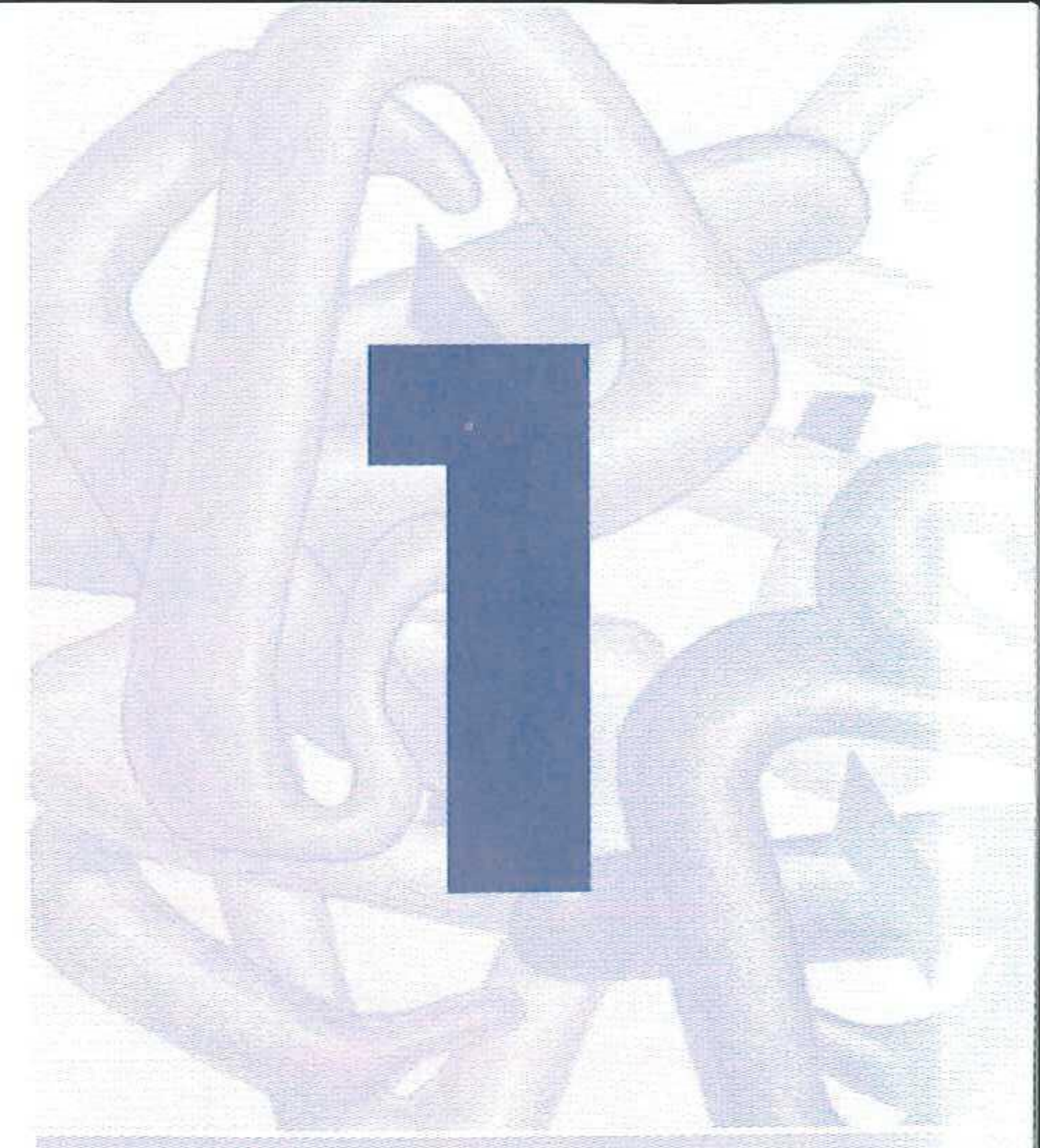


1

THE NUCLEIC ACID WORLD

When you have read Chapter 1, you should be able to:

State the chemical structures of nucleic acids.

Explain base-pairing and the double helix.

Explain how DNA stores genetic information.

Summarize the intermediate role of mRNA between DNA and proteins.

Outline how mRNA is translated into protein by ribosomes.

Outline how gene control is exercised by binding to short nucleotide sequences.

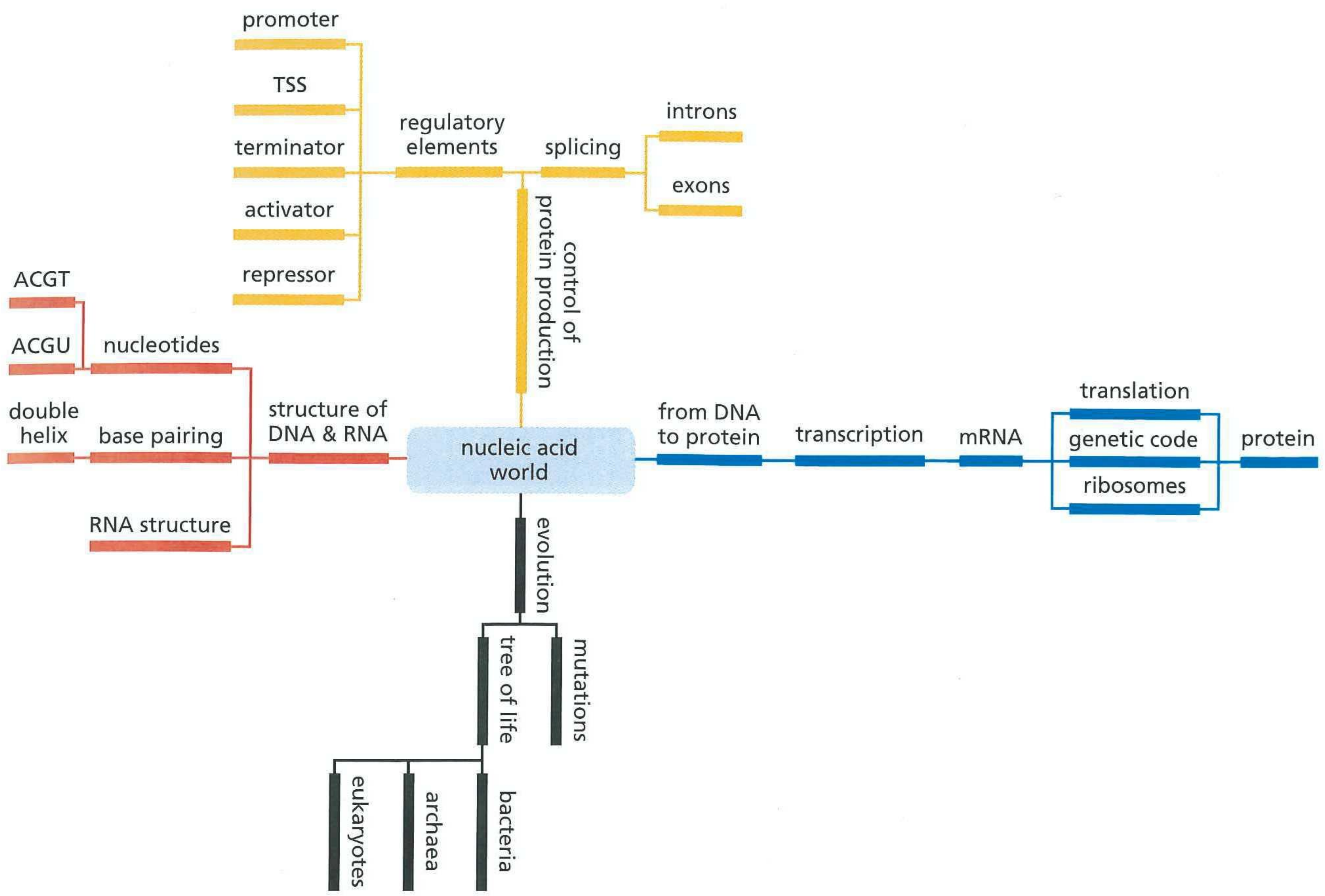
Show that eukaryotic mRNA often has segments (introns) removed before translation.

Discuss how all life probably evolved from a single common ancestor.

Summarize how evolution occurs by changes to the sequence of genomic DNA.

It is amazing to realize that the full diversity of life on this planet—from the simplest bacterium to the largest mammal—is captured in a linear code inside all living cells. In almost exactly the way the vivid detail of a musical symphony or a movie can be digitally recorded in a binary code, so the four base units of the DNA molecule capture and control all the complexity of life. The crucially important discovery of the link between DNA, proteins, and the diversity of life came during the twentieth century and brought about a revolution in the understanding of genetics. Since that time we have amassed increasing amounts of information on the sequences of DNA, RNA, and proteins and the great variety of these molecules in cells under differing conditions. The growth of this information threatens to outstrip our ability to analyze it. It is one of the key challenges facing biologists today to organize, study, and draw conclusions from all this information: the patterns within the sequences and experimental data, the structure and the function of the various types of molecules, and how everything interacts to produce a correctly functioning organism. Bioinformatics is the name we give to this study and our aim is to use it to obtain greater understanding of living systems.

Information about nucleic acids and proteins is the raw material of bioinformatics, and in the first two chapters of this book we will briefly review these two types of biomolecules and their complementary roles in reproducing and maintaining life. This review is not a comprehensive introduction to cell and molecular biology; for that you should consult one of the excellent introductory textbooks of cell and molecular biology listed in Further Reading (p. 24). Rather, it is intended simply as a reminder of those aspects of nucleic acids and proteins that you will need as a background to the bioinformatics methods described in the rest of the book. More



Mind Map 1.1
A mind map schematic of the topics covered in this chapter and divided, in general, according to the topic sections. This is to help you visualize the topics, understand the structure of the chapter, and memorize the important elements.

information about the biological context of the bioinformatics problems we discuss is also given in the biology boxes and glossary items throughout the book.

This chapter will deal with the nucleic acids—**deoxyribonucleic acid** (**DNA**) and **ribonucleic acid** (**RNA**)—and how they encode proteins, while the structure and functions of proteins themselves will be discussed in Chapter 2. In these two chapters we shall also discuss how DNA changes its information-coding and functional properties over time as a result of the processes of **mutation**, giving rise to the enormous diversity of life, and the need for bioinformatics to understand it.

The main role of DNA is information storage. In all living cells, from unicellular bacteria to multicellular plants and animals, DNA is the material in which genetic instructions are stored and is the chemical structure in which these instructions are transmitted from generation to generation; all the information required to make and maintain a new organism is stored in its DNA. Incredibly, the information required to reproduce even very complex organisms is stored on a relatively small number of DNA molecules. This set of molecules is called the organism's **genome**. In humans there are just 46 DNA molecules in most cells, one in each chromosome. Each DNA molecule is copied before cell division, and the copies are distributed such that each daughter cell receives a full set of genetic information. The basic set of 46 DNA molecules together encode everything needed to make a human being. (We will skip over the important influence of the environment and the nature–nurture debate, as they are not relevant to this book.)

Proteins are manufactured using the information encoded in DNA and are the molecules that direct the actual processes on which life depends. Processes essential to life, such as energy metabolism, biosynthesis, and intercellular communication, are all carried out through the agency of proteins. A few key processes such as the synthesis of proteins also involve molecules of RNA. Ignoring for a moment some of the complexity that can occur, a gene is the information in DNA that directs the manufacture of a specific protein or RNA molecular form. As we shall see, however, the organization of genes within the genome and the structure of individual genes are somewhat different in different groups of organisms, although the basic principles by which genes encode information are the same in all living things.

Organisms are linked together in evolutionary history, all having evolved from one or a very few ancient ancestral life forms. This process of evolution, still in action, involves changes in the genome that are passed to subsequent generations. These changes can alter the protein and RNA molecules encoded, and thus change the organism, making its survival more, or less, likely in the circumstances in which it lives. In this way the forces of evolution are inextricably linked to the genomic DNA molecules.

1.1 The Structure of DNA and RNA

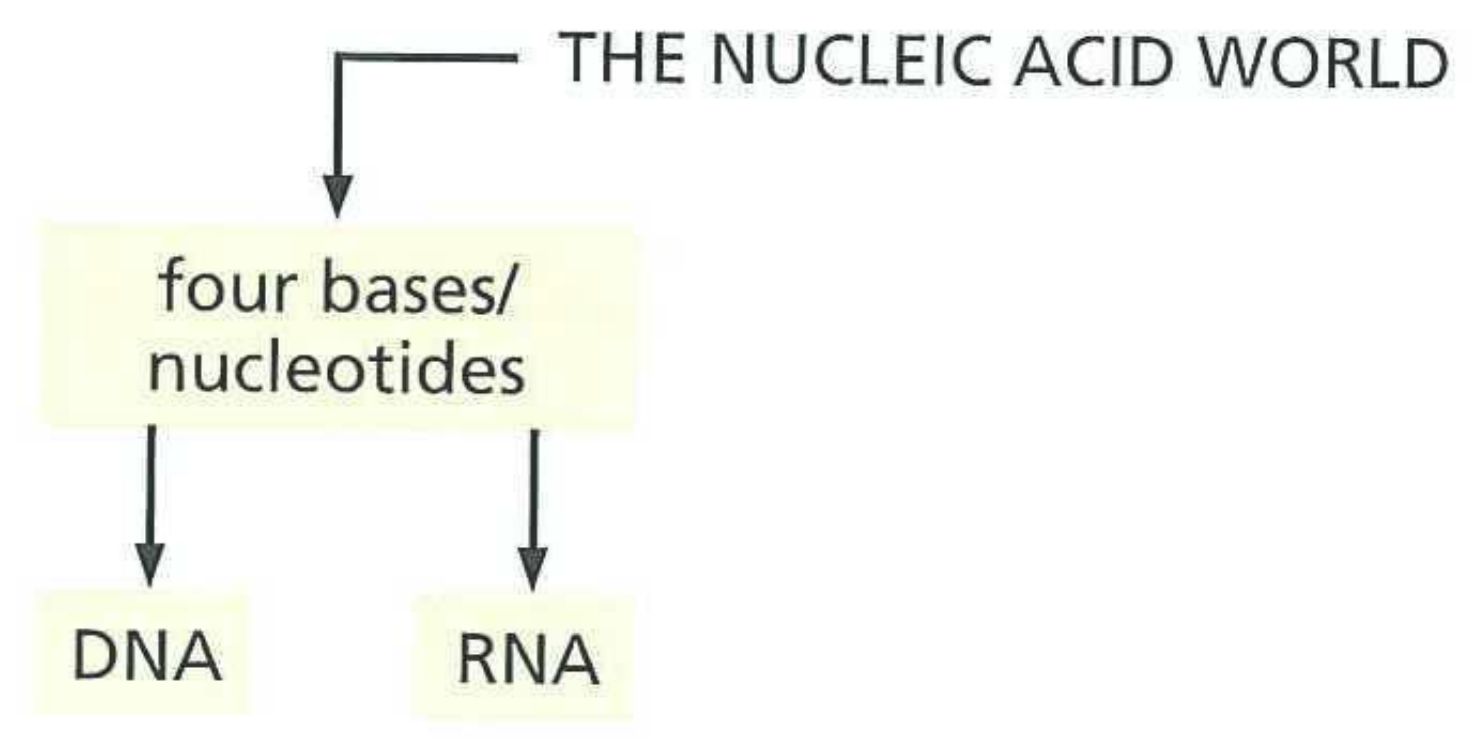
Considering their role as the carrier of genetic information, DNA molecules have a relatively simple chemical structure. They are linear polymers of just four different **nucleotide** building blocks whose differences are restricted to a substructure called the **base** (see Flow Diagram 1.1). For efficient encoding of information, one might have expected there to be numerous different bases, but in fact there are only four. This was one of the reasons why the true role of DNA as the carrier of genetic information was not appreciated until the 1940s, long after the role of the chromosomes in heredity was apparent. But although chemically simple, genomic DNA molecules are immensely long, containing millions of bases each, and it is the order of these bases, the **nucleotide sequence** or **base sequence** of DNA, which encodes the information for making proteins.

