is clear is that the double helix, discovered by Watson and Crick on Saturday, March 7, 1953, was the single most important breakthrough in biology during the twentieth century.

The discovery of the double helix can be looked on as one of the first multidisciplinary biological research projects. Watson and Crick used four quite different types of information to deduce the double-helix structure:

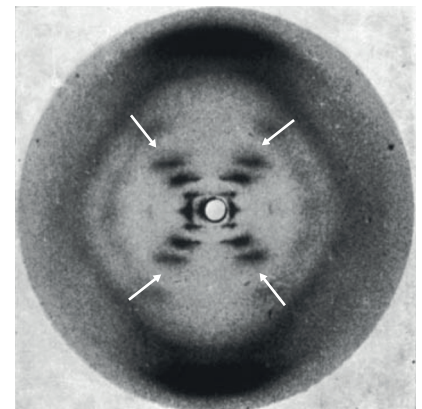
- Biophysical data of various kinds were used to infer some of the key features of the structure. The water content of DNA fibers was particularly important because it enabled the density of the DNA in a fiber to be estimated. The number of strands in the helix and the spacing between the nucleotides had to be compatible with the fiber density. Pauling's triple-helix model was based on an incorrect density measurement that suggested that the DNA molecule was more closely packed than is actually the case.

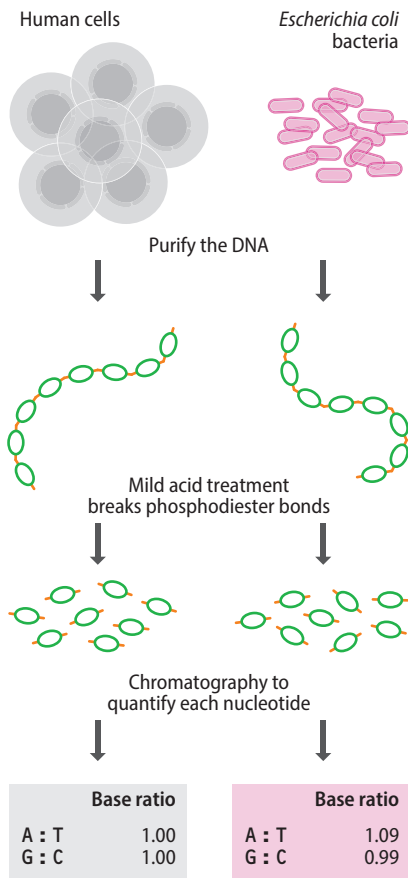


**Figure 1.6** The polymerization reaction that results in synthesis of a DNA polynucleotide. Synthesis occurs in the 5' → 3' direction, with the new nucleotide being added to the 3'-carbon at the end of the existing polynucleotide. The β- and γ-phosphates of the nucleotide are removed as a pyrophosphate molecule.

- **X-ray diffraction patterns** (Section 11.1), most of which were produced by Rosalind Franklin, revealed the detailed helical structure (Figure 1.7).
- The **base ratios**, which had been discovered by Erwin Chargaff of Columbia University in New York, enabled the pairing between the polynucleotides in the helix to be deduced. Chargaff had carried out a lengthy series of chromatographic studies of DNA samples from various sources and showed

**Figure 1.7** Franklin's photo 51 showing the X-ray diffraction pattern obtained with a fiber of DNA. The cross shape indicates that DNA has a helical structure, and the extent of the shadowing within the diamond spaces above, below, and to either side of the cross show that the sugar-phosphate backbone is on the outside of the helix (see Figure 1.9). The positions of the various smears that make up the arms of the cross enable dimensions such as the diameter, rise per base pair, and pitch (see Table 1.1) of the molecule to be calculated. The missing smears (the gap in each arm of the cross, marked by the arrows) indicate the relative positioning of the two polynucleotides. These missing smears enabled Watson and Crick to recognize that there are two grooves of different depths on the outer surface of the helix (see Figure 1.9). (From Franklin R & Gosling RG [1953] *Nature* 171:740–741. With permission from Macmillan Publishers Ltd.)





