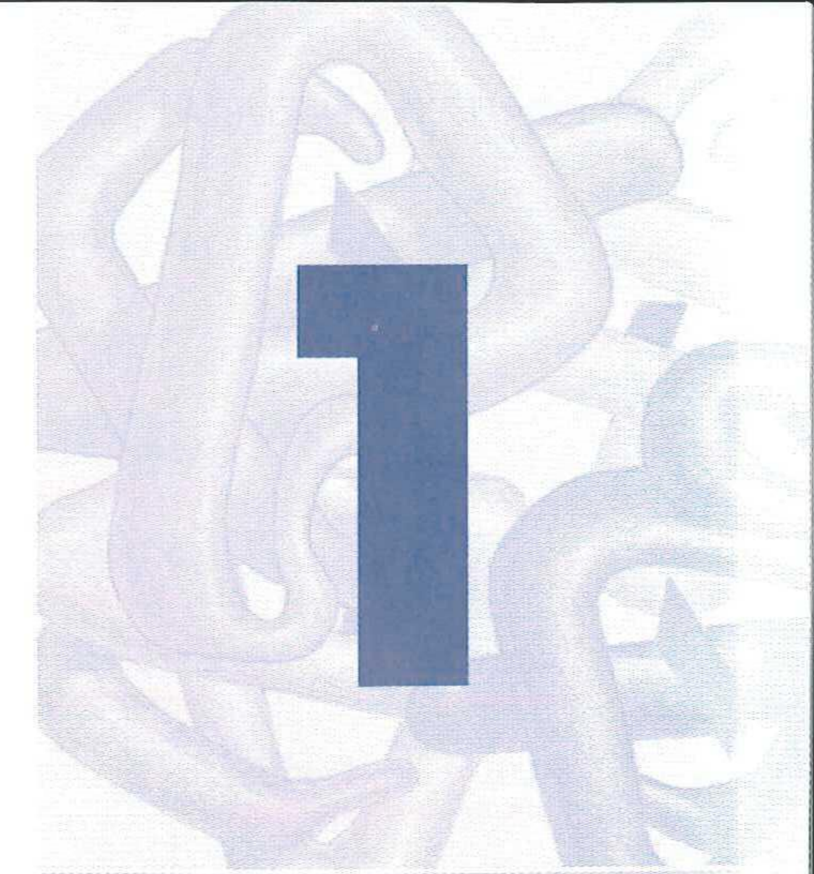


1

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.