The three-dimensional structure of DNA is also relatively simple, involving regular double helices. There are also larger-scale regular structures, but it has been clearly established that the information content of DNA is held at the level of the base sequence itself.

RNA molecules are also linear polymers, but are much smaller than genomic DNA. Most RNA molecules also contain just four different base types. However, several classes of RNA molecules are known, some of which have a small proportion of other bases. RNA molecules tend to have a less-regular three-dimensional structure than DNA.

DNA is a linear polymer of only four different bases

The building blocks of DNA and RNA are nucleotides, which occur in different but chemically related forms. A nucleotide consists of a nitrogen-containing base that is linked to a five-carbon sugar ring at the 1′ position on the ring, which also carries a phosphate group at the 5′ position (see Figure 1.1A). In DNA the sugar is deoxyribose, in RNA it is ribose, the difference between the two being the absence or presence,



Flow Diagram 1.1
The key concept introduced in this section is that DNA and RNA are composed of subunits called nucleotides, with only four different nucleotide types in a molecule, but a different set of four nucleotides in each of the two types of nucleic acid.

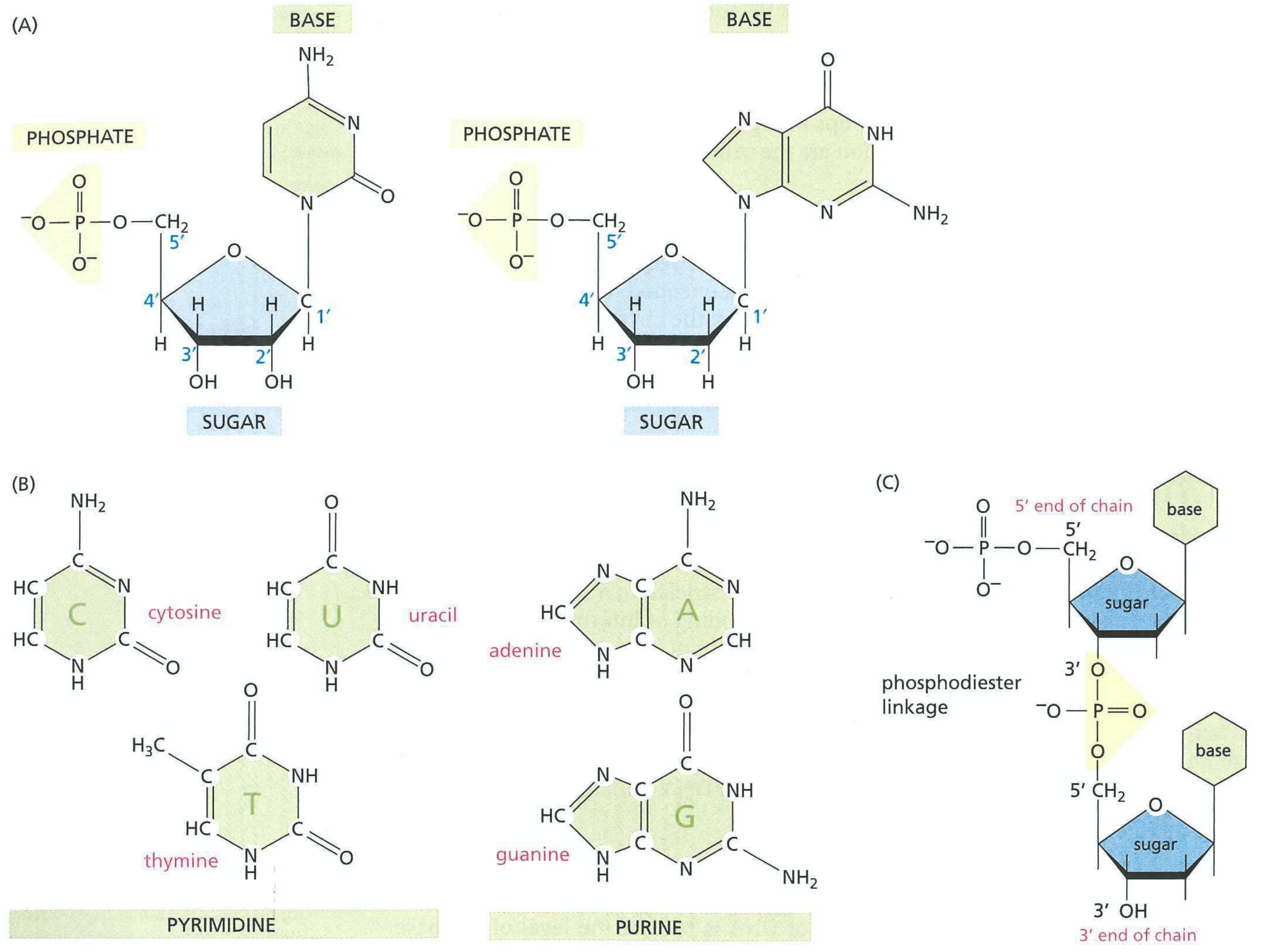


Figure 1.1
The building blocks of DNA and RNA. (A) Left, cytidylate (ribo-CMP); right, deoxyguanylate (dGMP). Each consists of three main parts: a phosphate group, a pentose sugar, and a base. It is the base part that is responsible for the differences in the nucleotides. (B) The bases fall into two groups: the purines and the pyrimidines. The purines consist of a 6- plus a 5-membered nitrogen-containing ring while the pyrimidines have only one 6-membered nitrogen-containing ring. (C) It is the phosphate group that is involved in linking the building blocks together by a phosphodiester linkage in DNA. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

respectively, of a hydroxyl group at the 2′ position. These two types of nucleotide are referred to as **deoxyribonucleotides** and **ribonucleotides,** respectively. Apart from this, the only difference between nucleotides is the base, which is a planar ring structure, either a pyrimidine or a purine (see Figure 1.1B). In DNA only four different bases occur: the purines guanine (G) and adenosine (A) and the pyrimidines cytosine (C) and thymine (T). In most forms of RNA molecule there are also just four bases, three being the same as in DNA, but thymine is replaced by the pyrimidine uracil (U).

Each nucleic acid chain, or strand, is a linear polymer of nucleotides linked together by phosphodiester linkages through the phosphate on one nucleotide and the hydroxyl group on the 3′ carbon on the sugar of another (see Figure 1.1C). This process is carefully controlled in living systems so that a chain is exclusively made with either deoxyribonucleotides (DNA) or ribonucleotides (RNA). The resulting chain has one end with a free phosphate group, which is known as the **5′ end**, and one end with a free 3′ hydroxyl group, which is known as the **3′ end**. The base sequence or nucleotide sequence is defined as the order of the nucleotides along the chain from the 5′ to the 3′ end. It is normal to use the one-letter code given above to identify the bases, starting at the 5′ end, for example AGCTTAG.

There are instances of bases within a nucleic acid chain being modified in a living cell. Although relatively rare, they can be of great interest. In the case of DNA, in vertebrates cytosine bases can be methylated (addition of a -CH_3 group). This can result in certain genes being rendered inactive, and is involved in the newly discovered phenomenon known as **genomic imprinting**, where the change that may occur in the offspring depends on whether the gene is maternally or paternally inherited. The class of RNA molecules called tRNA (see Section 1.2) have modifications to approximately 10% of their bases, and many different modifications have been seen involving all base types. These changes are related to the function of the tRNA.

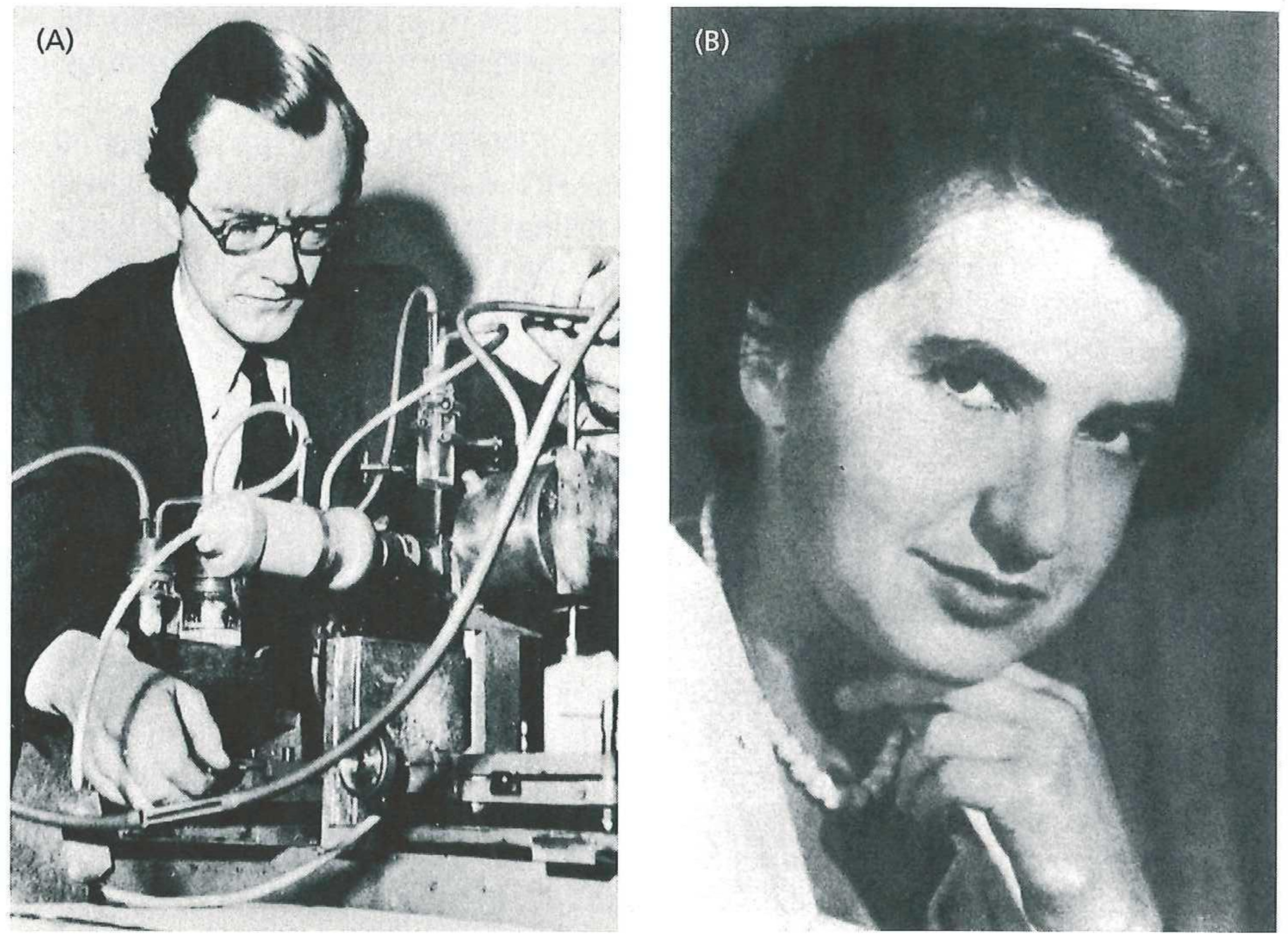


Figure 1.2
Two scientists whose work was influential on James Watson and Francis Crick when they elucidated the structure of DNA.
(A) Maurice Wilkins.
(B) Rosalind Franklin. (A and B courtesy of Science Photo Library.)

Two complementary DNA strands interact by base-pairing to form a double helix

The key discovery for which the Nobel Prize was awarded to James Watson and Francis Crick, who drew on the work of Rosalind Franklin and others (see Figure 1.2), was the elucidation in 1953 of the double-helical structure of the DNA molecule, in which two strands of DNA are wound around each other and are held together by hydrogen bonding between the bases of each strand. Their structure was an early example of model building of large molecules, and was based on knowledge of the chemical structure of the constituent nucleotides and experimental X-ray diffraction data on DNA provided by, among others, Maurice Wilkins, who was also awarded a share in the Nobel Prize. (Current methods of model building as applied to proteins are discussed in Chapter 13.)

All the bases are on the inside of the double helix, with the phosphate-linked sugars forming a backbone on the outside (see Figure 1.3A). Crucial to Watson and Crick's success was their realization that the DNA molecules contained two strands and that the base-pairing follows a certain set of rules, now called **Watson–Crick base-pairing**, in which a specific purine pairs only with a specific pyrimidine: A with T, and C with G. Each strand of a DNA double helix therefore has a base sequence that is **complementary** to the base sequence of its partner strand. The bases interact through hydrogen bonding; two hydrogen bonds are formed in a T–A base pair, and three in a G–C base pair (see Figure 1.3B). This complementarity means that if you know the base sequence of one strand of DNA, you can deduce the base sequence of the other. Note, however, that the two strands are antiparallel, running in opposite directions, so that the complementary sequence to AAG is not TTC but CTT (see Figure 1.3A).

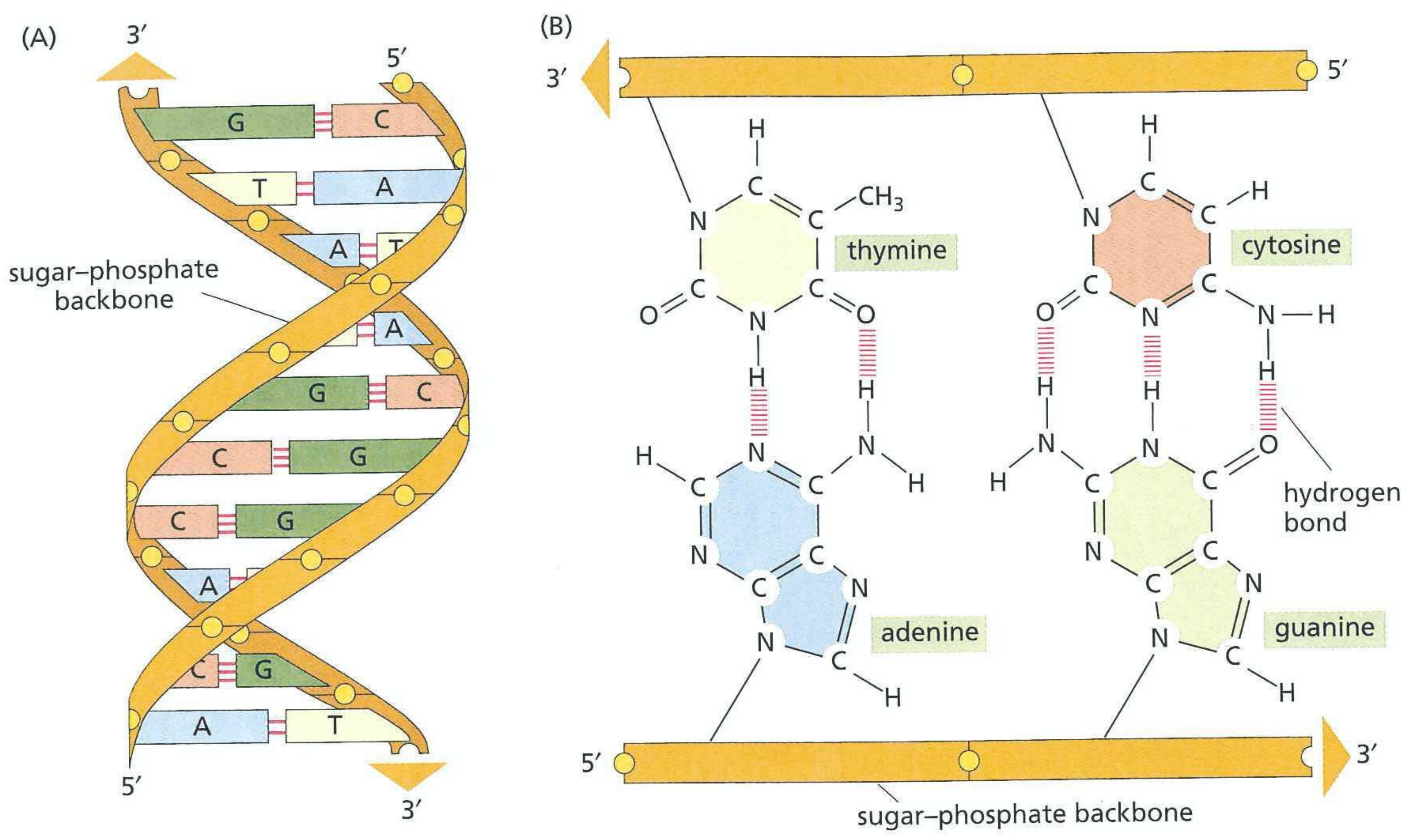


Figure 1.3
The double helical structure of DNA. (A) DNA exists in cells mainly as a two-stranded coiled structure called the double helix. (B) The two strands of the helix are held together by hydrogen bonds (shown as red lines) between the bases; these bonds are referred to as base-pairing. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

Hydrogen bonds are noncovalent bonds, which in biomolecules are much weaker than covalent bonds, often involving a binding energy of only 5–20 kJ mol^{-1}. As a consequence the two strands of the DNA double helix can be relatively easily separated. There are a number of processes in which this strand separation is required, for example when the molecules of DNA are copied as a necessary prelude to cell division, a process called **DNA replication**. The separated strands each serve as a template on which a new complementary strand is synthesized by an enzyme called DNA polymerase, which moves along the template successively matching each base in the template to the correct incoming nucleotide and joining the nucleotide to the growing new DNA strand. Thus each double helix gives rise to two new identical DNA molecules (see Figure 1.4). Genetic information is thus preserved intact through generations of dividing cells. The actual biochemical process of DNA replication is complex, involving a multitude of proteins, and will not concern us here, but at its heart is this simple structural property of complementarity.