**Figure 1.8 The base ratio experiments performed by Chargaff.** DNA was extracted from various organisms and treated with acid to hydrolyze the phosphodiester bonds and release the individual nucleotides. Each nucleotide was then quantified by chromatography. The data show some of the actual results obtained by Chargaff. These indicate that, within experimental error, the amount of adenine is the same as that of thymine and the amount of guanine is the same as that of cytosine.

that although the values are different in different organisms, the amount of adenine is always the same as the amount of thymine and the amount of guanine equals the amount of cytosine (Figure 1.8). These base ratios led to the base-pairing rules, which were the key to the discovery of the double-helix structure.

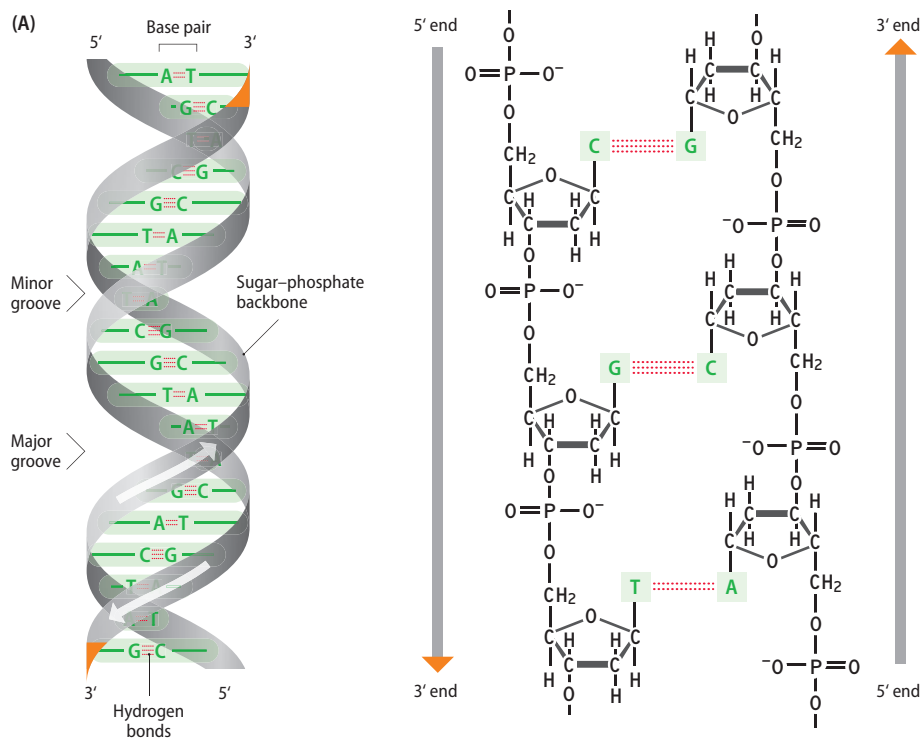
- The construction of scale models of possible DNA structures, which was the only major technique that Watson and Crick performed themselves, enabled the relative positioning of the various atoms to be checked, to ensure that pairs that formed bonds were not too far apart and that other atoms were not so close together as to interfere with one another.

### The double helix is stabilized by base pairing and base stacking

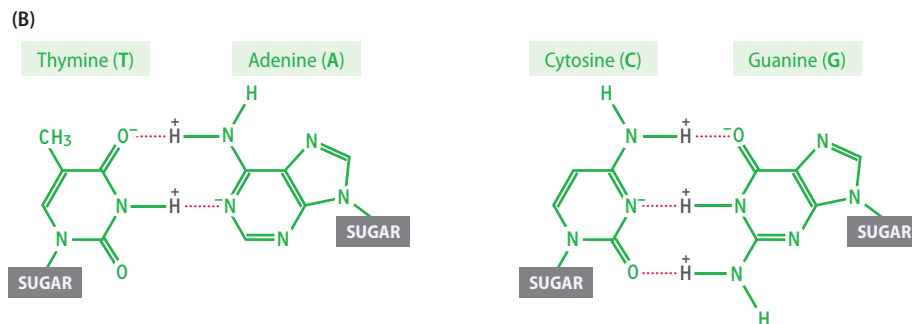
The double helix is right-handed, which means that if it were a spiral staircase and you were climbing upward, then the rail on the outside of the staircase would be on your right-hand side. The two strands run in opposite directions (Figure 1.9A). The helix is stabilized by two types of chemical interaction:

- **Base pairing** between the two strands involves the formation of **hydrogen bonds** between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine (Figure 1.9B). Hydrogen bonds are weak **electrostatic interactions** between an electronegative atom (such as oxygen or nitrogen) and a hydrogen atom attached to a second electronegative atom. Hydrogen bonds are longer than covalent bonds and are much weaker; typical bond energies are 8–29 kJ mol<sup>-1</sup> at 25°C, compared with up to 348 kJ mol<sup>-1</sup> for a single covalent bond between a pair of carbon atoms. As well as their role in the DNA double helix, hydrogen bonds stabilize protein secondary structures. The two base-pair combinations—A base-paired with T and G base-paired with C—explain the base ratios discovered by Chargaff. These are the only pairs that are permissible, partly because of the geometries of the nucleotide bases and the relative positions of the atoms that are able to participate in hydrogen bonds, and partly because the pair must be between a purine and a pyrimidine: a purine–purine pair would be too big to fit within the helix, and a pyrimidine–pyrimidine pair would be too small.
- **Base stacking** involves attractive forces between adjacent base pairs and adds stability to the double helix once the strands have been brought together by base pairing. Base stacking is sometimes called  **$\pi$ - $\pi$  interactions**, because it is thought to involve the p electrons associated with the double bonds of the purine and pyrimidine structures. However, this hypothesis is now being questioned, and the possibility that base stacking involves a type of electrostatic interaction is being explored.

Both base pairing and base stacking are important in holding the two polynucleotides together, but base pairing has added significance because of its biological implications. The limitation that A can only base-pair with T and G can only base-pair with C means that **DNA replication** can result in perfect copies of a parent molecule through the simple expedient of using the sequences of the preexisting strands to dictate the sequences of the new strands. This is **template-dependent DNA synthesis**, the system used by all cellular DNA polymerases



**Figure 1.9 Double-helix structure of DNA.** (A) Two representations of the double helix. On the left, the structure is shown with the sugar–phosphate backbone of each polynucleotide drawn as a gray ribbon with the base pairs in green. On the right, the chemical structure for three base pairs is given. (B) A base-pairs with T, and G base-pairs with C. The bases are drawn in outline, with the hydrogen bonding indicated by dotted lines. Note that a G–C base pair has three hydrogen bonds whereas an A–T base pair has just two.



(Section 2.1). Base pairing therefore enables DNA molecules to be replicated by a system that is so simple and elegant that as soon as the double-helix structure was publicized by Watson and Crick, every biologist became convinced that genes really are made of DNA.

### The double helix has structural flexibility

The double helix described by Watson and Crick, and shown in Figure 1.9A, is called the B-form of DNA or **B-DNA**. Its characteristic features lie in its dimensions: a helical diameter of 2.37 nm, a rise of 0.34 nm per base pair, and a pitch (the distance taken up by a complete turn of the helix) of 3.4 nm, corresponding to 10 base pairs (bp) per turn. The DNA in living cells is thought to be predominantly in this B-form, but it is now clear that genomic DNA molecules are not entirely uniform in structure. This is mainly because each nucleotide in the helix has the flexibility to take up a slightly different molecular shape. To adopt these different conformations, the relative positions of the atoms in the nucleotide must change slightly. There are a number of possibilities but the most important conformational changes are as follows:

- Rotation around the  $\beta$ -N-glycosidic bond changes the orientation of the base relative to the sugar: the two possibilities are called the *anti*- and *syn*-conformations (Figure 1.10A). Base rotation influences the positioning of the two polynucleotides.