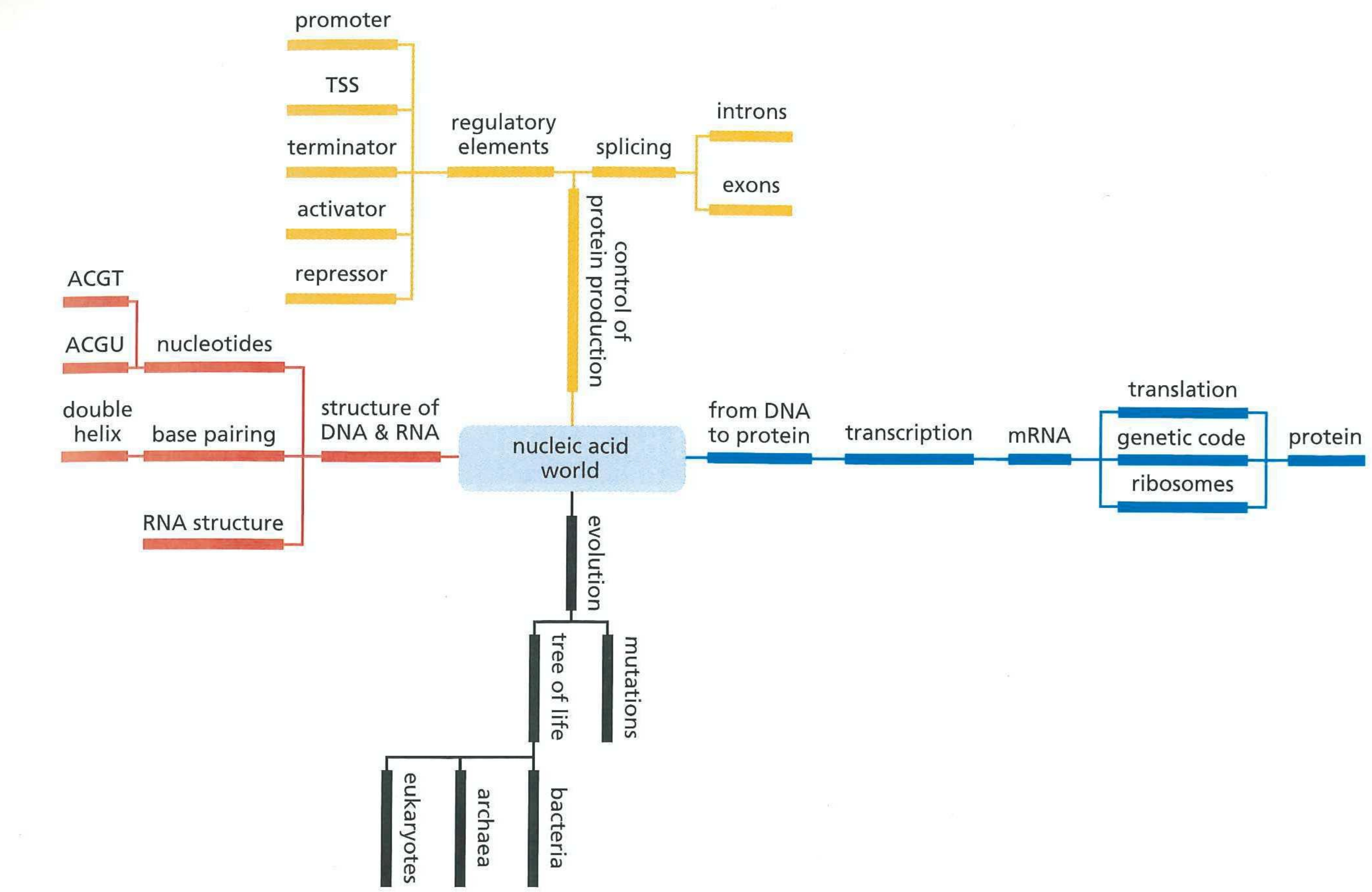
Show that eukaryotic mRNA often has segments (introns) removed before translation.

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

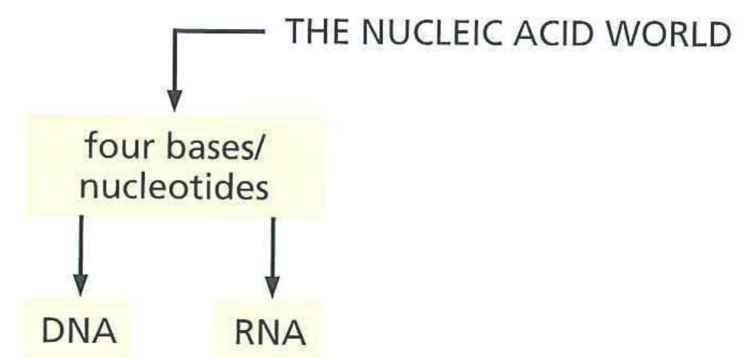
Considering their role as the carrier of genetic information, DNA molecules have a relatively simple chemical structure. They are linear polymers of just four different **nucleotide** building blocks whose differences are restricted to a substructure called the **base** (see Flow Diagram 1.1). For efficient 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.