One of the truly remarkable features of DNA in living systems is how few errors there are in the complementarity of the two strands. The error rate is approximately 1 base in 10^9. This is very important, as it is vital to transmit the genome as accurately as possible to subsequent generations. Many alternative base-pairings are possible, although these are less favorable than the Watson–Crick base-pairing. These energies can be used to predict the expected rate of incorrect base-pairing. However, genomic DNA shows a much lower error rate than expected. This is a result of the controlled way in which the DNA polymerase builds the second strand, including mechanisms for checking and correcting erroneous base-pairings.

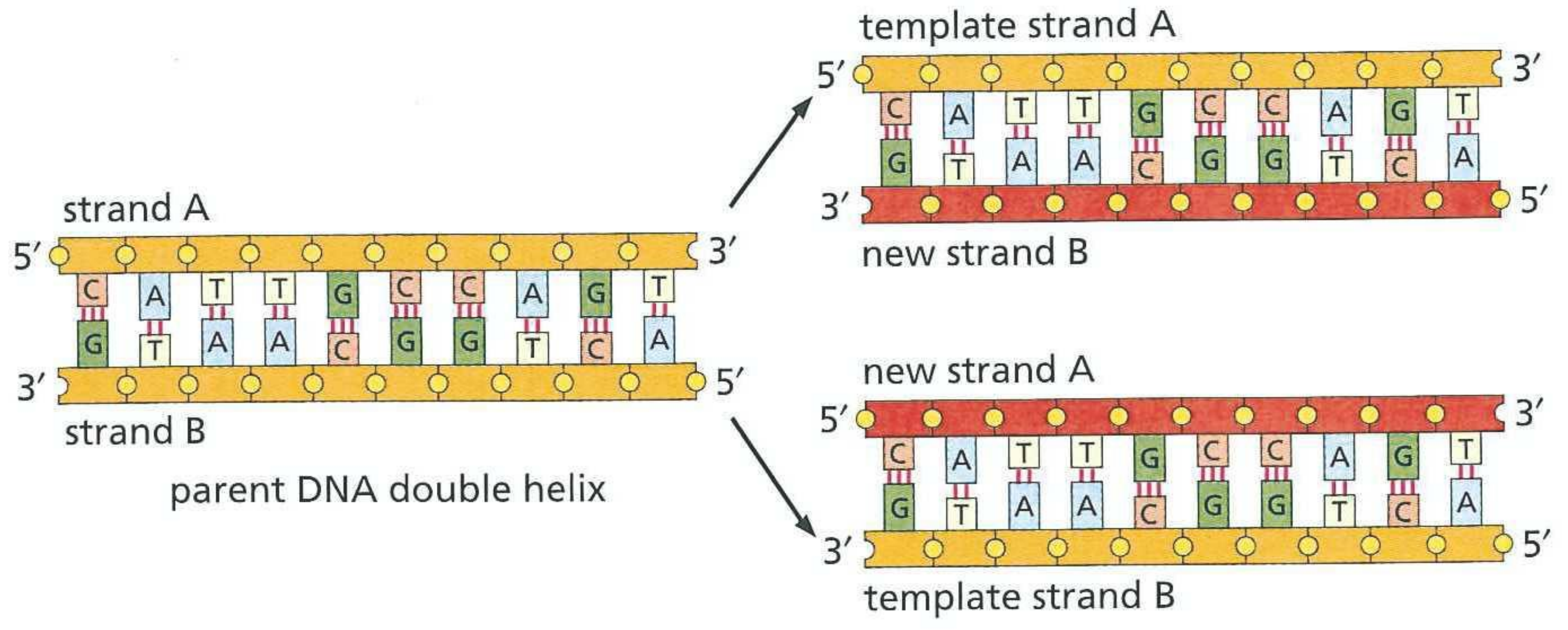


Figure 1.4
DNA replication. DNA duplicates itself by first separating the two strands, after which enzymes come and build complementary strands base-paired to each of the old strands. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

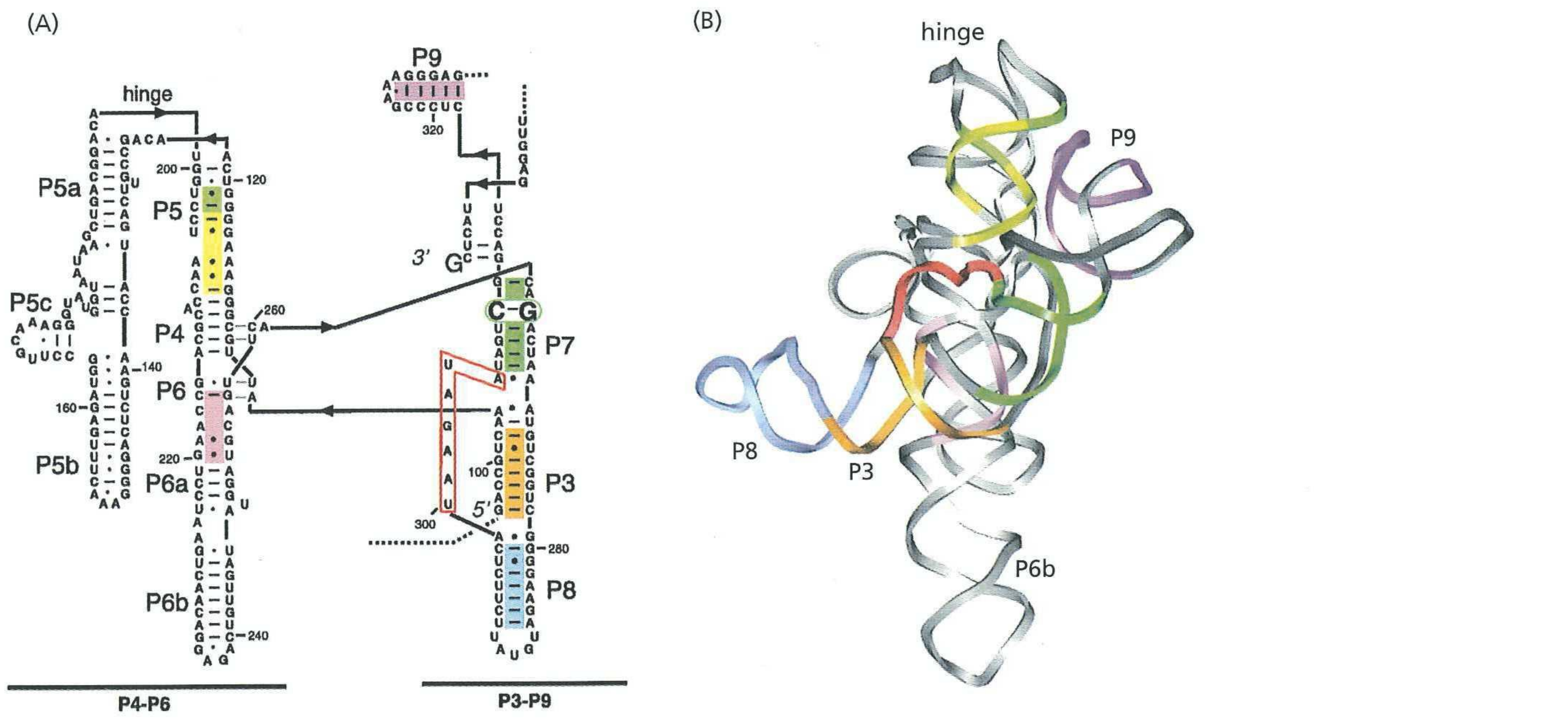
DNA strands can pair with a DNA or RNA strand of complementary sequence to make a double-stranded DNA or DNA/RNA hybrid. This property forms the basis of a set of powerful experimental molecular biology techniques. Known generally as nucleic acid **hybridization**, it is exploited in applications such as DNA **microarrays** (described in Chapter 15), *in situ* hybridization to detect the activity of specific genes in cells and tissues, and fluorescence *in situ* hybridization (FISH) for visually locating genes on chromosomes.

RNA molecules are mostly single stranded but can also have base-pair structures

In contrast to DNA, almost all RNA molecules in living systems are single stranded. Because of this, RNA has much more structural flexibility than DNA, and some RNAs can even act as enzymes, catalyzing a particular chemical reaction. The large number of hydrogen bonds that can form if the RNA molecule can double back on itself and create base-pairing makes such interactions highly energetically favorable. Often, short stretches of an RNA strand are almost or exactly complementary to other stretches nearby. Two such stretches can interact with each other to form a double-helix structure similar to that of DNA, with loops of unpaired nucleotides at the ends of the helices. The interactions stabilizing such helices are often not standard Watson–Crick pairing. These structures are often referred to as **RNA secondary structure** by analogy with protein secondary structure described in Section 2.1.

It is likely that all RNA molecules in a cell have some regions of stable three-dimensional structure due to limited base-pairing, the unpaired regions being very flexible and not taking up any specific structure. In some cases RNA folds up even further, packing helices against each other to produce a molecule with a rigid structure. The three-dimensional structures that result have a functional role for many RNA molecules. An example of such structures is shown in Figure 1.5, and is fairly typical in that a significant fraction of the sequence is involved in base-pairing interactions. The prediction of ribosomal secondary structures and three-dimensional structure is not covered in this book, but introductory references are given in Further Reading. Even more complex interactions are possible, one example of

Figure 1.5
Example three-dimensional structure of RNA. The structure shown is the Tetrahymena ribozyme. (A) Nucleotide sequence showing base-pairing and loops. (B) Three-dimensional structure as determined by x-ray crystallography. Entry 1GRZ of MSD database. (From B.L. Golden et al., A preorganized active site in the crystal structure of Tetrahymena ribozyme, *Science* 282:259–264, 1998. Reprinted with permission from AAAS.)



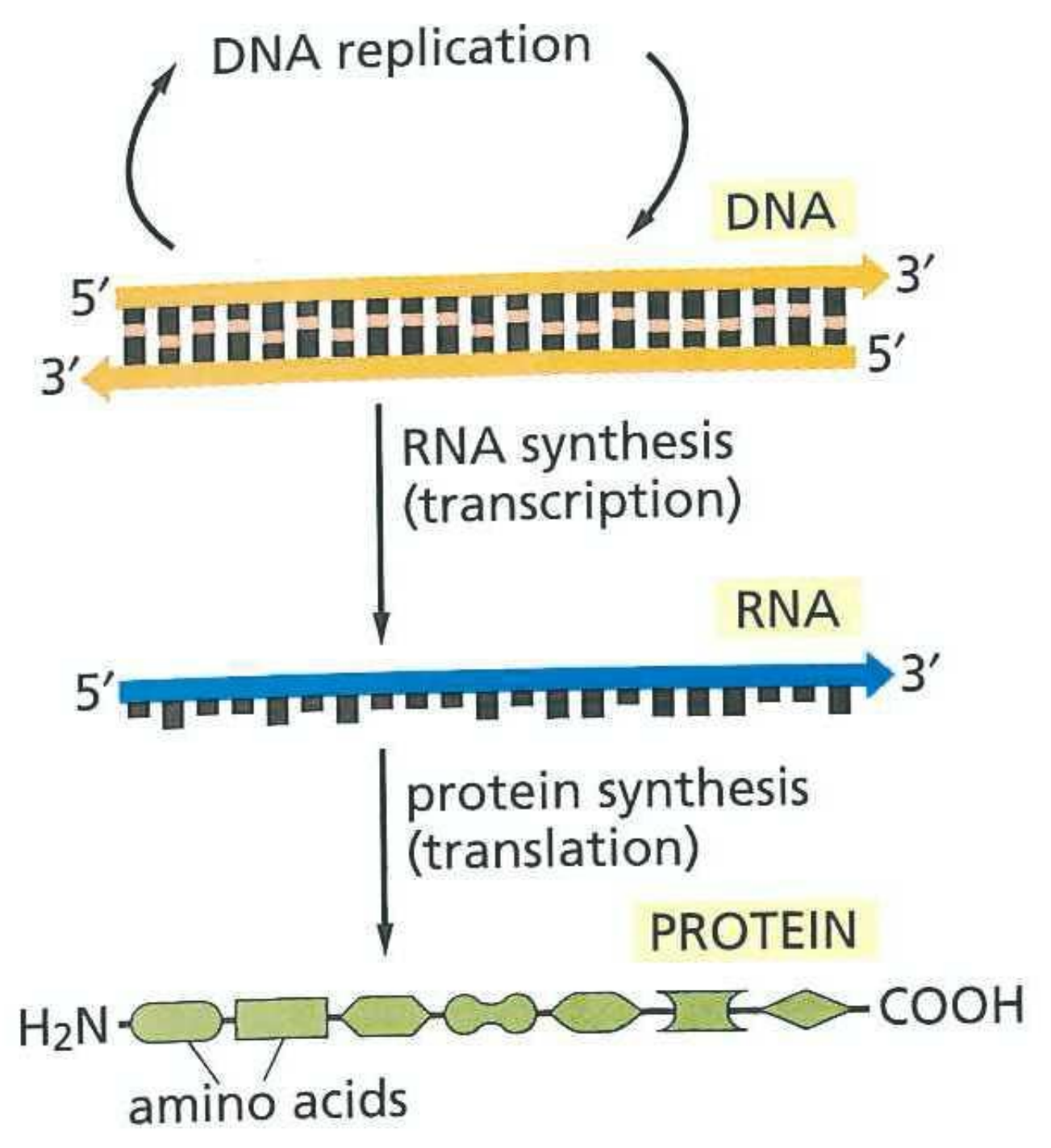

DNA replication
DNA
5′
3′
3′
5′
RNA synthesis (transcription)
RNA
5′
3′
protein synthesis (translation)
PROTEIN
H2N
COOH
amino acids

Figure 1.6
Central dogma. This is the basic scheme of the transcription of DNA to RNA (mRNA) which is then translated to proteins. Many enzymes are involved in these processes. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

