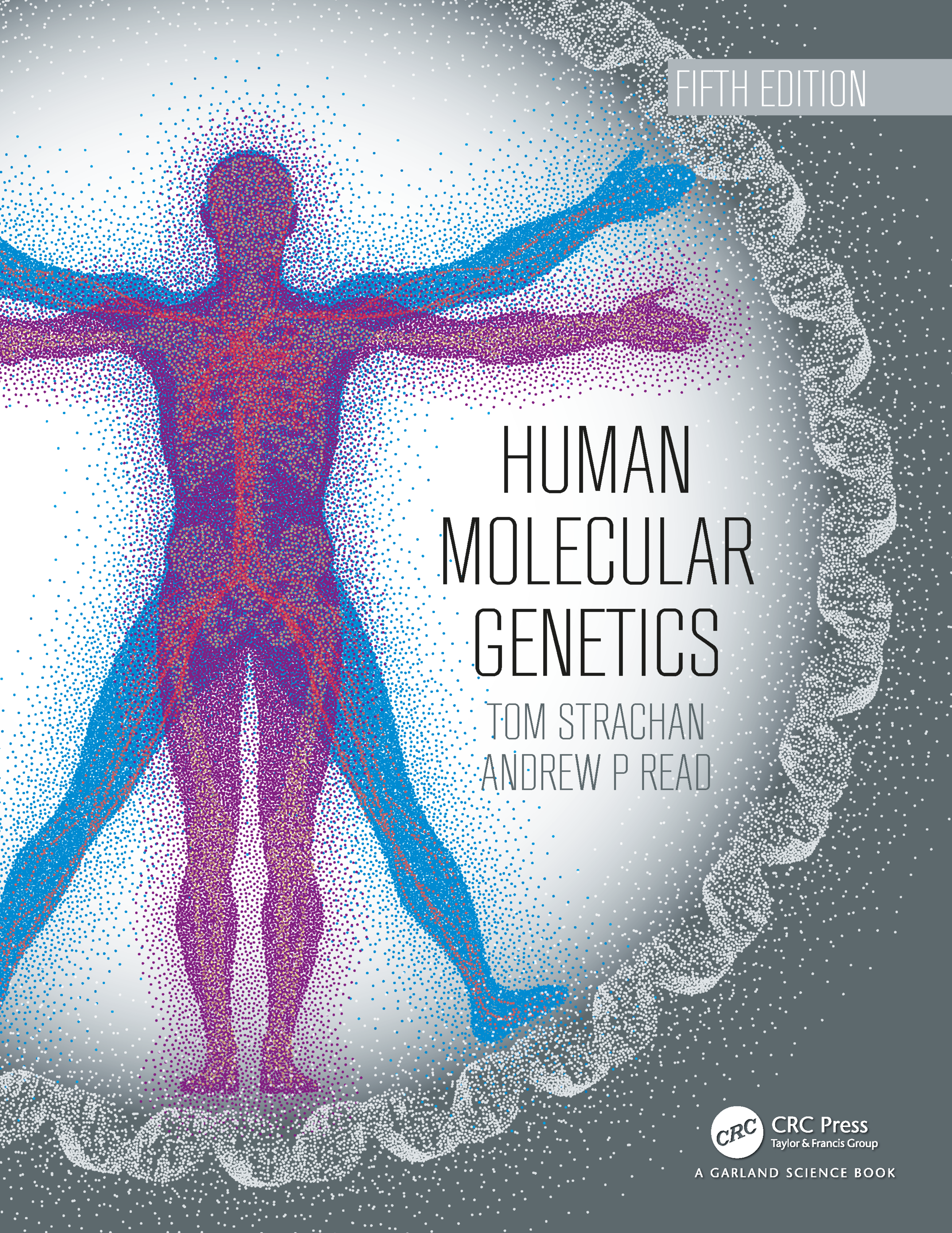

FIFTH EDITION

HUMAN MOLECULAR GENETICS

TOM STRACHAN
ANDREW P READ

CRC Press
Taylor & Francis Group
A GARLAND SCIENCE BOOK

FIFTH EDITION

HUMAN MOLECULAR GENETICS

FIFTH EDITION

HUMAN MOLECULAR GENETICS

TOM STRACHAN
BSc, PhD, FRSE, FMedSci
Emeritus Professor of Human Molecular Genetics
Newcastle University, Newcastle, UK

ANDREW P READ
MA, PhD, FRCPath, FMedSci
Emeritus Professor of Human Genetics
University of Manchester, Manchester, UK

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A GARLAND SCIENCE BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Printed on acid-free paper

International Standard Book Number-13: 978-0-815-34589-3 (Paperback)
978-0-367-00250-3 (Hardback)

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Contents

PART 5 APPLIED HUMAN MOLECULAR GENETICS

Preface

Much has changed since the fourth edition of *Human Molecular Genetics* (HMG4) appeared in 2011, so this fifth edition has seen a comprehensive rewrite and reorganization. Few of the chapters retain their identity from HMG4, but our aims throughout the book remain the same: to provide a framework of principles rather than to list facts (which are better found in online resources), to provide a bridge between basic textbooks and the research literature, and to communicate our excitement and enthusiasm for this very fast-moving area of science.

