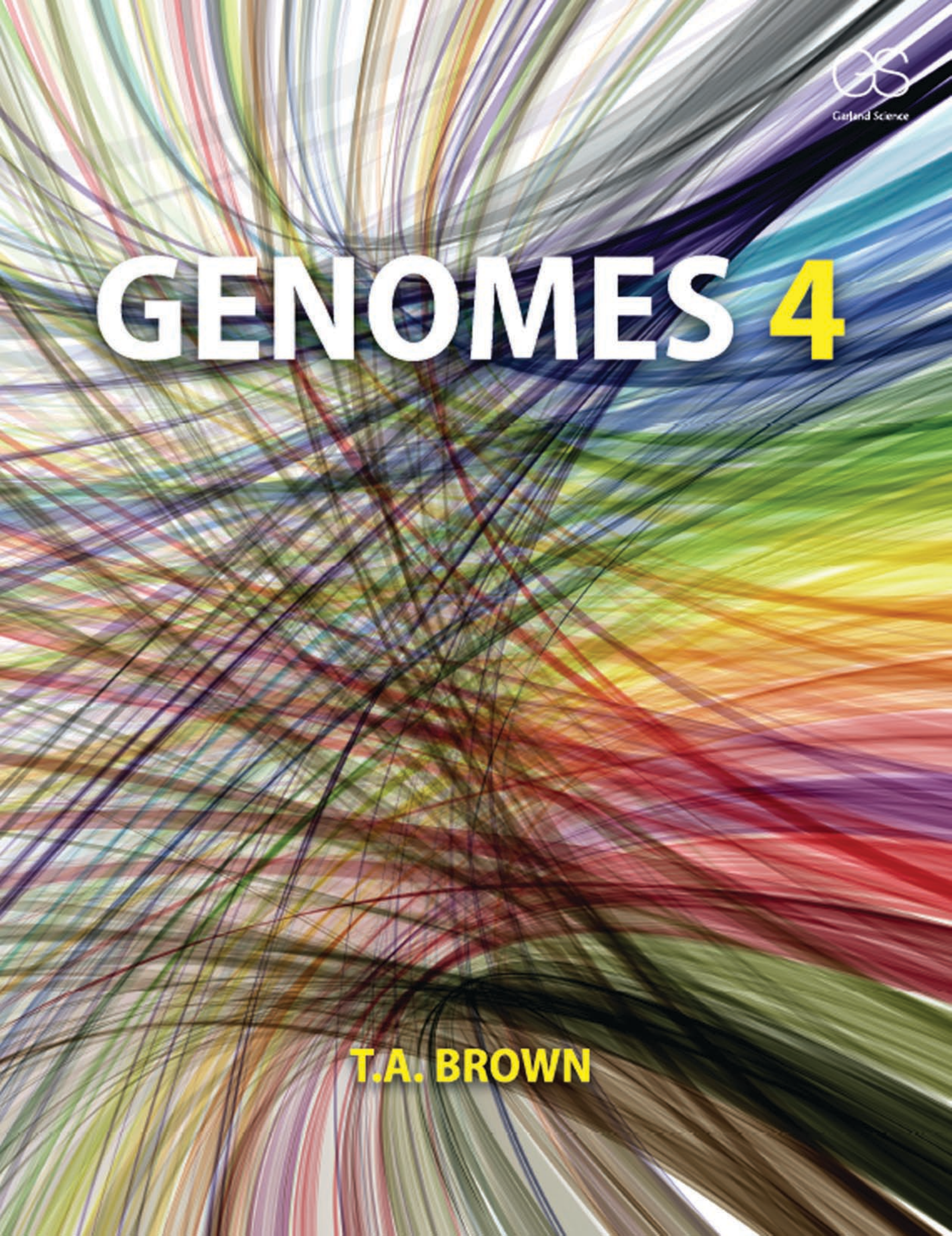
Garland Science

GENOMES 4



T.A. BROWN

GENOMES 4

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

GENOMES 4

T. A. BROWN

Garland Science
Taylor & Francis Group
NEW YORK AND LONDON

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.