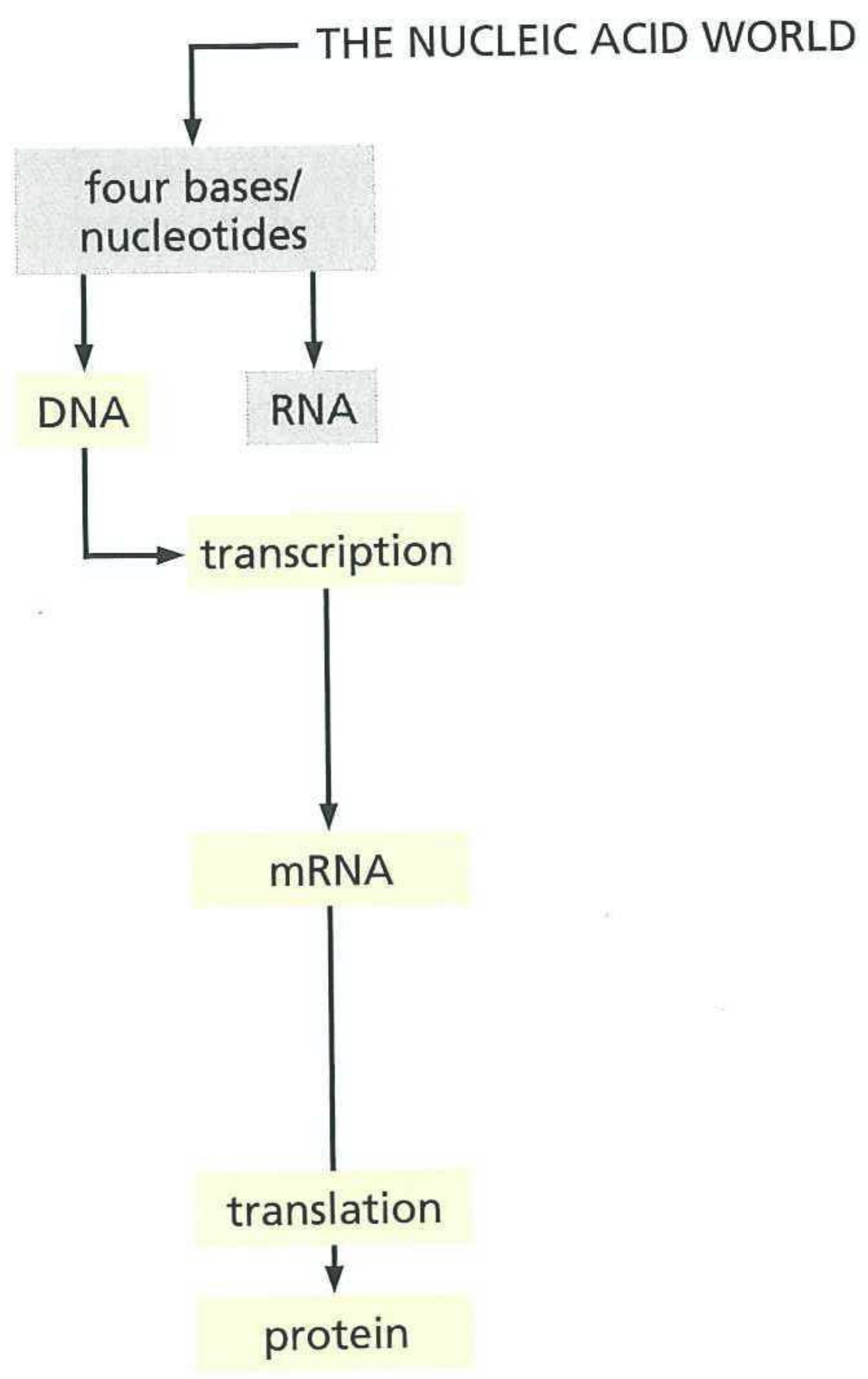
which is illustrated in Figure 1.5A. To the right of the P4 and P6 labels are four bases that are involved in two separate base-pairings, in each case forming a base triplet. Such interactions have been observed on several occasions, as have interactions involving four bases.

1.2 DNA, RNA, and Protein: The Central Dogma

There is a key relationship between DNA, RNA, and the synthesis of proteins, which is often referred to as the central dogma of molecular biology. According to this concept, there is essentially a single direction of flow of genetic information from the DNA, which acts as the information store, through RNA molecules from which the information is translated into proteins (see Figure 1.6 and Flow Diagram 1.2). This basic scheme holds for all known forms of life, although there is some variation in the details of the processes involved. The proteins are the main working components of organisms, playing the major role in almost all the key processes of life. However, not all the genetic information in the DNA encodes proteins. Molecules such as RNA can also be the end product, and other regions of genomes have as yet no known function or product. The genomic DNA encodes all molecules that are necessary for life, whether they are the proteins (such as enzymes) involved in nearly all biological activities or RNA important for translation and transcription.

For example, all the information needed to build and maintain a human being is contained in just 23 pairs of DNA molecules, comprising the chromosomes of the human genome. These molecules are amongst the largest and longest known, the smallest having 47 million bases and the largest 247 million bases, with the entire human genome being composed of approximately 3 billion bases. Even bacterial genomes, which are much smaller than this, tend to have several million bases. The DNA of each chromosome encodes hundreds to thousands of proteins, depending on the chromosome, each protein being specified by a distinct segment of DNA. In simple terms, this segment is the gene for that protein. In practice, a **gene** is considered also to include surrounding regions of noncoding DNA that act as control regions, as will be described in Section 1.3. These are involved in determining whether the gene is active—in which case the protein is produced—or is inactive.

Flow Diagram 1.2
The key concept introduced in this section is that DNA is transcribed into mRNA which is then translated to make protein molecules. The shaded boxes show the concepts introduced in the previous section, the yellow-colored boxes refer to concepts in this section. This color coding is used throughout the book.



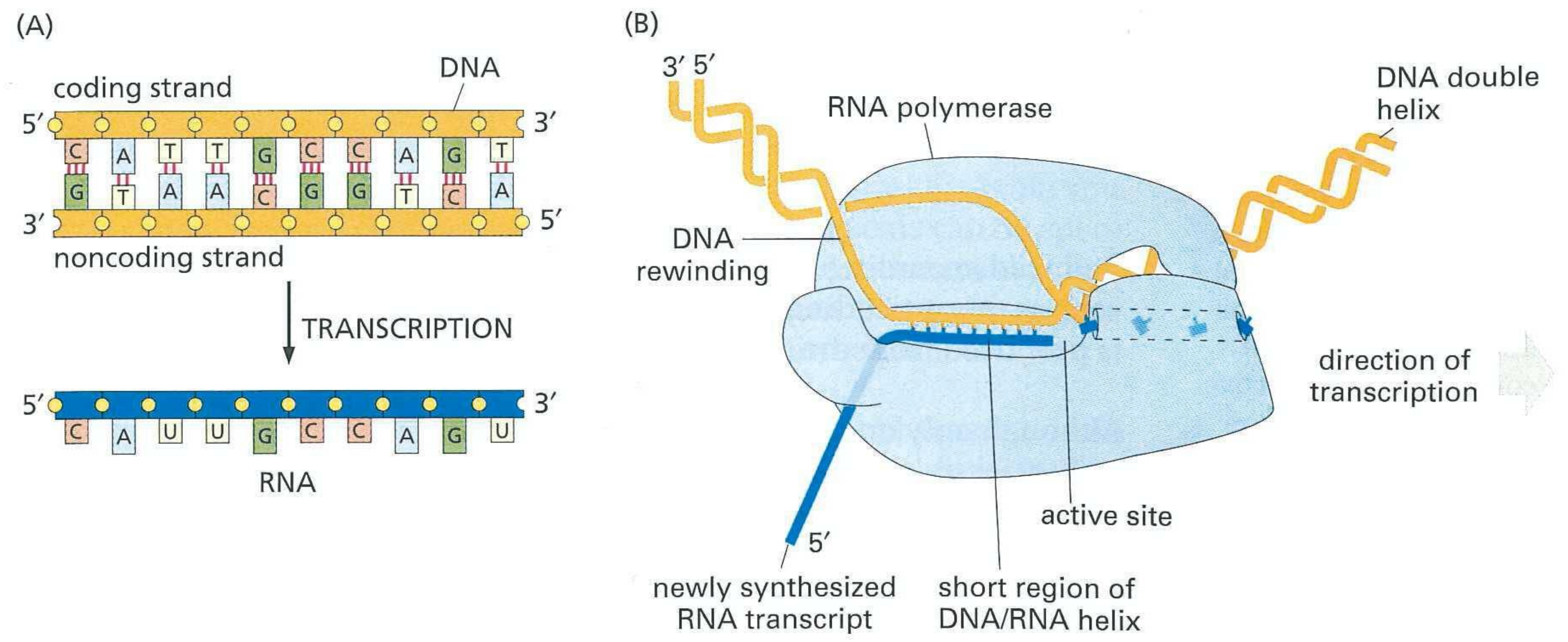


Figure 1.7
Transcription. (A) One strand of the DNA is involved in the synthesis of an RNA strand complementary to the strand of the DNA. (B) The enzyme RNA polymerase is involved in the transcription process. It reads the DNA and recruits the correct building blocks of RNA to string them together based on the DNA code. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

The sequence of the protein-coding region of a gene carries information about the protein sequence in a coded form called the genetic code. This DNA sequence is decoded in a two-stage process (see Figure 1.6), the stages being called transcription and translation. Both stages involve RNA molecules and will now be described.

DNA is the information store, but RNA is the messenger

The information encoded in DNA is expressed through the synthesis of other molecules; it directs the formation of RNA and proteins with specific sequences. As is described in detail in Chapter 2, proteins are linear polymers composed of another set of chemical building blocks, the **amino acids**. There are some 20 different amino acids in proteins, and their varied chemistry (see Figure 2.3) makes proteins much more chemically complex and biochemically versatile molecules than nucleic acids.

The sequence of bases in the DNA of a gene specifies the sequence of amino acids in a protein chain. The conversion does not occur directly, however. After a signal to switch on a gene is received, a single-stranded RNA copy of the gene is first made in a process called **transcription**. Transcription is essentially similar to the process of DNA replication, except that only one of the DNA strands acts as a template in this case, and the product is RNA not DNA (see Figure 1.7). RNA synthesis is catalyzed by enzymes called RNA polymerases, which, like DNA polymerases, move along the template, matching incoming ribonucleotides to the bases in the template strand and joining them together to make an RNA chain. Only the relevant region of DNA is transcribed into RNA, therefore the RNA is a much smaller molecule than the DNA it comes from. So while the DNA carries information about many proteins, the RNA carries information from just one part of the DNA, usually information for a single protein. RNA transcribed from a protein-coding gene is called **messenger RNA** (**mRNA**) and it is this molecule that directs the synthesis of the protein chain, in the process called translation, which will be described in more detail below. When a gene is being transcribed into RNA, which is in turn directing protein synthesis, the gene is said to be **expressed**. Expression of many genes in a cell or a set of cells can be measured using DNA or RNA expression microarrays (see Chapter 15).

Only one of the DNA strands in any given gene is transcribed into RNA. As the RNA must have a **coding** sequence that can be correctly translated into protein, the DNA strand that acts as the physical template for RNA synthesis does not carry the coding sequence itself, but its complement. It is therefore known as the **noncoding strand** or **anticoding** or **antisense strand**. The sequence of the other, non-template DNA strand is identical to that of the messenger RNA (with T replacing U), and this strand is called the coding or **sense strand**. This is the DNA sequence that is written out to represent a gene, and from which the protein sequence can be

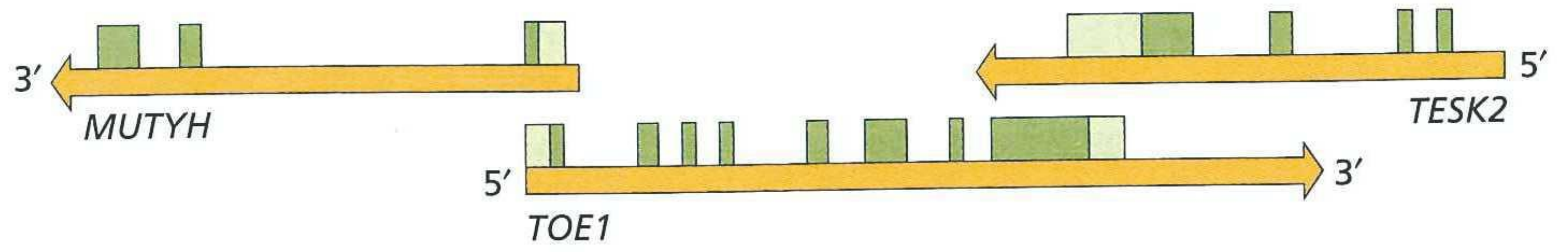


Figure 1.8
Overlapping genes. A schematic showing the overlap of three human genes. The dark green boxes show protein-coding regions of the DNA (exons) while the light green boxes show regions that are untranslated. (Adapted from V. Veeramachaneni et al., Mammalian overlapping genes: the comparative perspective, *Genome Res.* 14:280–286, 2004.)

deduced according to the rules of the genetic code. Note that the polymerase transcribes the anticoding strand in the direction from 3′ to 5′, so that the mRNA strand is produced from the 5′ to the 3′ end.

Although only one segment of the DNA strand is transcribed for any given gene, it is also possible for genes to overlap so that one or both strands at the same location encode parts of different proteins. This most commonly occurs in **viruses** as a means of packing as much information as possible into a very small genome. However, overlapping genes can also occur in mammals; recently 774 pairs of overlapping genes were identified in the human genome (see Figure 1.8).

The genomic DNA sequence contains more information than just the protein sequences. The transcriptional apparatus has to locate the sites where gene transcription should begin, and when to transcribe a given gene. At any one time, a cell is only expressing a few thousand of the genes in its genome. To accomplish this regulated gene expression, the DNA contains control sequences in addition to coding regions. We shall return to these regulatory regions later, after first discussing the details of the coding of protein sequences.

Messenger RNA is translated into protein according to the genetic code

The **genetic code** refers to the rules governing the correspondence of the base sequence in DNA or RNA to the amino acid sequence of a protein. The essential problem is how a code of four different bases in nucleic acids can specify proteins made up of 20 different types of amino acids. The solution is that each amino acid is encoded by a set of three consecutive bases. The three-base sets in RNA are called **codons**, and genetic code tables conventionally give the genetic code in terms of RNA codons. The standard genetic code is given in Table 1.1. From this table you can see that the genetic code is **degenerate**, in that most amino acids can be specified by more than one codon. The degeneracy of the genetic code means that you can deduce the protein sequence from a DNA or RNA sequence, but you cannot unambiguously deduce a nucleic acid sequence from a protein sequence.

Table 1.1
Standard genetic code. The corresponding amino acid is given next to each codon. The three-letter amino acid code defined in Table 2.1 is used.