The biggest single development since HMG4 has been the massive expansion of DNA sequencing in every area of human genetics. In response, we have provided a much-extended and updated coverage of massively-parallel sequencing technology, including the exciting new field of single-cell genomics. In some respects, the sequencing revolution has made things simpler. Many specific techniques covered in HMG4 have been largely or completely superseded by sequencing. The reader will note, however, that we still illustrate karyotypes. We make no apology for showing these where appropriate because they have educational value—it is often easier to understand what is going on by looking at a karyotype rather than sequence data, even though nowadays most laboratories would use sequencing rather than microscopy for these purposes.

In the preface to HMG4 we wrote, "we can confidently expect that the genomes of huge numbers of organisms and individuals will have been completed before the next edition of this book," and that expectation has been amply realized. Human genetics is now firmly in the world of Big Data and big international collaborations—and our coverage reflects this.

Among new or rearranged chapters we were particularly gratified when Mark Jobling and the team who produced the excellent *Human Evolutionary Genetics* (2nd edn, Garland Science, 2013) agreed to contribute a chapter on human evolution. Analysis of contemporary and ancient DNA has progressed enormously in the past few years and is revealing fascinating insights into our origins and history—but neither of us felt qualified to write with sufficient authority on this important topic.

Other developments include:

- A radical revision of coverage of early mammalian development and stem cells, with a detailed explanation of the origins of cellular differentiation and an in-depth survey of pluripotent stem cells as well as tissue stem cells and cell reprogramming.
- A specific chapter on genetic manipulation of mammalian cells, bringing together material from different HMG4 chapters and tracing the evolution of genome editing from a focus on simply using homologous recombination to the modern emphasis on using programmable nucleases.
- A chapter that deals with both the architecture of the human genome and also the ENCODE Project and other new initiatives to understand how our genome functions.
- A chapter giving unified coverage of gene regulation and epigenetics.
- A chapter giving an overview of human genetic variation that includes the origins of DNA sequence variation, DNA repair mechanisms, variant classes, population genomics, and functional genetic variation.

- A new chapter on human population genetics—a topic that we felt received inadequate coverage in previous editions.
- A specific chapter on molecular pathology, bringing together and expanding material from HMG4.
- Greatly expanded discussion of the achievements and limitations of genome-wide association studies (GWAS) in identifying susceptibility factors for common complex conditions. As the GWAS era is giving way to large-scale sequencing approaches, this seems a particularly apposite time for a critical analysis.
- New coverage of DNA diagnostics reflecting the major changes that have come with the routine use of whole exome and whole genome sequencing.
- Revised discussion of cancer genetics and genomics reflecting developments in multiplatform analyses, liquid biopsies, and targeted treatments.
- A new chapter that brings together model organisms and disease modeling, including the fast-moving new field of cellular disease modeling, especially organoid models that arose from basic developmental studies.

Apart from these specific topics, every page of the text has been revised and updated to provide an overview of human molecular genetics in 2018.

This book has only been possible because of the work of the team under Joanna Koster at Taylor & Francis who have converted our drafts and sketches into the finished product—Paul Bennett, Jordan Wearing, Matt McClements, Ruth Maxwell, Becky Hainz-Baxter, and probably others who have worked from time to time on the project. As ever, we are deeply grateful to our wives, Meryl and Gilly, for their forbearance and support during the long gestation of this book.

Tom Strachan
Andrew P Read

About the authors

Tom Strachan is Emeritus Professor of Human Molecular Genetics at Newcastle University, Newcastle, UK, and is a Fellow of the Royal Society of Edinburgh and a Fellow of the Academy of Medical Sciences. He was the founding Head of Institute at Newcastle University's Institute of Human Genetics (now the Institute of Genetic Medicine) and its Scientific Director from 2001 to 2009. Tom's early research interests were in multigene family evolution and interlocus sequence exchange, notably in the HLA and 21-hydroxylase gene clusters. While pursuing the latter, he became interested in medical genetics. His most recent research has focused on certain developmental disorders and developmental control genes.

Andrew Read is Emeritus Professor of Human Genetics at Manchester University, Manchester, UK, and a Fellow of the Academy of Medical Sciences. Andrew has been particularly concerned with making the benefits of DNA technology available to people with genetic problems. He established one of the first DNA diagnostic laboratories in the UK over 35 years ago (which became one of two National Genetics Reference Laboratories), and was founder chairman of the British Society for Human Genetics (now the British Society for Genetic Medicine), the main UK professional body in this area. His own research is on the molecular pathology of various hereditary syndromes, especially hereditary hearing loss.

Tom Strachan and **Andrew Read** were recipients of the European Society of Human Genetics Education Award in 2007.

Contributors

The authors are grateful to the following who contributed Chapter 14, *Human evolution*, in this edition:

Mark A Jobling DPhil

Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

Edward J Hollox PhD

Department of Genetics and Genome Biology, University of Leicester, Leicester, UK

Toomas Kivisild PhD

Department of Human Genetics, KU Leuven, Leuven, Belgium

Chris Tyler-Smith PhD

Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge, UK

BASICS OF DNA, CHROMOSOMES, CELLS, DEVELOPMENT AND INHERITANCE

PART ONE

1

Basic principles of nucleic acid structure and gene expression

Molecular genetics is largely defined by the interplay between three classes of macromolecule: the nucleic acid molecules, DNA (deoxyribonucleic acid) and RNA (ribonucleic acid), and proteins. In organisms and cells, DNA is the genetic (hereditary) material that is transmitted to daughter cells when cells replicate, and from one generation to the next when organisms reproduce. Viruses also have genetic material that is transmitted to viral progeny; according to the type of virus, the genetic material may be DNA or RNA. The term **genome** is the collective name for the set of different DNA molecules in an organism, cell, or DNA virus, or of RNA molecules in an RNA virus. All proteins have a polypeptide core that is synthesized using genetic information within DNA molecules (or within the hereditary RNA molecules of an RNA virus).