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.

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

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).

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://lccn.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

GS Garland Science
Taylor & Francis Group

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

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 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

CONTENTS

Taylor & Francis
Taylor & Francis Group
http://taylorandfrancis.com

PART I

STUDYING GENOMES

CHAPTER 1

GENOMES, TRANSCRIPTOMES, AND PROTEOMES

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**.

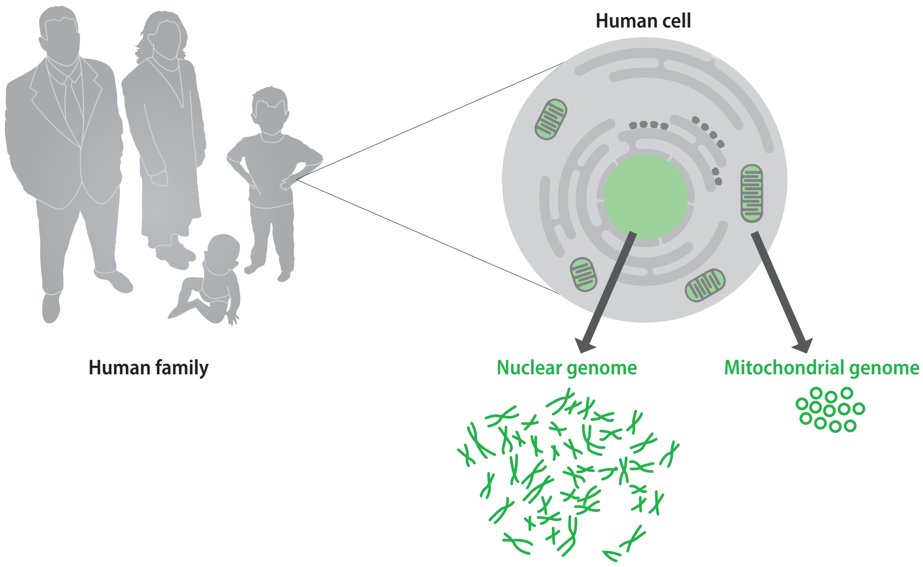


Figure 1.1 Nuclear and mitochondrial components of the human genome.

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.

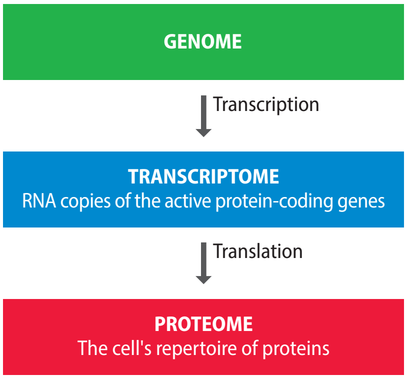

GENOME
Transcription
TRANSCRIPTOME
RNA copies of the active protein-coding genes
Translation
PROTEOME
The cell's repertoire of proteins