First letter of the codon		Second letter of the codon								Third letter of the codon
5′ end	U		C		A		G			3′ end
U	UUU	Phe	UCU	Ser	UAU	Tyr	UGU	Cys		U
	UUC	Phe	UCC	Ser	UAC	Tyr	UGC	Cys		C
	UUA	Leu	UCA	Ser	UAA	Stop	UGA	Stop		A
	UUG	Leu	UCG	Ser	UAG	Stop	UGG	Trp		G
C	CUU	Leu	CCU	Pro	CAU	His	CGU	Arg		U
	CUC	Leu	CCC	Pro	CAC	His	CGC	Arg		C
	CUA	Leu	CCA	Pro	CAA	Gln	CGA	Arg		A
	CUG	Leu	CCG	Pro	CAG	Gln	CGG	Arg		G
A	AUU	Ile	ACU	Thr	AAU	Asn	AGU	Ser		U
	AUC	Ile	ACC	Thr	AAC	Asn	AGC	Ser		C
	AUA	Ile	ACA	Thr	AAA	Lys	AGA	Arg		A
	AUG	Met	ACG	Thr	AAG	Lys	AGG	Arg		G
G	GUU	Val	GCU	Ala	GAU	Asp	GGU	Gly		U
	GUC	Val	GCC	Ala	GAC	Asp	GGC	Gly		C
	GUA	Val	GCA	Ala	GAA	Glu	GGA	Gly		A
	GUG	Val	GCG	Ala	GAG	Glu	GGG	Gly		G

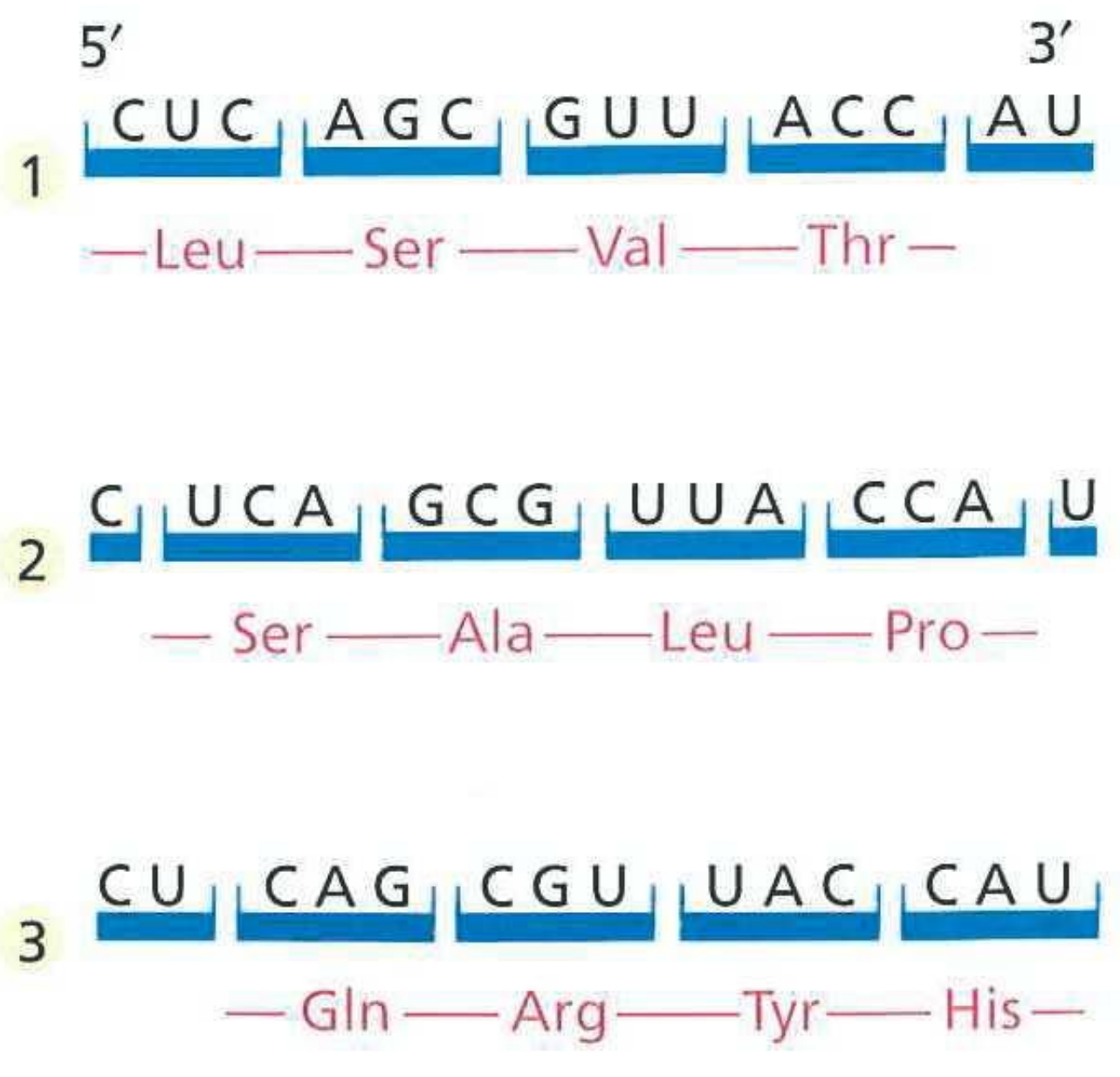


Figure 1.9
The three reading frames of a strand of mRNA. Each reading frame starts one nucleotide further giving rise to a different protein sequence. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

There are three codons that do not encode an amino acid but instead signal the end of a protein-coding sequence. These are generally called the stop codons. The signal to start translating is more complex than a single codon, but in most cases translation starts at an AUG codon, which codes for the amino acid methionine. This initial methionine residue is often subsequently removed from the newly synthesized protein. In general, all life uses the same genetic code, but there are some exceptions and care should be taken to use the appropriate code when deducing amino acid sequences from DNA sequences. The code tables can be accessed through many of the sequence databases and through the National Center for Biotechnology Information (NCBI) Taxonomy section.

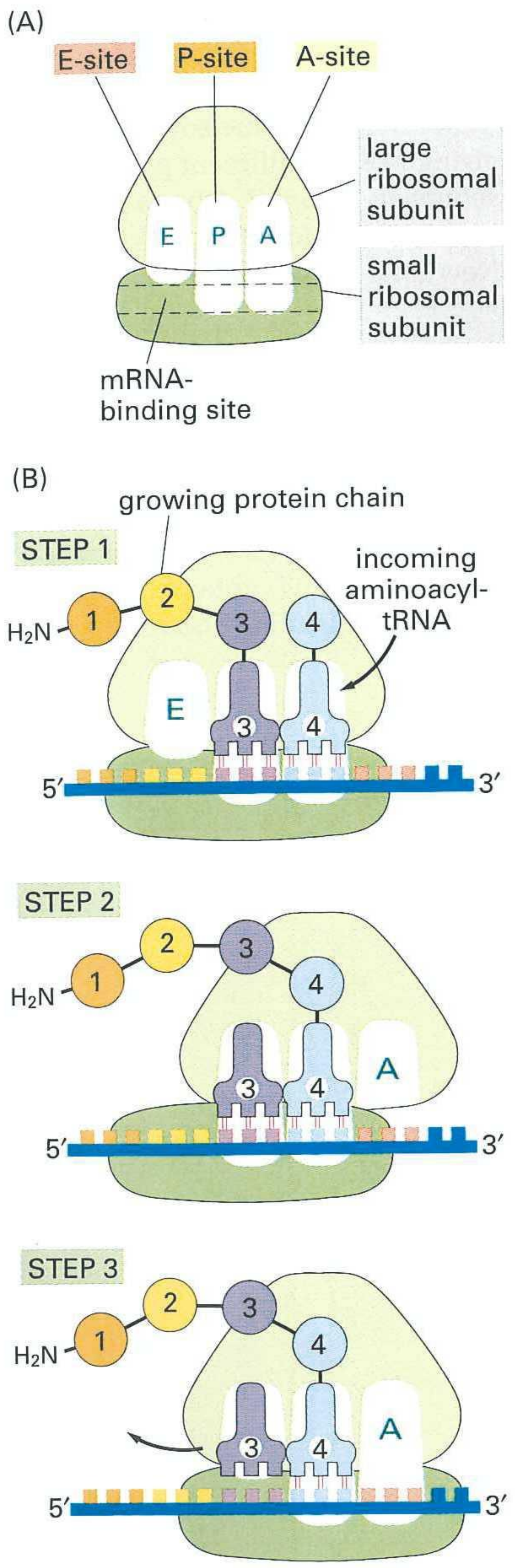
The translation of bases to amino acids occurs in nonoverlapping sets of three bases. There are thus three possible ways to translate any nucleotide sequence, depending on which base is chosen as the start. These three **reading frames** give three different protein sequences (see Figure 1.9). In the actual translation process the detailed control signals ensure that only the appropriate reading frame is translated into protein. When trying to predict protein-coding sequences in DNA sequences, information about the control signals is often lacking so that one needs to try six possible reading frames, three for each DNA strand. Usually, only one of these reading frames will produce a functional protein. Proteins tend (with notable exceptions) to be at least 100 amino acids in length. The incorrect reading frames often have a stop codon, which results in a much shorter translated sequence. When analyzing bacterial genome sequences, for example, to predict protein-coding sequences, reading frames are identified that give a reasonable length of uninterrupted protein code, flanked by appropriate start and stop signals, called **open reading frames** or **ORFs**. Gene prediction is discussed in Chapters 9 and 10.

Translation involves transfer RNAs and RNA-containing ribosomes

RNAs have a variety of roles in cells but are mainly involved in the transfer of information from DNA and the use of this information to manufacture proteins. There are three main classes of RNA in all cells—messenger RNA (mRNA), **ribosomal RNA (rRNA)**, and **transfer RNA (tRNA)**—as well as numerous smaller RNAs with a variety of roles, some of which we will encounter in this book. The role of mRNA has been described above. rRNAs and tRNAs are involved in the process of mRNA translation and protein synthesis.

The mRNA produced by transcription is translated into protein by ribosomes, large multimolecular complexes formed of rRNA and proteins, in a process called **translation**. Ribosomes consist of two main subunit complexes, one of which binds the mRNA (see Figure 1.10A). Amino acids do not recognize the codons in mRNA directly and their addition in the correct order to a new protein chain is mediated by the tRNA molecules, which transfer the amino acid to the growing protein chain when bound to the ribosome. These small tRNA molecules have a three-base **anticodon** at one end that recognizes a codon in mRNA, and at the other end a site to which the corresponding amino acid becomes attached by a specific enzyme. This system is the physical basis for the genetic code. There are different tRNAs corresponding to every

Figure 1.10
Translation. (A) Schematic structure of the ribosome showing the binding site for the mRNA and the three tRNA binding sites. (B) A simplified view of the three steps of translation during which residue 4 is added to the C-terminal end of the growing protein chain. These steps are repeated for every residue in the protein. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)



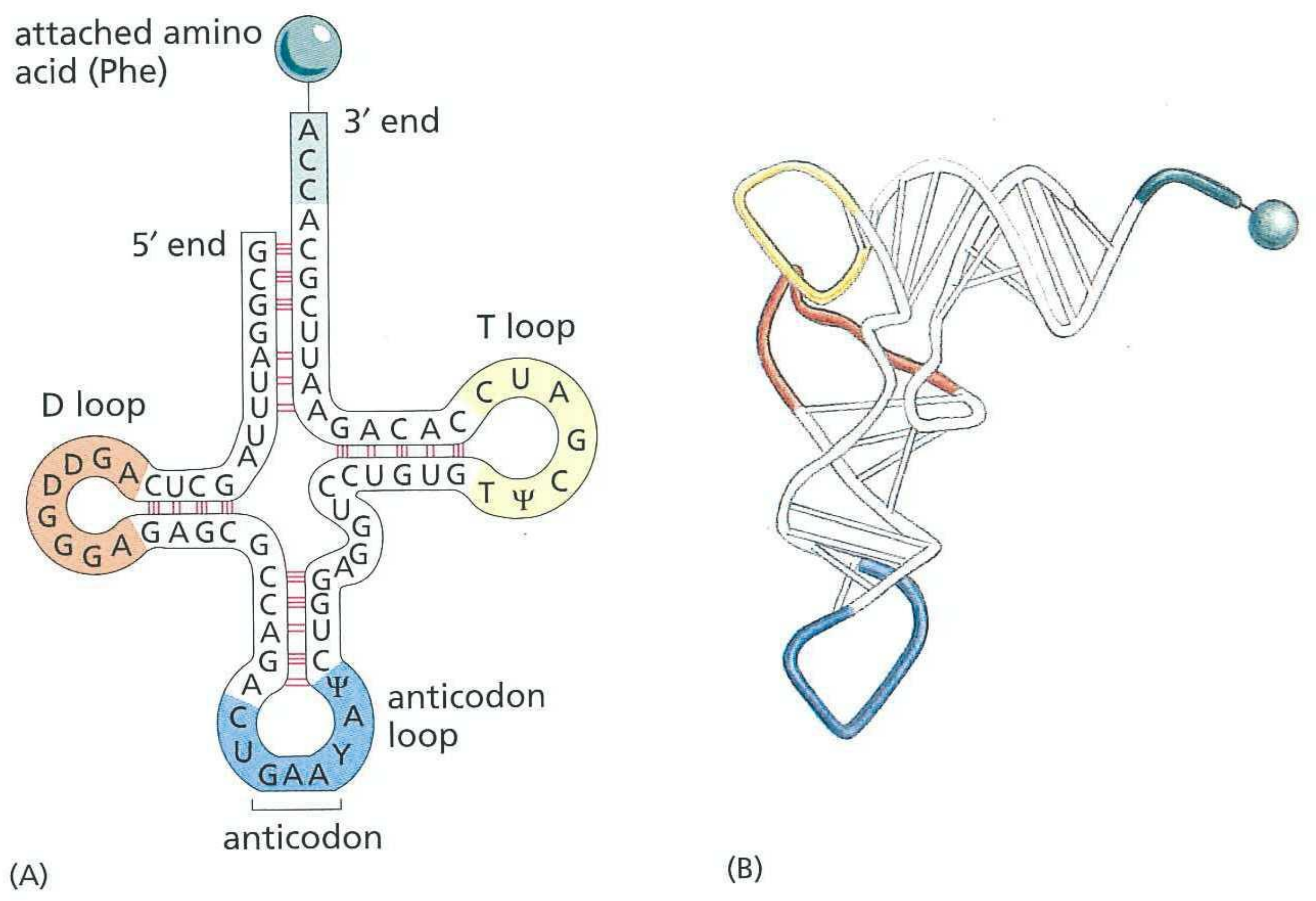
amino acid, but some tRNA anticodons can bind to several similar codons. One common mechanism for this flexibility, which is called **wobble base-pairing**, involves the occurrence of modified bases in the anticodon. The human tRNA set has 48 different anticodons, but some species have less than 40. tRNA is a good example of an RNA molecule with a specific three-dimensional structure that is crucial for its function, and an example structure is shown in Figure 1.11. The identification of tRNA genes is described in Chapter 10, and the generalized sequence and secondary structure of tRNA molecules are shown in Figure 10.2.