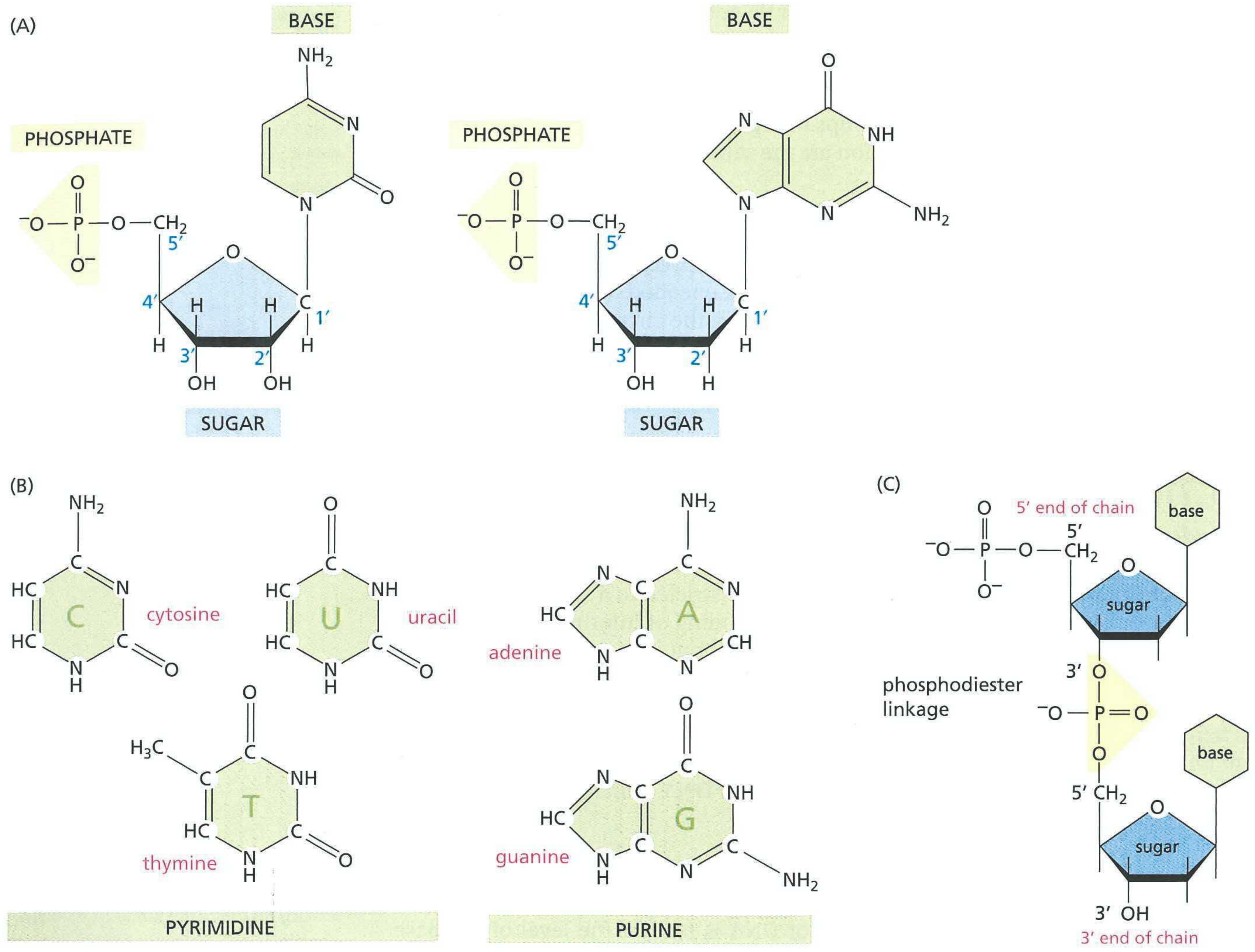


Figure 1.1
The building blocks of DNA and RNA. (A) Left, cytidylate (ribo-CMP); right, deoxyguanylate (dGMP). Each consists of three main parts: a phosphate group, a pentose sugar, and a base. It is the base part that is responsible for the differences in the nucleotides. (B) The bases fall into two 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