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

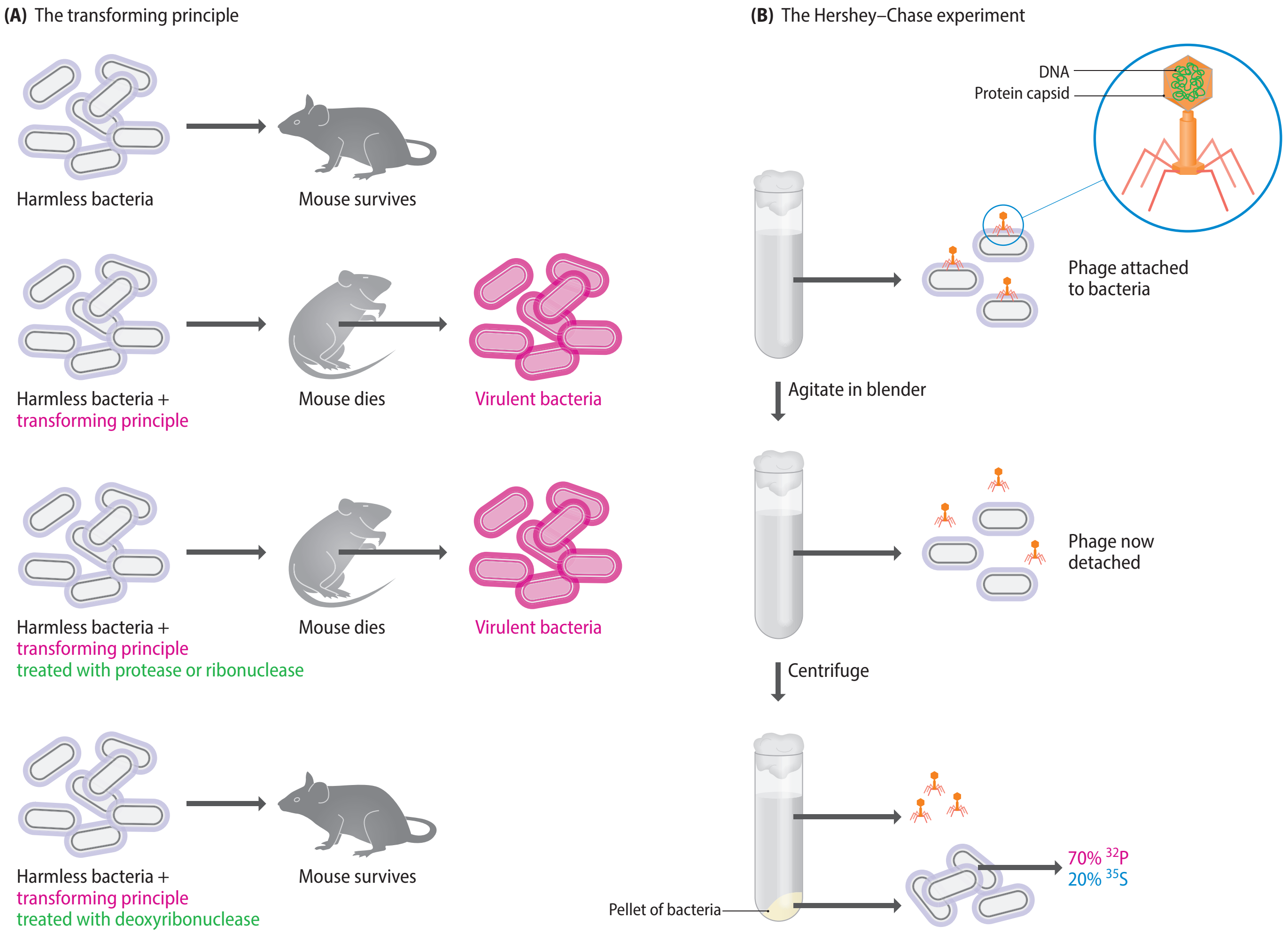


Figure 1.3 The two experiments that suggested that genes are made of DNA. (A) Avery and colleagues showed that the transforming principle is made of DNA. The top two panels show what happens when mice are injected with harmless *Streptococcus pneumoniae* bacteria with or without addition of the transforming 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}P, and the protein was labeled with ^{35}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}P-labeled component of the phages (the DNA) but only 20% of the ^{35}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