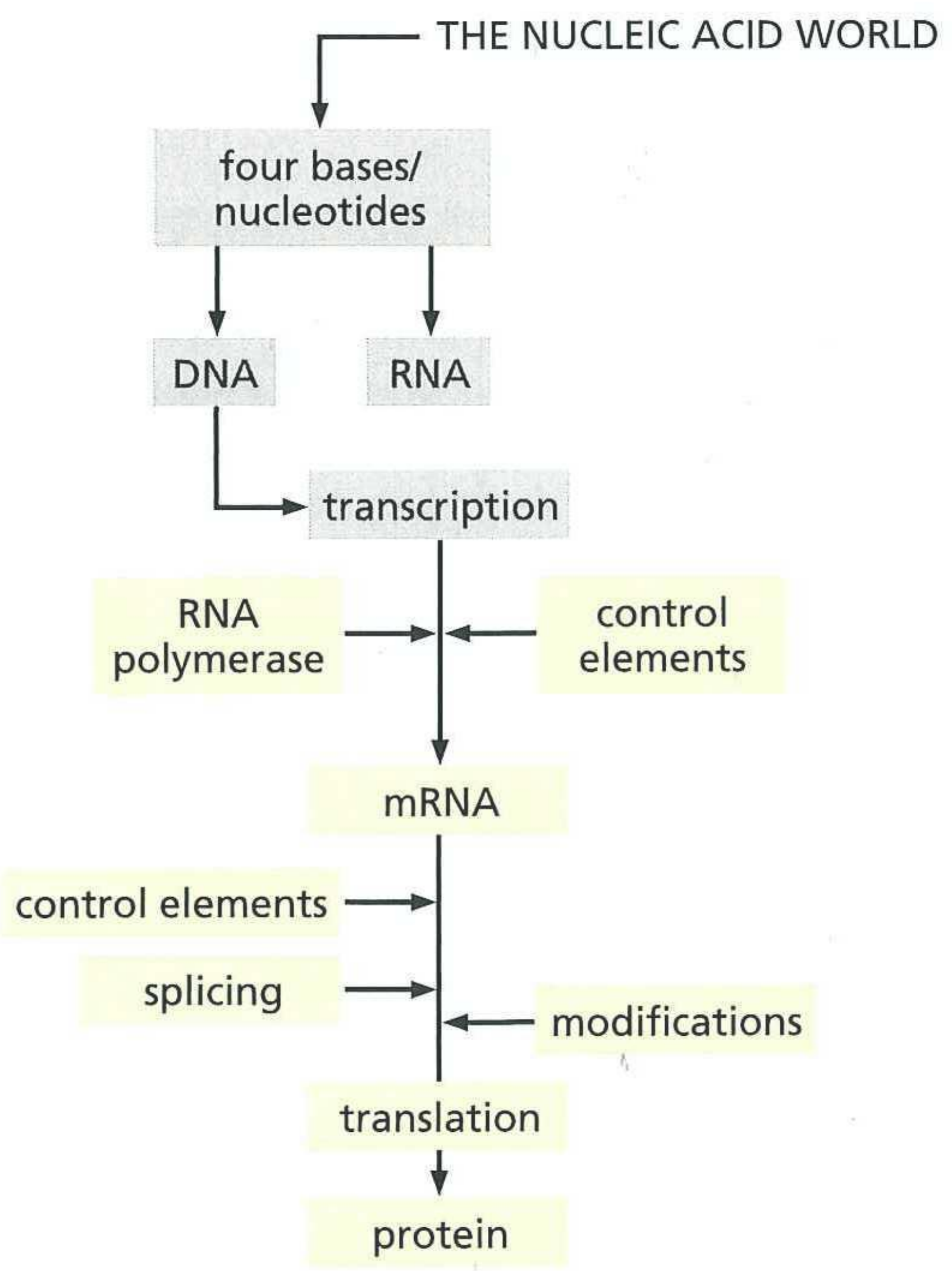
A simplified outline of the process of translation is given in Figure 1.10B and we shall not go into details of it here. The decoding of mRNA starts at its 5′ end, which is the end of the mRNA that is first synthesized at transcription. This is the key justification for the DNA sequence of a gene being conventionally written as the sequence of the sense or coding strand starting at its 5′ end. The site of binding of the ribosome gradually moves along the mRNA toward the 3′ end. The tRNA molecules bind to the ribosome and are the physical link between the mRNA and the growing protein chain. The enzymatic activity in the ribosome that joins amino acids together to form a protein chain is due to the rRNA, not to the ribosomal proteins.

1.3 Gene Structure and Control

The description in the previous section of the details of the central dogma focused almost exclusively on the way the protein sequence information is stored in the genome and its conversion from nucleotides to amino acids. Little attention was paid to the ways in which these processes are controlled. Additionally, there are further complications, especially in **eukaryotes**, whose genes often have a more complicated structure including noncoding regions called introns between protein-coding regions. Expressing such genes involves an extra step in converting the DNA information to proteins, called RNA splicing, in which the mRNA produced initially is modified to remove the introns.

Figure 1.11
The structure of a tRNA molecule. (A) A schematic of the Phe-tRNA molecule showing the arrangement of the base-pairing and loops. (B) The actual three-dimensional structure of the same tRNA molecule colored to show the equivalent regions as in the schematic. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)





Flow Diagram 1.3
The key concept introduced in this section is that the processes of transcription and translation are subject to several distinct and sophisticated control mechanisms.

The regulation of many processes that interpret the information contained in a DNA sequence relies on the presence of short signal sequences in the DNA (see Flow Diagram 1.3). There are many different signal sequences, the general term for which is a **regulatory element**. For example, the molecules involved in transcription and translation require signals to identify where they should start and stop. In addition there are signals that are involved in the control of whether transcription occurs or not. The majority of these regulatory sequences are binding sites for specialized regulatory proteins that interact with the DNA to regulate processes such as transcription, RNA splicing, and translation. In order to interpret a DNA sequence correctly, it is important to have some understanding of the nature of these signals and their roles, although the precise mechanisms by which gene regulatory proteins act are not relevant to this book and are not discussed further. In this section we will survey the main aspects of the control of genes and how control structures occur in gene structure.

There are two distinct classes of organism—**prokaryotes** and eukaryotes—whose translation and transcription processes need to be considered separately. For a description of the general characteristics of prokaryotic and eukaryotic organisms see Section 1.4. We shall briefly review the general types of noncoding regulatory sequences found in prokaryotic and eukaryotic genes and introduce their roles. As prokaryotic control regions are, in general, less complicated than those of eukaryotic genes, we shall use them first to describe the basic signals that direct RNA polymerase to bind and start transcription in the appropriate place.

RNA polymerase binds to specific sequences that position it and identify where to begin transcription

The control regions in DNA at which RNA polymerase binds to initiate transcription are called **promoters**. RNA polymerase binds more tightly to these regions than to the rest of the DNA and this triggers the start of transcription. Additional proteins binding with the polymerase are also required in order to activate transcription. Bacterial promoters typically occur immediately before the position of the **transcription start site** (**TSS**) and contain two characteristic short sequences, or motifs, that are the same or very similar in the promoters for different genes.

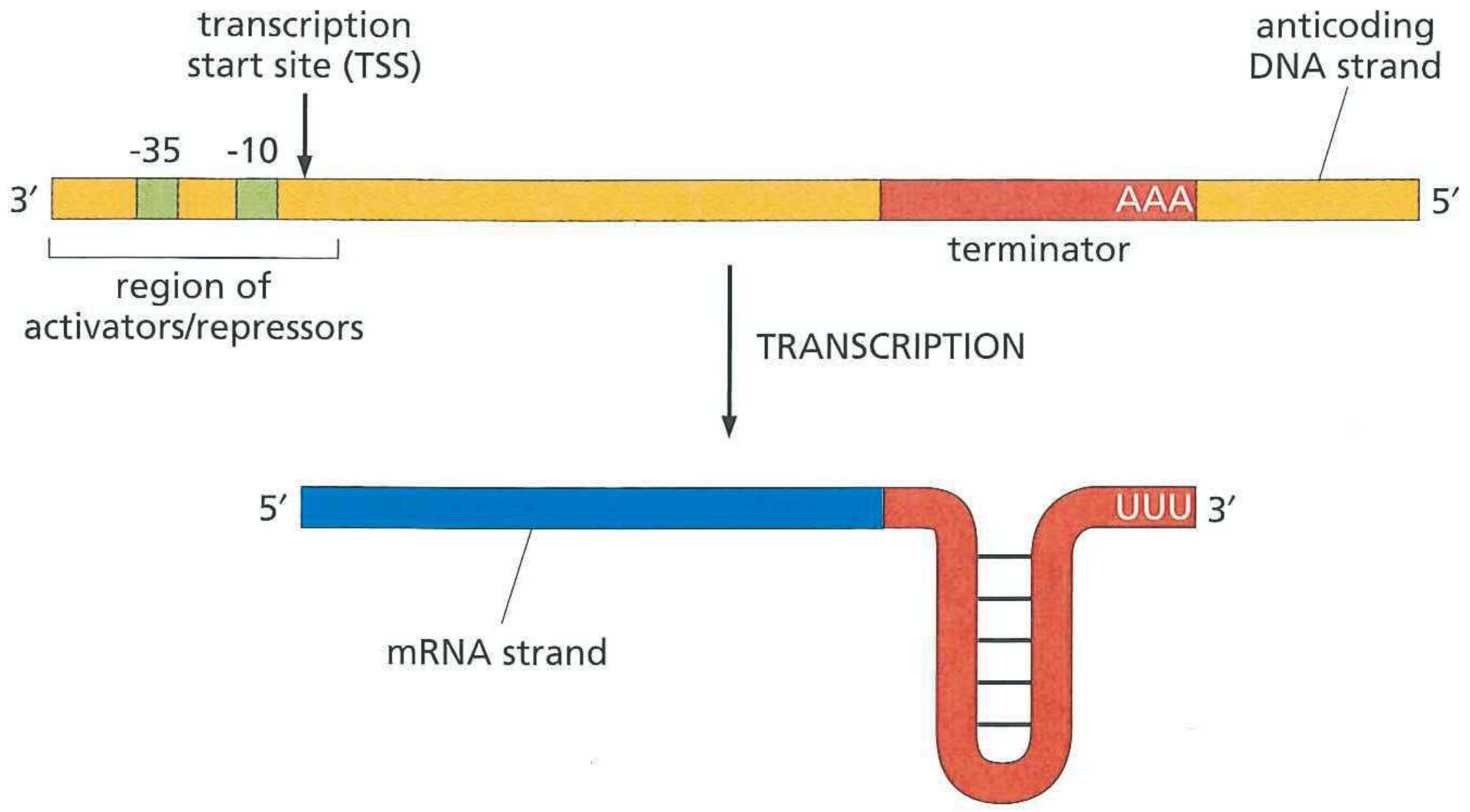


Figure 1.12
The start and stop signals for prokaryotic transcription. The signals to start transcription are short nucleotide sequences that bind transcription enzymes. The signal to stop transcription is a short nucleotide sequence that forms a loop structure preventing the transcription apparatus from continuing.

One of these motifs is centered at approximately 10 bases before the start, conventionally written as –10, and the other at approximately 35 bases before, written as –35 (see Figure 1.12). These two sequences are essential for the tight binding of the RNA polymerase, and they position it at the appropriate location and in the correct orientation to start transcription on the right strand and in the right direction. Sequences located before the start point of transcription are commonly called **upstream sequences**. Sequences located after the start point of transcription are called **downstream sequences**. This terminology is easy to remember if one thinks of the direction of transcription as the flow of a river.

One of the problems in finding promoters in DNA sequences is that the sequence outside the particular conserved motifs varies considerably, and even the motifs vary somewhat from gene to gene. Motifs like these are often described in terms of their **consensus sequence,** which is made up of the bases that are most frequently found at each position. In *Escherichia coli*, for example, the consensus sequence for the –10 motif is TATAAT, and that of the –35 motif is TTGACA. Furthermore, the separation between these two motifs can be between 15 and 19 bases. Note that by convention these are the sequences found on the coding DNA strand, whereas in reality the polymerase binds to double-stranded DNA, which also contains the complementary anticoding strand sequence. RNA polymerase binds to a variety of sequences but binds more strongly the closer the promoter is to a consensus sequence. The tighter the binding, the more frequently the region will be transcribed. Such variation in sequence is a common feature of many control sites, and makes it harder to locate them.

The termination of transcription is also controlled by sequence signals. In bacteria the **terminator signal** is distinct from the promoter in that it is active when transcribed to form the end of the mRNA strand. In the mRNA the terminator sequence produces two short stretches of complementary sequence that can base-pair to form an RNA double helix, usually involving at least four to five CG base pairs; this is followed, usually within five bases, by at least three consecutive U nucleotides (see Figure 1.12). The prokaryotic terminator sequence is more variable than the promoter sequence, and is usually not included in genome annotations, as described in Chapter 9.

In addition to the promoter sequences, many bacterial genes have additional controls, including binding sites for proteins other than the RNA polymerase. Some of these proteins improve the efficiency of correct binding of the RNA polymerase and are called **activators**, while others called **repressors** block the promoter sites and thus inhibit the expression of the gene. These additional repressor and activator proteins have a profound influence on whether, and when, transcription actually occurs, and so are of crucial biological importance. In this book, however, we shall only be concerned with locating genes, not in trying to dissect their higher-level control, so this aspect of gene control will not be discussed in any detail.

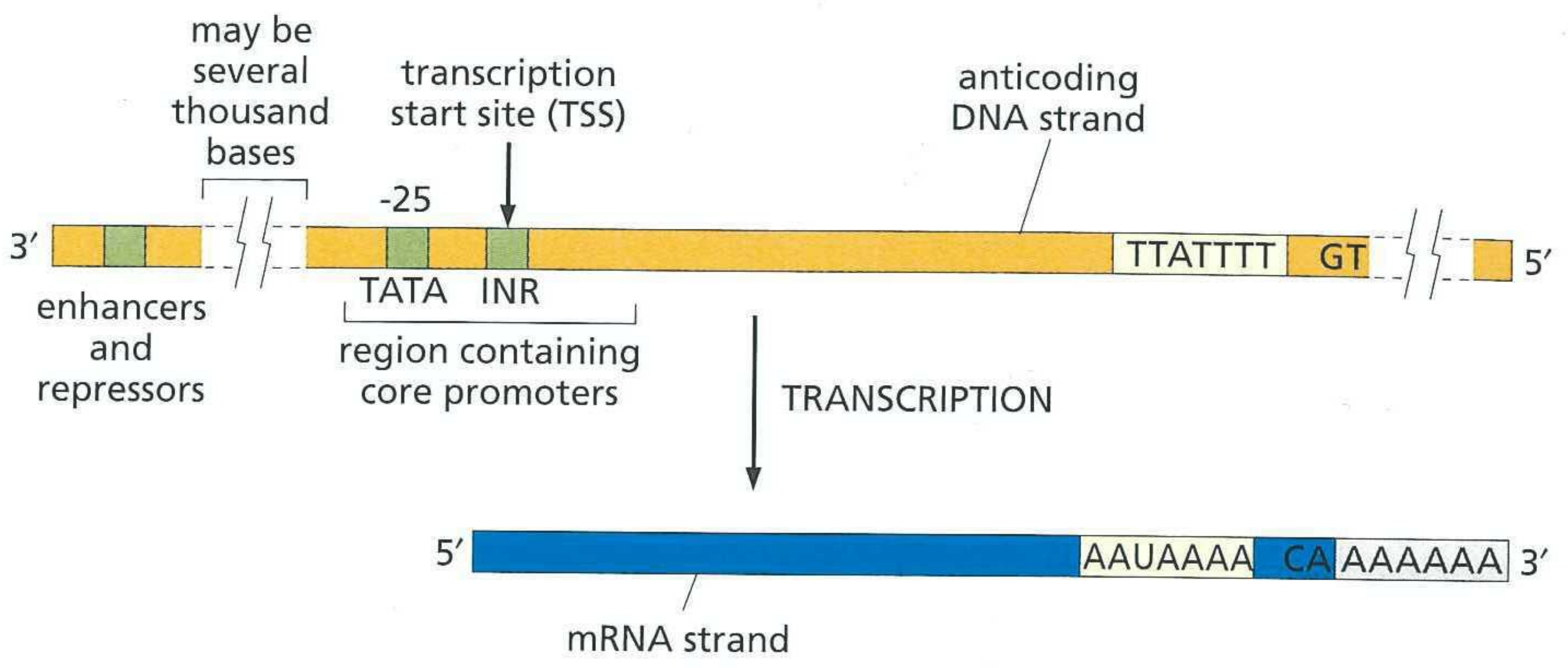


Figure 1.13
The start and stop signals for eukaryotic transcription. All the signals are short sequences that bind enzymes involved in this complex process.

The signals initiating transcription in eukaryotes are generally more complex than those in bacteria

In bacteria all genes are transcribed by a single type of RNA polymerase, whereas in eukaryotes there are three different RNA polymerases, each of which transcribes a different type of gene. All the mRNA that codes for proteins is produced by RNA polymerase II, and in this and later chapters we shall focus mainly on this class of genes. The other RNA polymerases are concerned with transcribing genes for tRNAs, rRNAs, and other types of RNA and these have different types of promoters from genes transcribed by RNA polymerase II.