RNA may have been the hereditary material at a very early stage of evolution, but now, except in certain viruses, it no longer serves this role. Instead, the genetic information in cells came to be stored in DNA molecules (which are more chemically stable than RNA and can be copied accurately and transmitted to daughter cells, and from one generation to the next). In eukaryotes, DNA molecules are found mainly in the chromosomes of the nucleus, but the mitochondria of all eukaryotic cells also have small DNA molecules, as do the chloroplasts of plant cells.

Genes are segments of hereditary DNA or RNA molecules that are used to make one or both of two types of functional end product: a polypeptide or a mature functional RNA. Both types of product are then subject to processing reactions to make a working molecule. For example, a polypeptide may be subject to cleavage and/or to minor chemical changes to its constituent components, and may often also be complexed with other molecules including carbohydrates, lipids, or other polypeptides in order to make a working protein.

In simple organisms, the DNA is packed with genes (bacteria typically have from several hundred up to a few thousand different genes packed within 1–10 Mb [megabases] of DNA). In the more complex cells of eukaryotes, the genes are usually much more sparsely distributed within the DNA, and in complex multicellular eukaroytes much of the DNA consists of highly-repetitive sequences (whose functions are often not so readily easily identified).

There are many different types of RNA molecule, but according to their function they can be divided into two broad classes. A **coding RNA** sequence, popularly called a **messenger RNA** (**mRNA**), carries genetic information from DNA to the protein synthesis machinery. Messenger RNA made in the nucleus needs to be exported to the cytoplasm to make proteins, but the messenger RNA synthesized in mitochondria and chloroplasts is used to make certain proteins within these organelles.

Mature **noncoding RNA** sequences are the second broad class of RNA. They are not used as a template to make polypeptides. Instead, they often assist the expression of other genes, sometimes acting in a fairly general way and sometimes by regulating the expression of a small set of target genes. Because most gene expression is ultimately dedicated to making polypeptides, either directly or by regulating how they are produced, proteins represent the major functional endpoint of the information stored in DNA.

The central dogma of molecular biology

Genetic information generally flows in a one-way direction: DNA is decoded to make RNA, and then coding RNA (messenger RNA) is used to make polypeptides that subsequently form proteins. Because of its universality, this flow of genetic information has been described as the central dogma of molecular biology. Two sequential processes are essential in all cellular organisms:

1. Transcription, by which a sequence of bases on a DNA strand is used as a template by an RNA polymerase to synthesize an RNA; the RNA product is processed to make a messenger RNA (coding RNA) or noncoding RNA;
2. Translation, by which a messenger RNA is decoded to make polypeptides at ribosomes, large RNA–protein complexes found in the cytoplasm and also in mitochondria and chloroplasts.

Genetic information is encoded in the linear sequence of nucleotides in DNA. That information is copied during transcription to specify a linear sequence of nucleotides in the RNA product. In the case of a coding RNA, groups of three nucleotides at a time (codons) are read in a linear sequence to specify a linear sequence of amino acids in the polypeptide product.

The central dogma is now recognized to be not strictly valid. A class of RNA virus known as retroviruses provided the first evidence. These viruses have an RNA genome with a gene that makes a reverse transcriptase, a DNA polymerase that uses an RNA template to make a DNA sequence copy. Thereafter, it became clear that cellular reverse transcriptases also exist. We now know that many DNA sequences in our cells specify reverse transcriptases to allow DNA copies to be made from different RNAs. This reverse flow of genetic information from RNA to DNA has been important in the evolution of our genome (as described in Chapter 13), and in replicating the DNA sequences at the very ends of linear chromosomes (described in Chapter 2).

1.1 COMPOSITION OF NUCLEIC ACIDS AND POLYPEPTIDES

We describe below the structure of nucleic acids and proteins. All proteins have a linear polypeptide backbone (encoded by a gene) to which carbohydrate, lipid, and small chemical groups may be added at the post-translational level. Here we describe the component units of nucleic acids and polypeptides, and the different types of chemical bonding within these macromolecules.

Nucleic acids and polypeptides are linear sequences of simple repeat units

DNA and RNA strands are large polymers that have very similar structures. Each has a linear sugar–phosphate backbone that has alternating residues of a five-carbon sugar and a phosphate, with a nitrogenous base attached to each sugar residue (**Figure 1.1A**). The sugars are ribose in RNA and deoxyribose in DNA, and they differ in either lacking or possessing an –OH group at their 2′-carbon positions (**Figure 1.1B**).

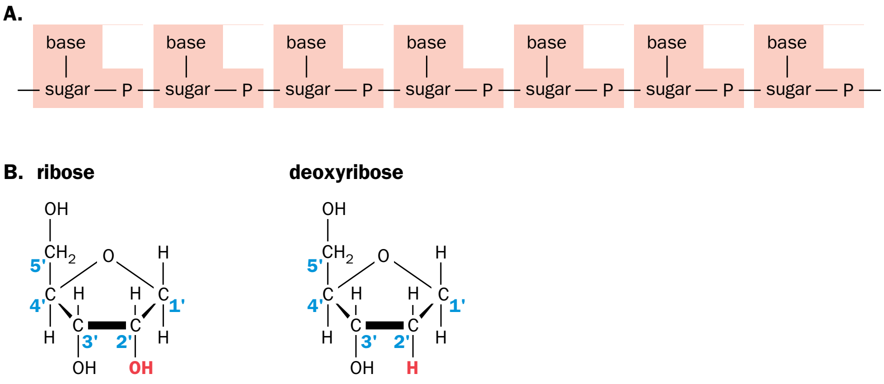

Figure 1.1 Repeat units in nucleic acids. (**A**) The linear backbone of nucleic acids consists of alternating phosphate (P) and sugar residues. Attached to each sugar is a base. The basic repeat unit (pale peach shading) consists of a base + sugar + phosphate = a nucleotide. (**B**) Ribose, the sugar in RNA, and deoxyribose, the sugar in DNA, both have five carbon atoms numbered 1′ to 5′. Deoxyribose lacks the hydroxyl (OH) group attached to carbon 2 of ribose (the proper name is 2′-deoxyribose).