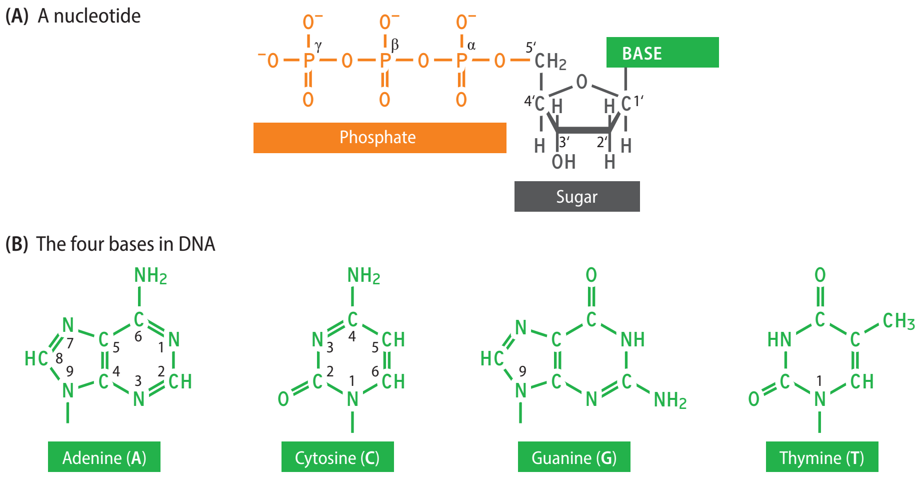


Figure 1.4 Structure of a nucleotide. (A) General structure of a deoxyribonucleotide, which is the type of nucleotide found in DNA. (B) The four bases that occur in deoxyribonucleotides.

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 **β-*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 α, β, and γ, with the α-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

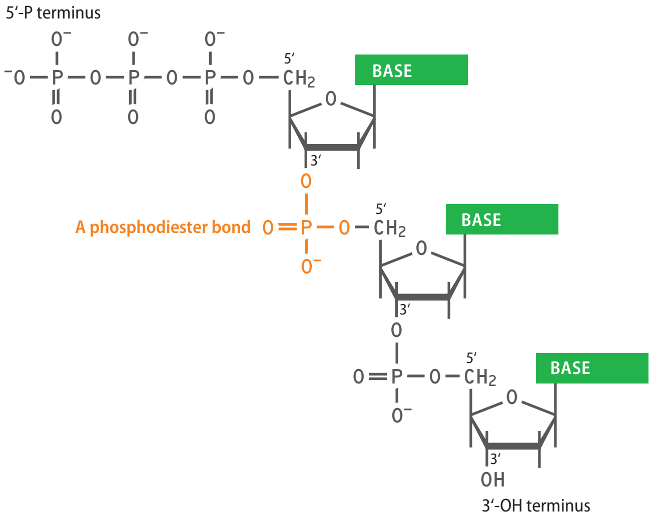


Figure 1.5 A short DNA polynucleotide showing the structure of the phosphodiester bond. Note that the two ends of the polynucleotide are chemically distinct.

of this linkage, we can see that the polymerization reaction (**Figure 1.6**) involves removal of the two outer phosphates (the β- and γ-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′ → 3′ (down in Figure 1.5) or 3′ → 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′ → 3′ direction is different from that needed to make a 3′ → 5′ extension. The **DNA polymerase** enzymes present in living organisms are only able to carry out 5′ → 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.