There are differences in the mechanisms for initiating transcription in eukaryotes compared with bacteria, but the principles are the same. Perhaps the most important difference is the much greater degree of variation in the set of promoters present to initiate transcription. There is a set of **core promoter** DNA sequence signals located in the region of the transcription start site (see Figure 1.13) that are initially bound by a complex of proteins known as **general transcription initiation factors**, which in turn recruit the RNA polymerase to the DNA in the correct position to start transcription. The most important core promoter sequence in genes transcribed by RNA polymerase II is about 25 nucleotides upstream from the start of transcription; this is called the **TATA box** and is characterized by a TATA sequence motif. Details of the sequence variation of this promoter are given in Figure 10.12A, but it should also be noted that this signal is not present in all eukaryotic genes. The transcription factor that binds this motif is called the TATA-binding protein (TBP). Many other protein components are involved in initiating and regulating RNA polymerase activity, but any given gene will only require a small subset for activation, and thus only have a subset of the promoter signals. There appears to be no promoter ubiquitous in eukaryotic genes. In further contrast to the situation in prokaryotes, some of the eukaryotic protein-binding sites that modify RNA polymerase activity can be more than a thousand bases away from the promoter. It is thought that the intervening DNA loops round so that all the gene regulatory proteins come together in a complex that regulates the activity of the RNA polymerase and determines whether or not it starts transcription.

Although termination signals have been identified for both RNA polymerases I and III, no specific signal has been identified for RNA polymerase II. However, it may yet remain to be identified because following transcription the mRNA transcript contains a AAUAAA sequence signal that results in cleavage of the 3′ end of the transcript at a site some 10–30 bases after the signal and addition of a series of adenosine nucleotides to form a new 3′ end. This occurs very quickly and removes information about any signal which may be present to terminate transcription. In eukaryotes, this initial mRNA transcript is further modified before translation, as will be described below.

Eukaryotic mRNA transcripts undergo several modifications prior to their use in translation

The major difference between eukaryotes and prokaryotes in terms of their transcription and translation processes is that the eukaryotic mRNA transcripts are substantially modified before translation. Two of these modifications have no effect on the final protein product. The first modification, which occurs whilst transcription is still in progress, involves the addition of a modified guanosine nucleotide (7-methylguanosine) to the 5′ end of the transcript, a process called **RNA capping**. The last modification, which also occurs while transcription continues, is that which produces the 3′ end of the transcript as mentioned previously; this modification consists of two separate steps. The first step is the cleavage of the mRNA transcript after a CA sequence. The second step, called **polyadenylation**, results in approximately 200 adenosine nucleotides being added to the 3′ end.

The other mRNA modification that occurs in eukaryotes has a significant effect on the final protein products. The major structural difference between the protein-coding genes of prokaryotes and those of eukaryotes is that the protein-coding DNA of most plant and animal genes is interrupted by stretches of noncoding DNA called **introns**; the blocks of protein-coding sequence between the introns are known as **exons** (see Figure 1.14). Introns have lengths which vary from 10 to over 100,000 nucleotides, whereas exons tend to an average length of 100–200 necleotides and rarely exceed 1000 nucleotides. Most bacterial protein-coding genes, on the other hand, have an uninterrupted coding sequence. Introns are found in the genes of most eukaryotes but are less frequent in some genomes, for example that of the yeast *Saccharomyces cerevisiae*. They occur very rarely in prokaryotes.

The existence of introns necessitates an extra step between transcription and translation in eukaryotes, which is known as **RNA splicing**. The complete gene is initially transcribed into RNA; the introns are then excised and the exons joined, or spliced, together to provide a functional mRNA that will give the correct protein sequence when translated (Figure 1.14). In most protein-coding genes, RNA splicing is carried out by a complex called a spliceosome, which consists of small nuclear RNA molecules (snRNAs) and proteins. This complex contains the enzymatic activity that cleaves and rejoins the RNA transcript. The excised intron forms a circular structure called a lariat with a branching point usually at an adenine base (Figure 1.14). The lariat RNA is subsequently degraded. There are particular sequence motifs present at the sites at which the RNA transcript is spliced, as well as the position which will become the lariat branch point. However, these sequence motifs show considerable variability, with the exception of the first and last two bases of each intron. In most cases these are GU and AG as shown in Figure 1.14, but a few instances of another signal, namely AU and AC, have been found in some complex eukaryotes. These are known as AT–AC or U12 introns, after the DNA sequences or one of the components of the spliceosome, respectively. In some even rarer cases the intron RNA itself has splicing activity.

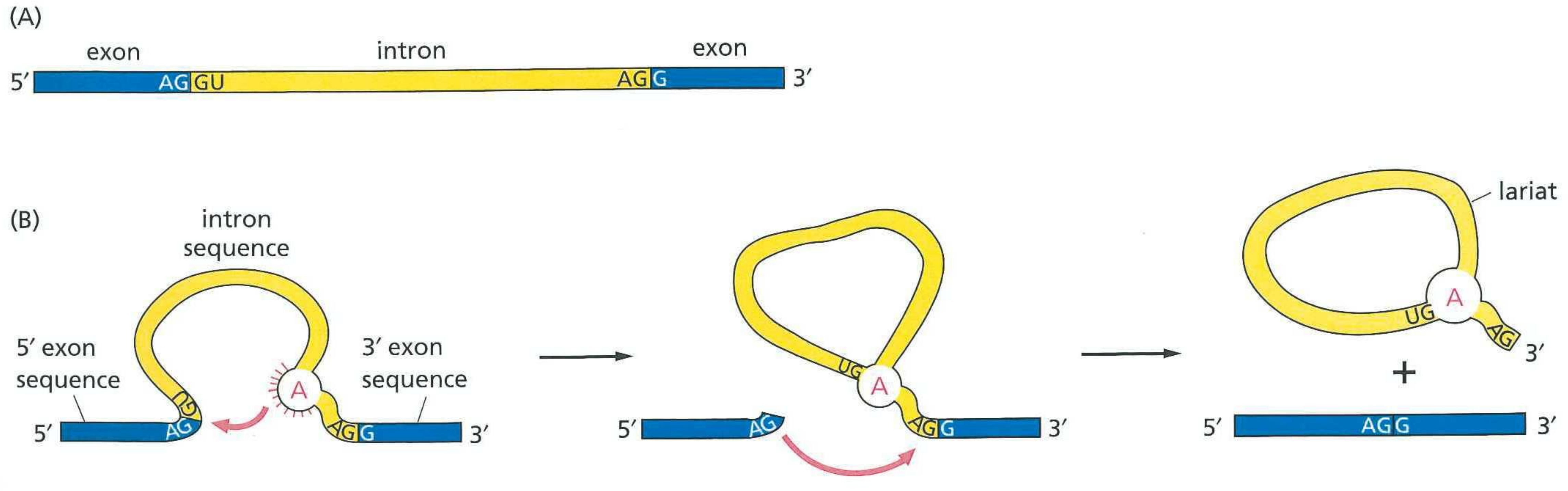


Figure 1.14
The simple schematic of the splicing of an intron. (A) A linear schematic that shows a segment of pre-mRNA with an intron in yellow. The donor splice site has the conserved dinucleotide GU at the start of the intron. The acceptor splice site has the conserved dinucleotide AG at the end of the intron. (B) The intron is spliced out in a two-stage process: firstly creating a loop structure, the lariat, involving the adenine base (colored red), followed by joining the two exons together releasing the intron. In this case the intron occurred within the codon AGG, which is formed when the exons are spliced together. (B, from B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

It should be noted that although protein-coding sequences require a whole number of codons to encode the protein sequence, and thus are a multiple of three bases in length, individual exons do not have this requirement. Codons can be derived from the spliced ends of two consecutive exons, as shown in Figure 1.14. This can lead to further complications in gene prediction, as it adds to the difficulty of correctly identifying the exons and from them the protein sequence. For this and other reasons the possible existence of introns in eukaryotic genes significantly complicates the process of gene prediction in eukaryotic compared with prokaryotic genomes, as is discussed in more detail in Chapters 9 and 10.

Although usually all the exons are retained and joined together during RNA splicing, there are cases where this does not happen and **alternative splicing** occurs, excluding some exons or parts of exons, and thus producing different versions of a protein from the same gene. Alternative splicing is quite common in the genes of humans and other mammals, and is thought to be one means by which the relatively small number of genes present in the genome can specify a much greater number of proteins.

The control of translation

There are various sequence motifs in the mRNA transcripts that indicate the beginning and end of a protein-coding sequence and the site at which the mRNA initially binds to a ribosome. Most protein-coding sequences start with a methionine codon, typically AUG, and invariably end with one of the stop codons of the genetic code (see Table 1.1). In bacterial DNA there is also a distinct short sequence at the 5′ end of the mRNA known as the **Shine–Dalgarno sequence** that indicates the ribosome-binding site. This has a typical consensus sequence of AGGAGGU and occurs a few bases upstream of the AUG translation start codon.

Eukaryotes do not use a Shine–Dalgarno sequence to position the ribosome, but instead have components that bind specifically to the 7-methylguanosine nucleotide at the 5′ end of all eukaryotic mRNA transcripts. The ribosome binds to this and then starts to search for an AUG codon. Occasionally, the first AUG is missed and another downstream codon is used instead. Termination of translation occurs on encountering a stop codon as in prokaryotes.

There is one feature of gene organization in bacteria that is rarely found in eukaryotes, and which relates to the way the ribosome binds near the translation start site. Functionally related protein-coding sequences are often clustered together into **operons** (see Figure 1.15). Each operon is transcribed as a single mRNA transcript and the proteins are then separately translated from this one long molecule. This has the advantage that only one control region is required to activate the simultaneous

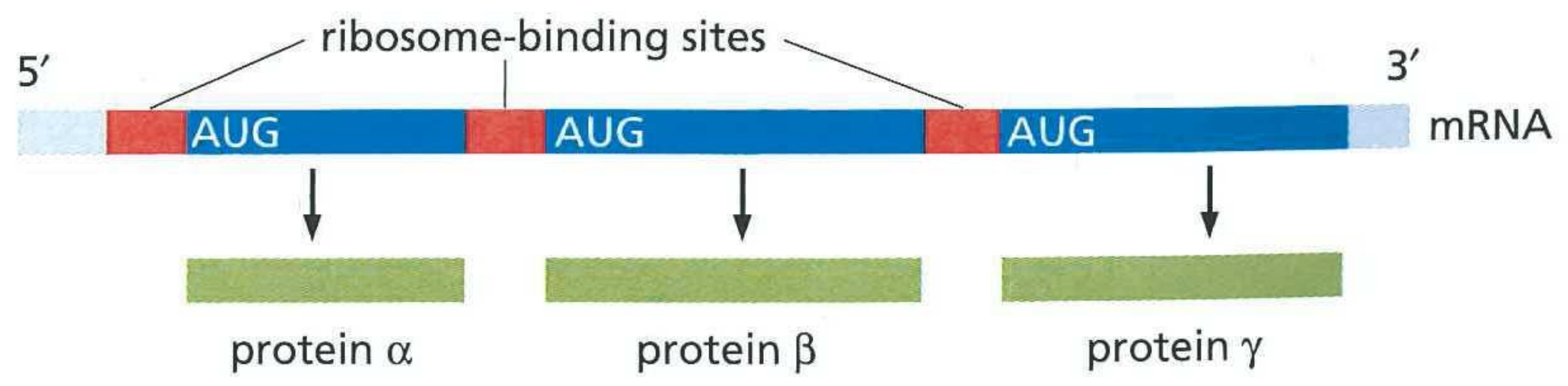
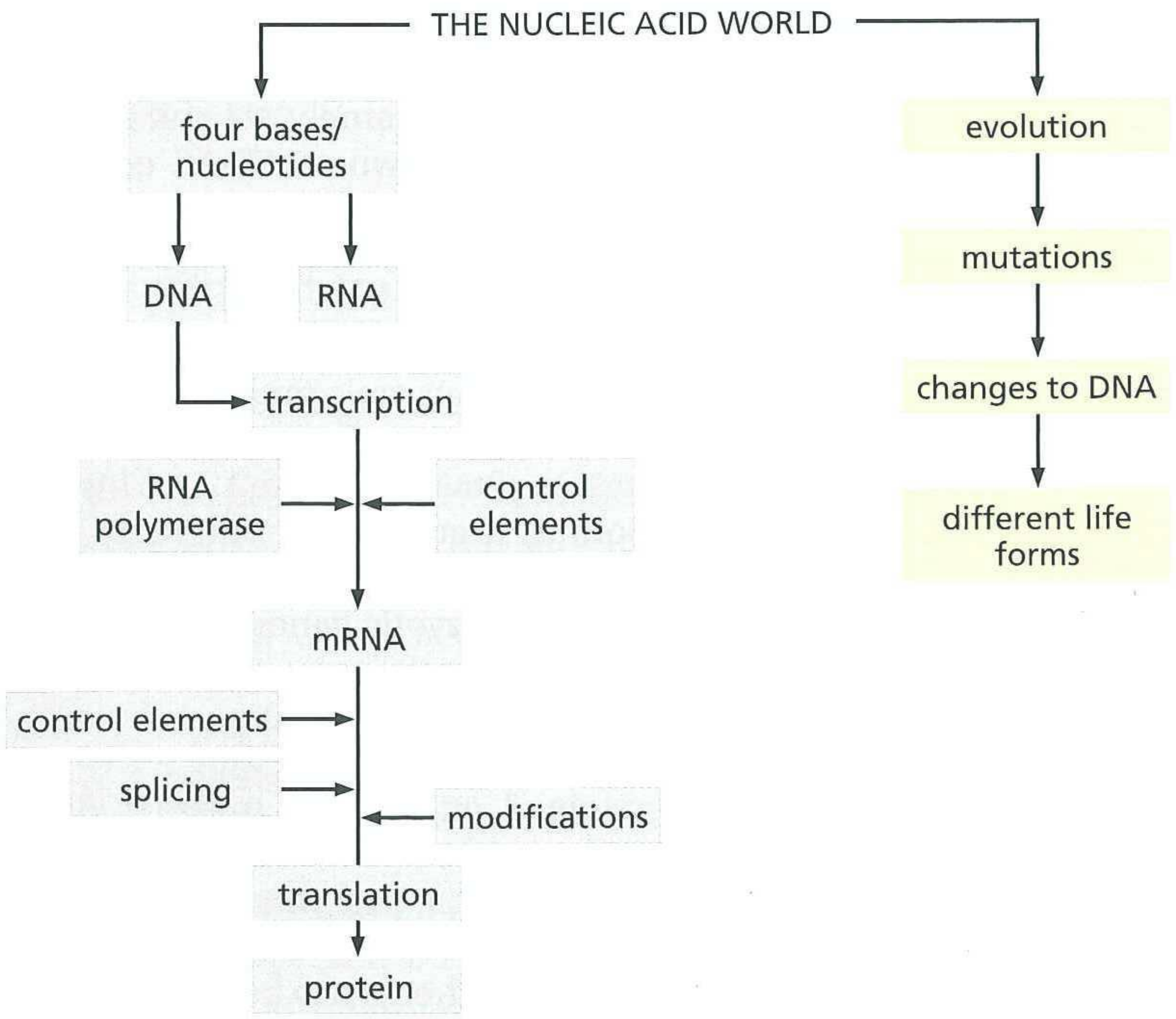


Figure 1.15
A schematic of operon structure. This single mRNA molecule contains three separate protein-coding regions that produce proteins α, β, and γ on translation. The mRNA contains three separate ribosome-binding sites (red).

Flow Diagram 1.4
The key concept introduced in this section is that the process of evolution is based on mutations of the DNA that result in different life forms, which may then be subjected to different evolutionary selective pressures.



expression of, say, all the enzymes required for a particular metabolic pathway. Not all bacterial genes are contained in operons; many are transcribed individually and have their own control regions.