Unlike the sugar and phosphate residues, the bases of a nucleic acid molecule vary, and it is the sequence of bases that identifies the nucleic acid and determines its function. The bases of a nucleic acid each consist of heterocyclic rings of carbon and nitrogen atoms and can be divided into two structural classes: **purines**, which have two interlocked rings, and **pyrimidines**, which have a single ring. In both DNA and RNA there are four principal types of base, two purines and two pyrimidines. Three types of base adenine (A), cytosine (C), and guanine (G) are common to both DNA and RNA. The fourth base is thymine (T) in DNA and the closely related uracil (U) in RNA. Uracil lacks the 5-methyl group found in thymine (**Figure 1.2A**).

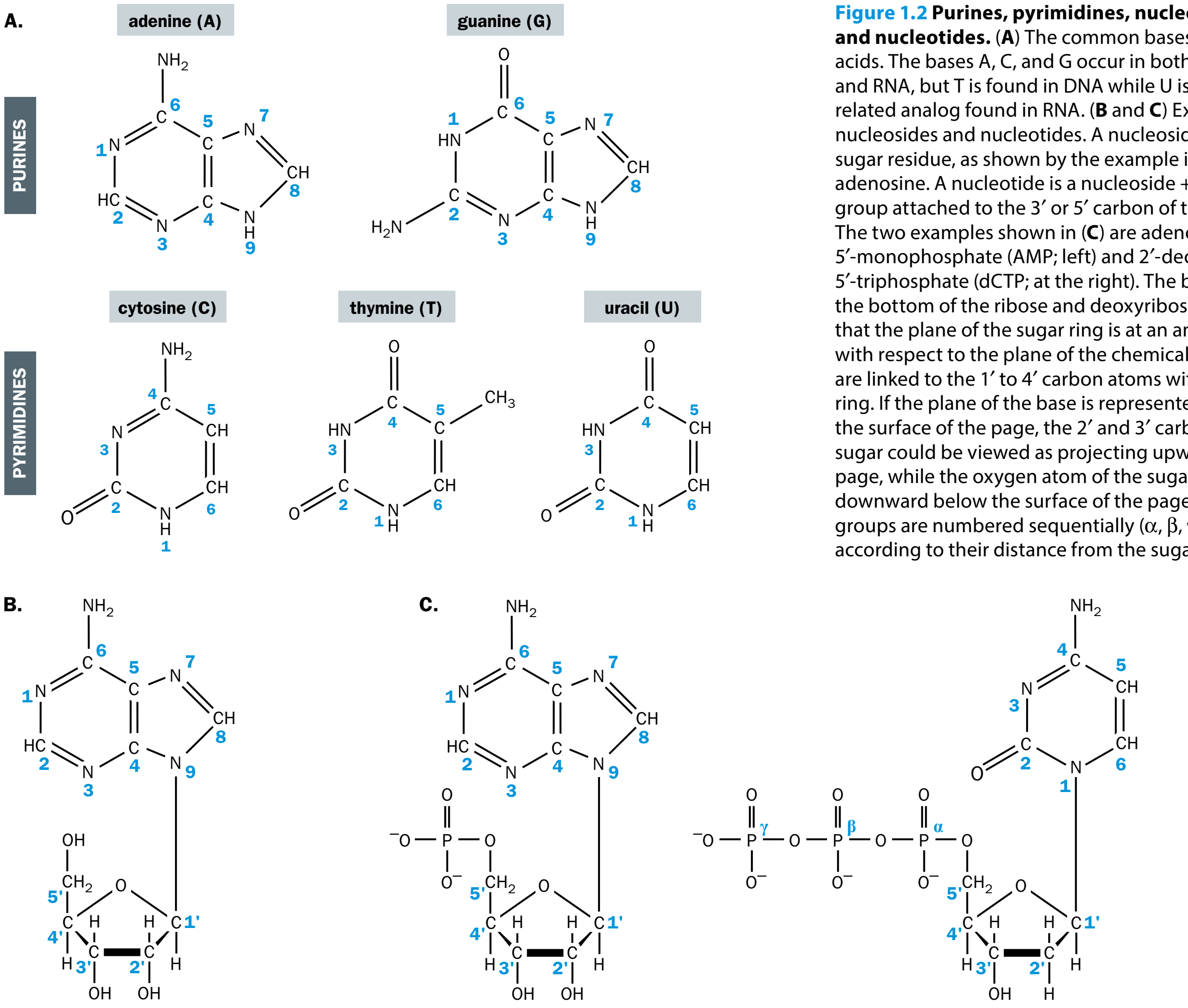

Figure 1.2 Purines, pyrimidines, nucleosides, and nucleotides. (**A**) The common bases in nucleic acids. The bases A, C, and G occur in both DNA and RNA, but T is found in DNA while U is a closely related analog found in RNA. (**B** and **C**) Examples of nucleosides and nucleotides. A nucleoside is a base + sugar residue, as shown by the example in (**B**), which is adenosine. A nucleotide is a nucleoside + phosphate group attached to the 3′ or 5′ carbon of the sugar. The two examples shown in (**C**) are adenosine 5′-monophosphate (AMP; left) and 2′-deoxycytidine 5′-triphosphate (dCTP; at the right). The bold lines at the bottom of the ribose and deoxyribose rings mean that the plane of the sugar ring is at an angle of 90° with respect to the plane of the chemical groups that are linked to the 1′ to 4′ carbon atoms within the ring. If the plane of the base is represented as lying on the surface of the page, the 2′ and 3′ carbons of the sugar could be viewed as projecting upward out of the page, while the oxygen atom of the sugar ring projects downward below the surface of the page. Phosphate groups are numbered sequentially (α, β, γ, etc.), according to their distance from the sugar ring.

In nucleic acids, each base is covalently attached to the sugar by an ***N*-glycosidic bond** that joins a nitrogen atom (nitrogen 1 of a pyrimidine or nitrogen 9 of a purine) to the carbon 1′ (one prime) of the sugar. A sugar with an attached base is called a nucleoside (**Figure 1.2B**). A nucleoside with a phosphate group attached at the 5′ or 3′ carbon of the sugar is the basic repeat unit of a DNA strand, and is called a **nucleotide** (**Figure 1.2C** and **Table 1.1**).

As described below, DNA also contains a few types of minor base produced by chemical modification, but base modification is much more common in RNA where a large variety of chemical modifications of both bases and ribose sugars are known to occur.

Polypeptides

Proteins are composed of one or more **polypeptide** chains that may be modified by the addition of carbohydrate side chains or other chemical groups. Like DNA and RNA, polypeptides are polymers that have a linear sequence of repeating units. The basic repeat unit is called an **amino acid**.