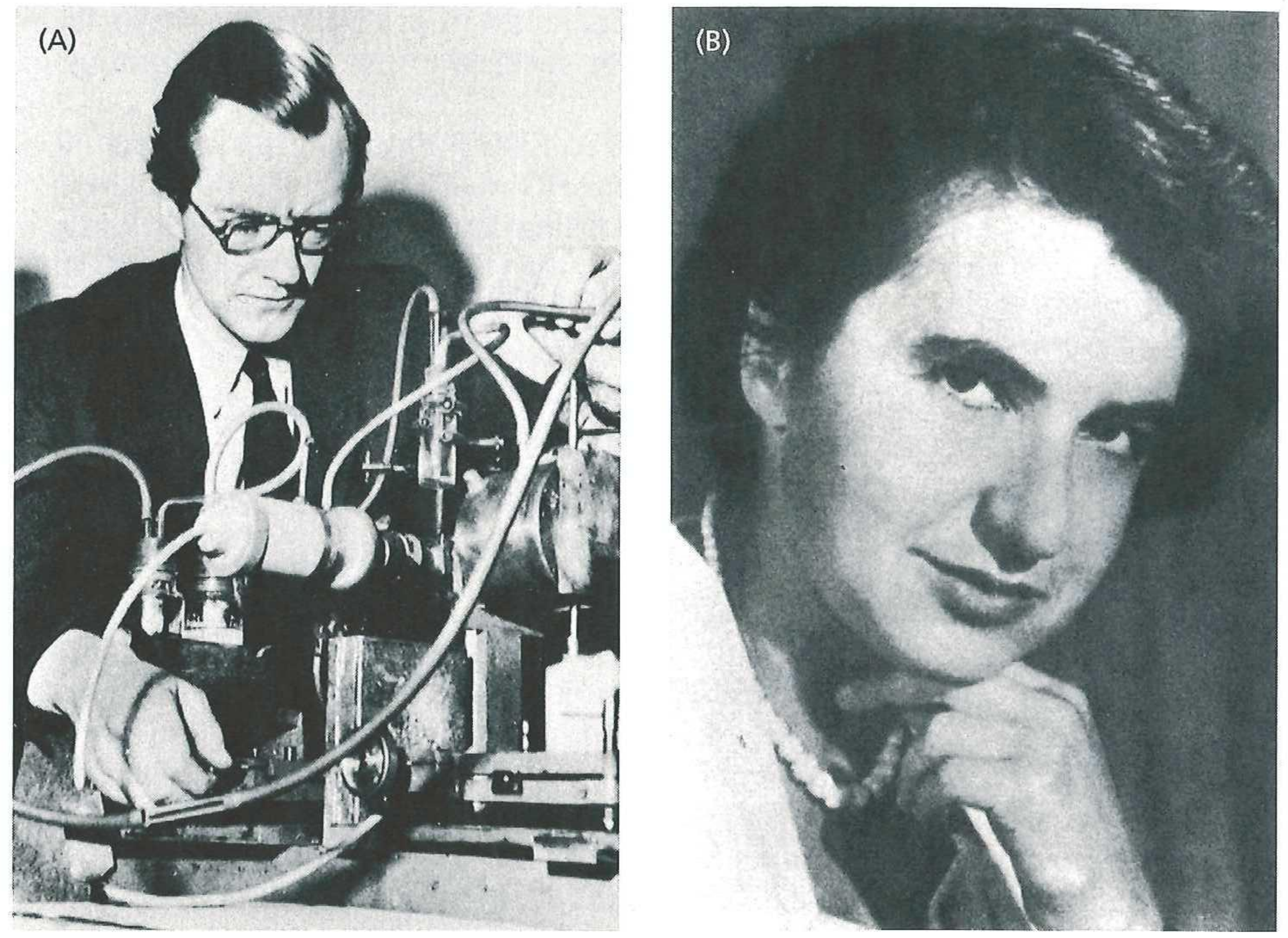


Figure 1.2
Two scientists whose work was influential on James Watson and Francis Crick when they elucidated the structure of DNA.
(A) Maurice Wilkins.
(B) Rosalind Franklin. (A and B courtesy of Science Photo Library.)

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.

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).