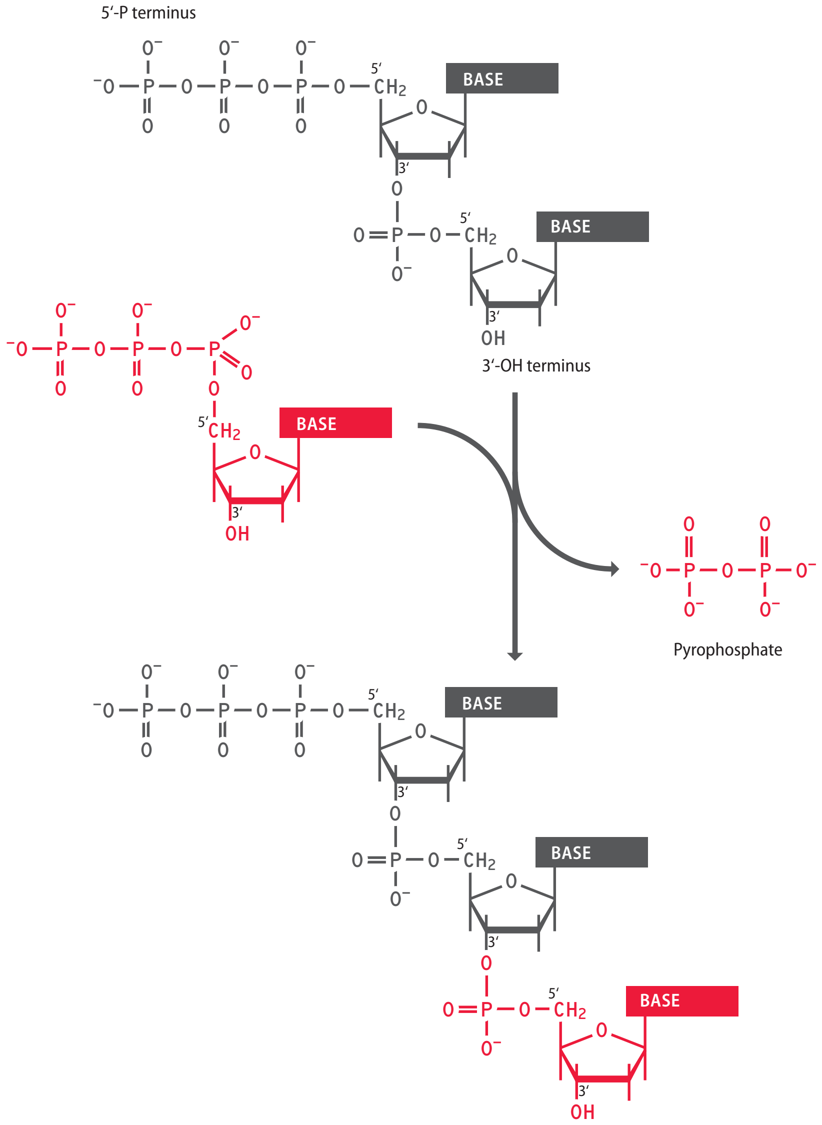


Figure 1.6 The polymerization reaction that results in synthesis of a DNA polynucleotide. Synthesis occurs in the 5′ ⟶ 3′ direction, with the new nucleotide being added to the 3′-carbon at the end of the existing polynucleotide. The β- and γ-phosphates of the nucleotide are removed as a pyrophosphate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

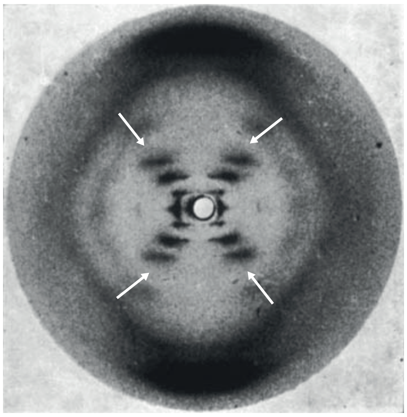


Figure 1.7 Franklin's photo 51 showing the X-ray diffraction pattern obtained with a fiber of DNA. The cross shape indicates that DNA has a helical structure, and the extent of the shadowing within the diamond spaces above, below, and to either side of the cross show that the sugar–phosphate backbone 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.)

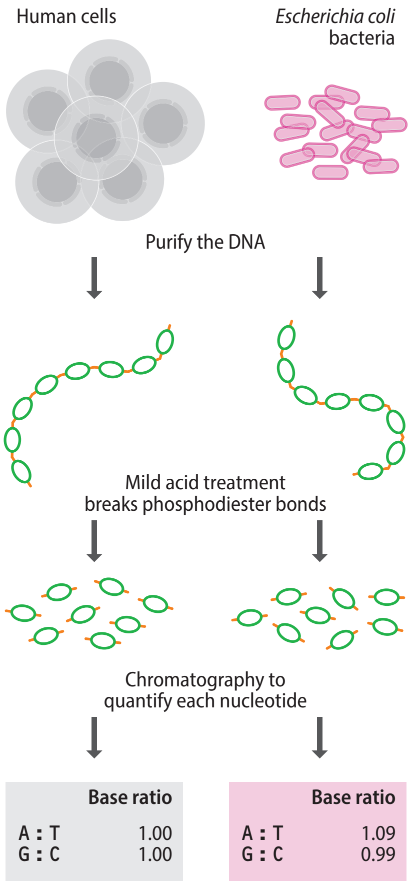

Human cells
Escherichia coli bacteria
Purify the DNA
Mild acid treatment breaks phosphodiester bonds
Chromatography to quantify each nucleotide
Base ratio
A : T 1.00
G : C 1.00
Base ratio
A : T 1.09
G : C 0.99