TABLE 1.1 NOMENCLATURE FOR BASES, NUCLEOSIDES, AND NUCLEOTIDES

		Nucleoside (= base + sugar)		Nucleotide (= nucleoside + phosphate)		
	Base	**Ribose**	**Deoxyribose**	**Monophosphate**	**Diphosphate**	**Triphosphate**
PURINE	Adenine	Adenosine	Deoxyadenosine	Adenosine monophosphate (AMP)[a,b] Deoxyadenosine monophosphate (dAMP)[a]	Adenosine diphosphate (ADP) Deoxyadenosine diphosphate (dADP)	Adenosine triphosphate (ATP) Deoxyadenosine triphosphate (dATP)
	Guanine	Guanosine	Deoxyguanosine	Guanosine monophosphate (GMP)[b] Deoxyguanosine monophosphate (dGMP)	Guanosine diphosphate (GDP) Deoxyguanosine diphosphate (dGDP)	Guanosine triphosphate (GTP) Deoxyguanosine triphosphate (dGTP)
PYRIMIDINE	Cytosine	Cytidine	Deoxycytidine	Cytidine monophosphate (CMP)[b] Deoxycytidine monophosphate (dCMP)	Cytidine diphosphate (CDP) Deoxycytidine diphosphate (dCDP)	Cytidine triphosphate (CTP) Deoxycytidine triphosphate (dCTP)
	Thymine	Thymidine	Deoxythymidine	Thymidine monophosphate (TMP)[b] Deoxythymidine monophosphate (dTMP)	Thymidine diphosphate (TDP) Deoxythymidine diphosphate (dTDP)	Thymidine triphosphate (TTP) Deoxythymidine triphosphate (dTTP)
	Uracil	Uridine	Deoxyuridine	Uridine monophosphate (UMP)[b] Deoxyuridine monophosphate (dUMP)	Uridine diphosphate (UDP) Deoxyuridine diphosphate (dUDP)	Uridine triphosphate (UTP) Deoxyuridine triphosphate (dUTP)

[a] Where the sugar is ribose, the nucleotide is AMP; where the sugar is deoxyribose, the nucleotide is dAMP. This pattern applies throughout the table. Note that TMP, TDP, and TTP are not normally found in cells.

[b] Nucleoside monophosphates are alternatively named as follows: AMP, adenylate; GMP, guanylate; CMP, cytidylate; TMP, thymidylate; UMP, uridylate.

Amino acids get their name because in its electrically neutral form a single unbound amino acid has an amino group (–NH_2) connected by a central α-carbon atom to a carboxyl group (–COOH). The central carbon atom also bears an identifying side chain that determines the chemical nature of the amino acid. At physiological pH, the amino group acquires a proton and becomes positively charged and the carboxyl group loses a proton and becomes negatively charged (**Figure 1.3A**). According to the type of amino acid, the side chain may or may not have a charge, as detailed below.

Polypeptides are formed by sequential condensation reactions between the amino group of one amino acid and the carboxyl group of the next amino acid to be incorporated into the polymer. As a result, a polypeptide has a repeating backbone where the amino acid residues are linked by amide groups (–CO–NH–) that are referred to as **peptide bonds** (**Figure 1.3B**), and where the side chain (generally called an R-group) can differ from one amino acid to another (**Figure 1.4**).