1.4 The Tree of Life and Evolution

The integrity of the genetic information in DNA is carefully protected, and without this protection life would soon disintegrate. Nevertheless, errors inevitably arise in the genome, and these are extremely important as they also provide the genetic variation on which natural selection and evolution can act. Over very long periods of time some of those changes that do not prove to cause their carriers a disadvantage are likely to spread and eventually occur in all genomes of the species (see Flow Diagram 1.4). In this way species can evolve, and ultimately they can evolve into entirely new species. It is generally thought that all existing life has evolved from a single common, very distant ancestor. The evolutionary relationship of known species to each other is commonly described as the tree of life (see Figure 1.16). In this section we will briefly describe the most basic divisions of the tree of life, and some of the modifications which are seen in genomes and their consequences.

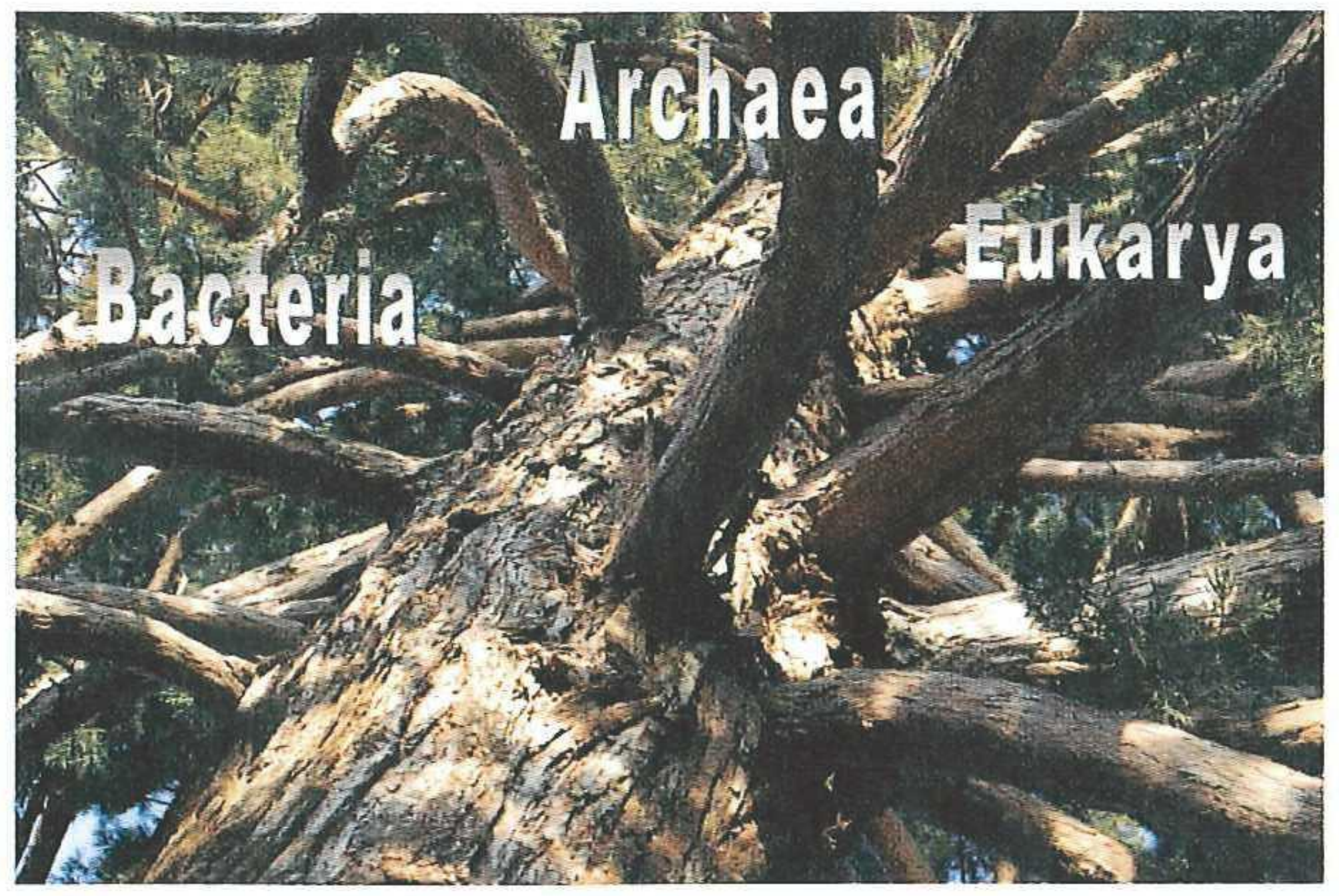


Figure 1.16
Tree of life. Evolution branches out like a real tree where the roots are the origin and the branches are the different groups of life form.

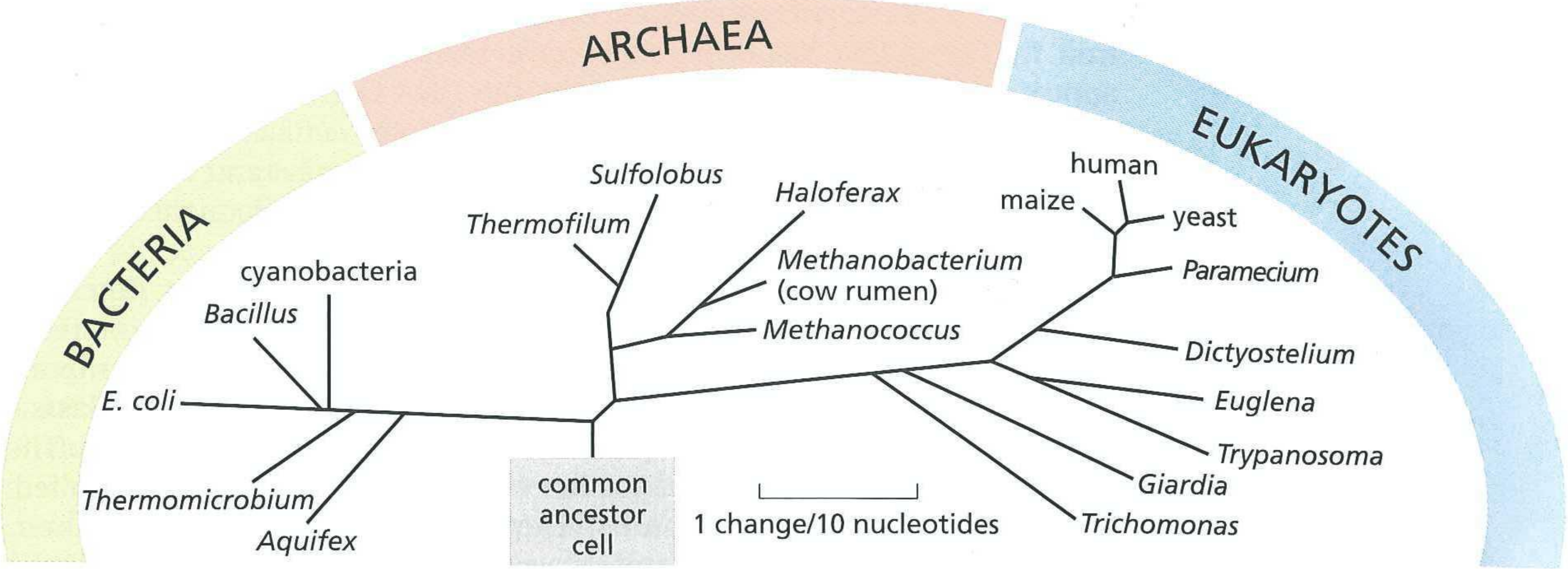


Figure 1.17
The tree of life based on analysis of molecular sequences. Note that multicellular organisms are all in a small region of the tree. (From B. Alberts et al., Molecular Biology of the Cell, 4th ed. New York: Garland Science, 2002.)

A brief survey of the basic characteristics of the major forms of life

In this context, all living organisms can be divided into two vast groups: the prokaryotes, which are further divided into the unicellular bacteria and archaea, and the eukaryotes, which include all other living organisms (see Figure 1.17). Another class of objects that contain nucleic acid instructions for their reproduction is the viruses. These have very small genomes that encode the proteins that make up the virus structure, but viruses can only replicate inside a living cell of another organism, as they hijack the cell's biochemical machinery for replicating DNA and synthesizing proteins. Viruses may have either DNA or RNA genomes. Although viral genes follow the basic rules by which DNA encodes RNAs and proteins, it is worth noting that some viral genomes have unusual features not commonly present in cellular genomes, such as overlapping genes, which need careful interpretation.

The prokaryotes are a vast group of unicellular microorganisms. Their cells are simple in structure, lacking a nucleus and other intracellular organelles such as mitochondria and chloroplasts. Taxonomically, the prokaryotes comprise two **superkingdoms**, or **domains**, called the Bacteria and the Archaea, which in evolutionary terms are as distinct from each other as both are from the rest of the living world. Their DNA is usually in the form of a single circular chromosome (although linear chromosomes are known in prokaryotes), containing a single circular DNA molecule, and is not enclosed in a nucleus. In favorable growing conditions, prokaryotes reproduce rapidly by simple cell division, replicating their DNA at each division so that each new cell receives a complete set of genetic instructions. Many bacteria also contain extrachromosomal DNAs in the form of **plasmids**, small circular DNAs that carry additional genes and which can often be transmitted from bacterium to bacterium. Genes for drug resistance, the ability to utilize unusual compounds, or the ability to cause disease are often carried on plasmids. The best-studied bacterium, and the one that for many years provided virtually all our knowledge about the processes of transcription and translation, is the gut bacterium *Escherichia coli*, abbreviated to *E. coli.*

All other living organisms are eukaryotes and belong to the domain **Eukarya**. All animals, plants, fungi, algae, and protozoa are eukaryotes. The eukaryotes include both multicellular and unicellular organisms. Unicellular eukaryotes widely used as model organisms for genetic and genomic studies are the yeasts and unicellular algae such as *Chlamydomonas*. Eukaryotic cells are larger and more complicated than those of prokaryotes. The DNA is contained inside a nucleus, and is highly packaged with histones and other proteins into a number of linear chromosomes, ranging from two to hundreds depending on the organism. Humans have 46 chromosomes in their body cells (made up of two sets of 23 chromosomes inherited from each parent), the fruit fly *Drosophila* has 8, petunias have 14, while the king crab has 208. There appears to be no particular reason why the DNA is divided up into such different numbers of chromosomes in different organisms; the actual numbers of genes in the genomes of multicellular organisms are much more similar and vary between 20,000 and 30,000 for those organisms whose genomes have been sequenced to date.

Eukaryotic cells are highly compartmentalized, with different functions being carried out in specialized organelles. Two of these are of particular interest here, as they contain their own small genomes. **Mitochondria** contain the components for the process of energy generation by aerobic respiration, and **chloroplasts** contain the molecular components for photosynthesis in plant and algal cells. These two organelles are believed to be the relics of prokaryotic organisms engulfed by the ancestors of eukaryotic cells and still retain small DNA genomes of their own—mitochondrial DNA (mitDNA) and chloroplast DNA—and the protein machinery to transcribe and translate them. These genomes encode some of the proteins specific to mitochondria and chloroplasts, but most of their proteins are now encoded by genes in the eukaryotic cell nucleus.

Nucleic acid sequences can change as a result of mutation

There are a number of occasions, such as DNA replication, when the genomic DNA is actively involved in processes that leave it vulnerable to damage. Sometimes this damage will be on a large scale, such as the duplication of whole genes, but often it involves just a single base being incorrectly replicated. The general term used to describe such damage is mutation. Depending on which part of the DNA sequence is affected, mutations can have a drastic effect on the information encoded, leading to changes in the sequence of encoded proteins, or the loss of control signals. Genes can be rendered inactive or proteins dysfunctional, although mutations can also have beneficial effects (see Box 1.1). In organisms that use sexual reproduction, unless the DNA affected is in a germ cell, the DNA will not be used to generate the genomes of future generations, and so will only affect the organism in which the damage occurred. In such cases, the organism might suffer, especially if the mutation causes uncontrolled cell growth and division, as happens in tumors. The alternative is that the mutation is transmitted through to the next generation, in which case it has a chance to eventually become part of the normal genome of the species.

Box 1.1 There is more to genomes than protein and RNA genes

Not all the DNA sequence in a genome contains a meaningful message or a known function. Regions without a message are sometimes referred to as junk DNA, although this term should not be taken too literally as we have much still to learn, and these regions may yet come to be seen as functional in some way. Mammalian genomes contain large amounts of this type of DNA, both in the form of introns and between genes. Simpler eukaryotic organisms have less, and bacteria have very little.

Much of the so-called junk DNA is in the form of highly repeated DNA sequences, which form significant percentages of the genomes of many higher organisms. Many of these **DNA repeats** are due to DNA elements known as transposons, which can copy themselves and insert themselves into other parts of the genome. Transposons are present in both bacteria and eukaryotes, but in mammalian genomes they have multiplied to a much greater extent and appear to have largely lost their function, existing now simply as apparently functionless sequence repeats.

On a final note, changes in DNA sequences that occur during evolution occasionally destroy some of the control sequences needed for a gene to be expressed. When this happens, the resultant inactive gene is called a **pseudogene**. Soon after the initial inactivation, pseudogenes can be hard to distinguish from active genes, but over time they accumulate many more mutations than active genes, and so diverge to (ultimately) random sequence.

The fate of the mutation, to be lost or to be retained, depends on the process of natural selection that is the cornerstone of the theory of evolution.

The existence of similar DNA and protein sequences in different organisms is a consequence of the process of evolution that has generated the multitude of living organisms from an ancient ancestor held in common. The sequence similarities reveal details of the ways that mutations arise and of the balance of forces which will result in only a small group of mutations being preserved through evolutionary selection. Therefore some details of the mechanism of mutation are of relevance in a number of areas of bioinformatics, including sequence alignment (see especially Sections 4.3 and 5.1) and phylogenetic analysis (see especially Sections 7.2 and 8.1). In the phylogenetic analysis described in Chapters 7 and 8 an attempt is made to reconstruct the evolutionary history of a set of sequences. This requires a detailed model of evolution, which requires a comprehensive understanding of the kinds of mutations that occur, their effects, and the process of natural selection by which they are either accepted or lost.

Summary

We have tried in this chapter to give a brief introduction to the nucleic acids and their role in living systems. We have focused exclusively on the role of nucleic acids in genomes, although a look at any introductory textbook on molecular biology will show that, in addition, single nucleotides play a part in many other processes. We have described how the sequence of DNA can encode proteins, and how simple sequence signals are used to control the interpretation of the genomic DNA. The material is sufficient to allow a novice to appreciate the techniques discussed in this book, particularly those for gene detection. You should be aware, however, that there are probably exceptions to every general statement we have made in this chapter! (See Box 1.2.) For example, under certain circumstances, the codon UGA can code for the unusual amino acid selenocysteine instead of being understood as a stop signal. Many organisms have their own peculiarities, and one should preferably seek out an expert for advice.

Box 1.2 Things are usually not that simple!

This chapter has given a brief introduction to the key role of nucleic acids in the storage, interpretation, and transmission of genetic information. The many genome sequencing projects are producing results at a phenomenal rate, and the techniques of bioinformatics, such as are described in Chapters 9 and 10, are required to identify and characterize the functional components of the genomes. When working on such projects, it must be remembered that the descriptions given in this and the other chapters are general, and wherever possible care must be taken to discover the specific details applicable to the organism of interest.

A further warning is required in that even some of the fundamental concepts described in this chapter are much less well defined than might be supposed. Two brief illustrations of this will be given, dealing with the definition of a gene and of the human genome. Processes such as alternative splicing are making it ever harder to agree on a definition of a gene. The human genome is another concept whose definition is proving harder to agree on than expected. Recent studies have shown that there is much greater variation in the genome in humans than was expected. As well as many small point mutations that were anticipated, it was found that there were a surprising number of large-scale differences between humans. As a result it is no longer clear how to define a fully representative human genome. There are reviews in Further Reading that explain these points in more detail.