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^{-1} at 25°C, compared with up to 348 kJ mol^{-1} 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 **π–π 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

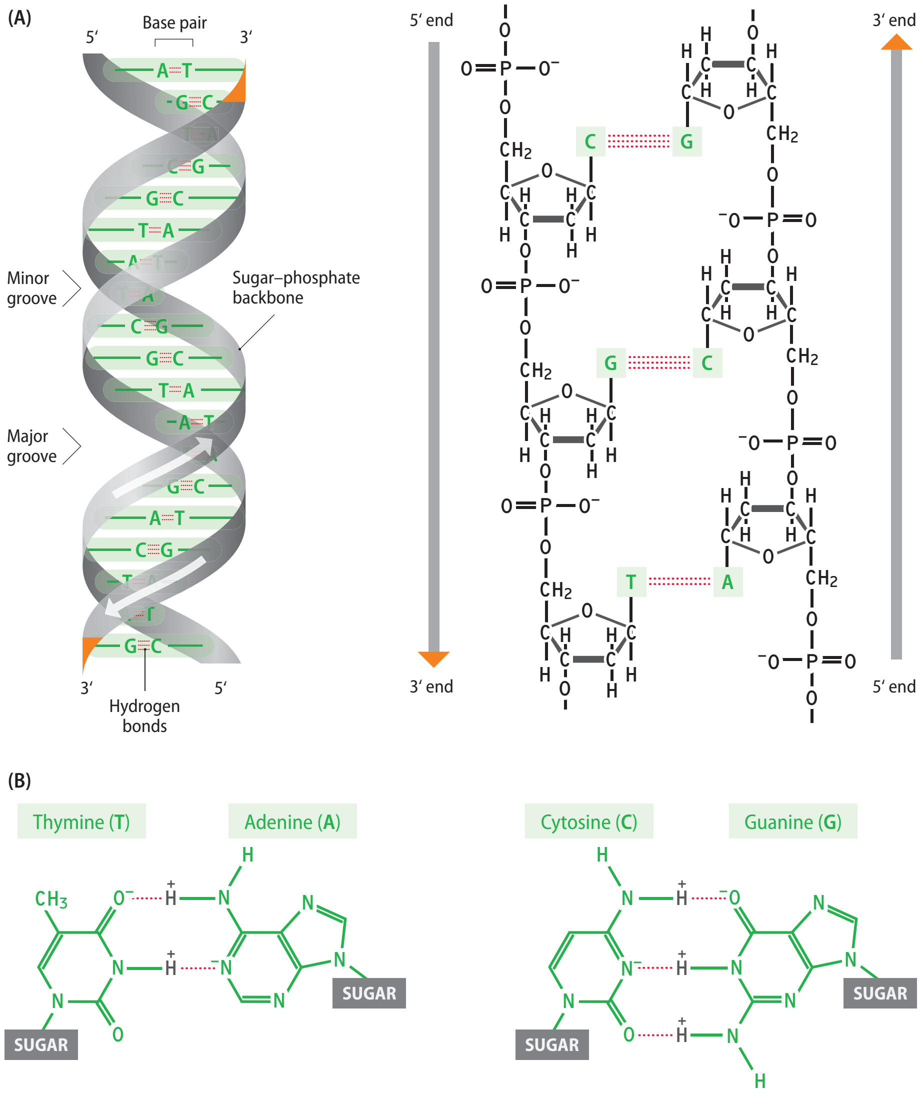


Figure 1.9 Double-helix structure of DNA. (A) Two representations of the double helix. On the left, the structure is shown with the sugar–phosphate backbone of each polynucleotide drawn as a gray ribbon with the base pairs in green. On the right, the chemical structure for three base pairs is given. (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 β-*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.