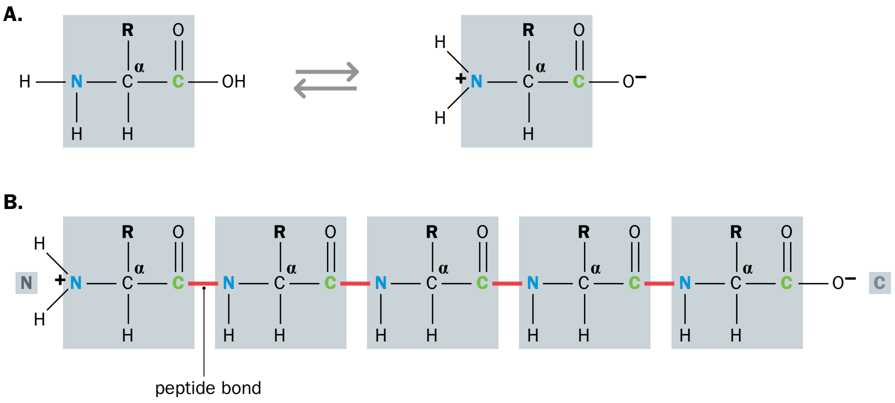

Figure 1.3 The general structure of an amino acid and a polypeptide. (**A**) Amino acid structure. At the left is the uncharged form of a generalized individual amino acid. A central α carbon is linked to three major groups: an amino group (NH_2), a carboxyl group (COOH), and a side chain R, giving the general formula $H_2N–CH(R)–COOH$. At physiological pH, as shown at the right, the end groups are ionized: the amino group acquires a positive charge and the carboxyl group acquires a negative charge. The gray shading shows an amino acid repeating unit as found in a polypeptide. (**B**) Polypeptide structure. A polypeptide forms by sequential addition of amino acid monomers in a condensation reaction involving the carboxyl group of the last amino acid to be incorporated and the amino group of the next amino acid to be incorporated. The amino acid monomers (highlighted by gray shading) are therefore connected by amide bonds (–CO–NH–), known in this context as peptide bonds. One end of the polypeptide backbone will retain the charged amino group of the original amino acid and is known as the N-terminal end; the other end has the charged carboxyl group of the last amino acid to be incorporated, and is the C-terminal end.

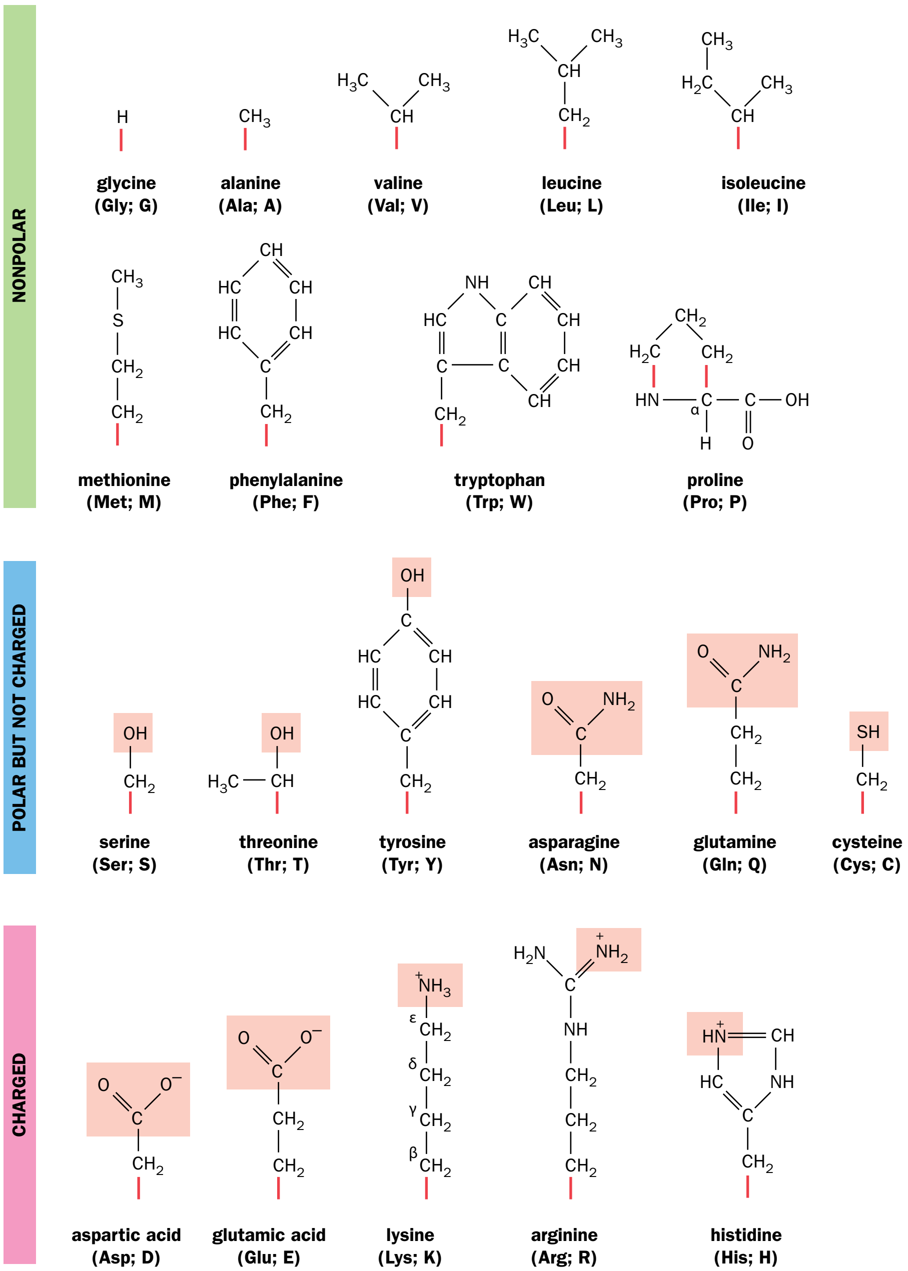

Figure 1.4 Side chains of the 20 common amino acids, grouped according to chemical class. In 19 of the 20 common amino acids the side chain is connected by a single covalent bond (red) to the α-carbon atom of the amino acid backbone; for these, we give the structure of the side chain only. Proline is the exception and we give its full structure here. Its side chain (-CH_2-CH_2-CH_2-) is connected to the backbone by two covalent bonds (red), with one end joined to the central α carbon atom, and the other end to the nitrogen atom of the backbone amino group. The convention for naming carbon atoms in a side chain is to use sequential Greek letters, counting out from the central α carbon atom (β, γ, δ, and so on; for example, in lysine's side chain, the carbon atom joined to the amino group, is the ε or epsilon carbon atom). Some amino acids have side chains with polar groups (pale peach shading) that may be uncharged or charged. The uncharged polar amino acids comprise three with a free hydroxyl group (serine, threonine, and tyrosine), two with amide groups (asparagine and glutamine), plus cysteine (which is only weakly polar). The charged amino acids comprise two acidic amino acids, aspartic acid (= aspartate) and glutamic acid (= glutamate), with a negatively-charged carboxyl ion on their side chain at physiological pH, plus three basic amino acids. The latter include two strongly basic amino acids, lysine and arginine, each with a positively-charged nitrogen atom in the side chain at physiological pH, plus the very weakly basic histidine. Note: at physiological pH histidines are predominantly neutral, but at low pH they can be positively charged (as shown here).

Twenty different amino acids are common in nature and can be classified into three main groups according to their side chains (see Figure 1.4). Nine amino acids have a nonpolar side chain. In most of these cases the side chain is a simple aliphatic group, but phenylalanine and tryptophan have aromatic side chains and proline has a very unusual side chain that connects the central carbon atom to the N-terminal amino group (see Figure 1.4). The nonpolar neutral amino acids are hydrophobic (repel water), often interacting with one another and with other hydrophobic groups.

Six amino acids are polar but electrically neutral overall. Their side chains carry polar groups with fractional electrical charges (often denoted as δ^+ or δ^-). Five amino acids have a charged side chain that either has a negative charge at physiological pH (acidic) or a net positive charge (basic, see Figure 1.4). In general, charged and uncharged polar amino acids are hydrophilic while nonpolar amino acids are hydrophobic. However, glycine and cysteine occupy intermediate positions on the hydrophilic–hydrophobic scale (glycine has just a single hydrogen as its side chain, and the –SH group is not so polar as an –OH group).

The amino acids of proteins often undergo chemical modification of the side chains. Quite often a very simple chemical group is added to the side chain of the amino acid, but sometimes a large carbohydrate, lipid, or even another protein is joined to the side chain, as described below.

The role of chemical bonding in the stability and function of macromolecules

The stability of nucleic acid and protein polymers is primarily dependent on strong covalent bonds between the atoms of their linear backbones. In addition to covalent bonds, weak noncovalent bonds (Table 1.2) are important in stabilizing molecules and in allowing a variety of transient interactions between diverse molecules within cells. Whereas covalent bonds are comparatively stable, and require a high input of energy to break them, individual noncovalent bonds are typically >10 times weaker than individual covalent bonds. As a result, they are constantly being made and broken at physiological temperatures.

TABLE 1.2 WEAK NONCOVALENT BONDS AND FORCES

Type of bond	Nature of bond
Hydrogen	Hydrogen bonds form when a hydrogen atom interacts with electron-attracting atoms, usually oxygen or nitrogen atoms
Ionic	Ionic interactions occur between charged groups. They can be very strong in crystals but in an aqueous environment the charged groups are shielded by both H_2O molecules and ions in solution and so are quite weak. Nevertheless, they can be very important in biological function, as in enzyme–substrate recognition
Van der Waals forces	Any two atoms in close proximity show a weak attractive bonding interaction due to their fluctuating electrical charges (van der Waals attraction). When atoms become extremely close, they repel each other very strongly (van der Waals repulsion). Although the forces are individually very weak, van der Waals attraction can be important when there is a very good fit between the surfaces of two macromolecules
Hydrophobic forces	Water is a polar molecule. Hydrophobic molecules or chemical groups in an aqueous environment tend to cluster. This minimizes their disruptive effects on the complex network of hydrogen bonds between water molecules. Hydrophobic groups are said to be held together by hydrophobic bonds, although the basis of their attraction is their common repulsion by water molecules

The cellular environment is an aqueous one and the structure of water is particularly complex, with a rapidly fluctuating network of noncovalent bonding occurring between water molecules. The predominant force in this structure is the **hydrogen bond**, a weak electrostatic bond between fractionally positive hydrogen atoms and fractionally negative atoms (oxygen atoms, in the case of water molecules).

Charged molecules are highly soluble in water. Because of the phosphate groups in their component nucleotides, both DNA and RNA are negatively-charged polyanions. Depending on their amino acid composition, proteins may be electrically neutral, or they may carry a net positive charge (**basic protein**) or a net negative charge (**acidic protein**). All of these molecules can form multiple interactions with the water during their solubilization. Even electrically neutral proteins are readily soluble if they contain sufficient charged or neutral polar amino acids. In contrast, membrane-bound proteins with many hydrophobic amino acids are thermodynamically more stable in a hydrophobic environment.

Although individually weak, the combined action of numerous noncovalent bonds can make large contributions to the stability of the **conformation** (structure) of macromolecules and are important for specifying their shape. We describe in the next section how hydrogen bonds between pairs of bases are essential for maintaining the structure of DNA and RNA molecules; and in the final section of this chapter we illustrate the central role of hydrogen bonding in determining the shape of diverse structural motifs in proteins, including the classic α-helices, β-sheets, and so on.

Because noncovalent bonds are fragile and able to be broken and remade easily, they also allow transient interactions between different molecules. Hydrogen bonding is especially important in allowing transient interactions between different nucleic acids, facilitating the recognition by regulatory RNAs of target sequences in other RNAs or in DNA. We provide examples in different chapters, notably when we consider gene regulation.

1.2 BASE PAIRING IN DNA AND RNA, THE DOUBLE HELIX, AND DNA REPLICATION

As described above, nucleic acids have a sugar–phosphate backbone with alternating sugar residues and phosphate groups. Neighboring sugar residues are linked by **3′–5′ phosphodiester bonds**, in which a phosphate group links the 3′ carbon atom of one sugar to the 5′ carbon atom of the next sugar in the sugar–phosphate backbone (Figure 1.5).

The genetic material of certain viruses is single-stranded DNA, but each cellular DNA species has two DNA strands (a DNA duplex). The two DNA strands are structured as a **double helix**: they curve around each other and each base on one DNA strand is noncovalently linked (by hydrogen bonding) to a laterally opposed base on the opposite DNA strand, forming a **base pair** (Figure 1.6).

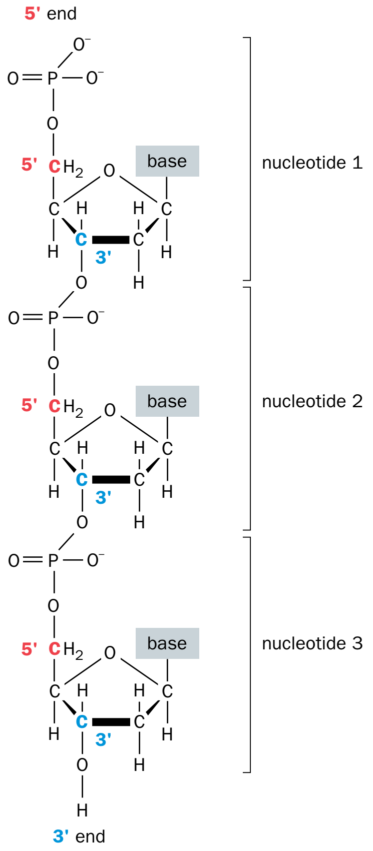

Figure 1.5 Repeating structure and asymmetric 5′ and 3′ ends of a nucleic acid strand. The repeat unit of a DNA or RNA strand is a nucleotide, consisting of a sugar with an attached base and phosphate group and, for simplicity, we show here a trinucleotide in which the 5′ carbons (red) and 3′ carbons (blue) are highlighted. There is asymmetry in how neighboring sugars are joined by the intervening phosphate group. That is, a phosphodiester bond connecting two sugars in a nucleic acid joins the carbon 3′ of one sugar to the carbon 5′ of a neighbor (a 3′–5′ phosphodiester bond). This results in asymmetry at the linear ends of the strand where the terminal nucleotides will have a sugar with either a carbon 3′ or a carbon 5′ atom that is not joined to a neighboring sugar. At the 5′ end of a nucleic acid strand the carbon 5′ of the sugar has a free phosphate group, and at the 3′ end the carbon 3′ of the sugar is attached to a hydroxyl group only.

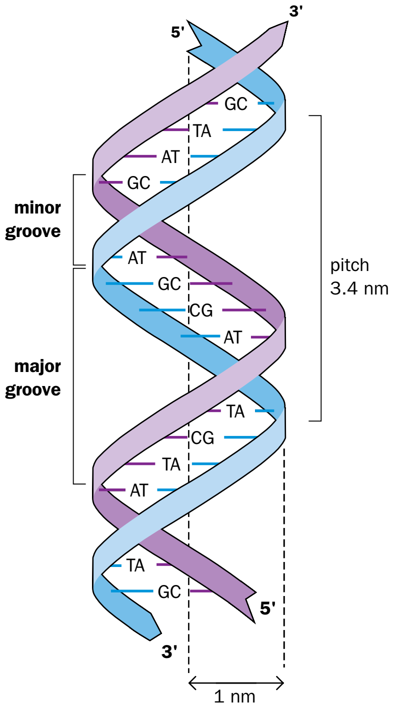

Figure 1.6 Structural features of B-DNA, the most common form of a DNA double helix. The two DNA strands of a double helix wind round each other. Under physiological conditions the B-form of a double helix is the most common form in bacterial and eukaryotic cells. It is a right-handed helix (imagine looking from one end of the helix as it spirals away from you into the distance: if the DNA strands spiral away from you in a clockwise direction you have a right-handed helix; if they spiral away in an anticlockwise direction, the helix is left-handed). B-DNA has a pitch of 3.4 nm, a radius of 1 nm per turn, and 10 base pairs per turn, and has a minor groove and a broader major groove (which facilitates access to DNA-binding proteins). See **Figure 1.13** for alternative structures for a double helix.