Genomes, Transcriptomes, and Proteomes

1

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

Define the terms "genome," "transcriptome," and "proteome," and state how these are linked in the process of genome expression.

Describe the two experiments that led molecular biologists to conclude that genes are made of DNA, and explain the limitations of those experiments.

Give a detailed description of the structure of a polynucleotide, and summarize the chemical differences between DNA and RNA.

Discuss the evidence that Watson and Crick used to deduce the double helix structure of DNA and describe the key features of this structure.

Distinguish between coding and functional RNA and give examples of each type.

Describe in outline how RNA is synthesized and processed in the cell.

Give a detailed description of the various levels of protein structure, and explain why amino acid diversity underlies protein diversity.

Describe the key features of the genetic code.

Explain why the function of a protein is dependent on its amino acid sequence.

List the major roles of proteins in living organisms and relate this diversity to the function of the genome.

Life as we know it is specified by the **genomes** of the myriad organisms with which we share the planet. Every organism possesses a genome that contains the **biological information** needed to construct and maintain a living example of that organism. Most genomes, including the human genome and those of all other cellular life forms, are made of **DNA** (deoxyribonucleic acid) but a few viruses have **RNA** (ribonucleic acid) genomes. DNA and RNA are polymeric molecules made up of chains of monomeric subunits called **nucleotides**.

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

- The **nuclear genome** comprises approximately 3,200,000,000 nucleotides of DNA, divided into 24 linear molecules, the shortest 50,000,000 nucleotides in length and the longest 260,000,000 nucleotides, each contained in a different **chromosome**. These 24 chromosomes consist of 22 **autosomes** and the two sex chromosomes, X and Y. Altogether, some 35,000 **genes** are present in the human nuclear genome.
- The **mitochondrial genome** is a circular DNA molecule of 16,569 nucleotides, multiple copies of which are located in the energy-generating organelles called mitochondria. The human mitochondrial genome contains just 37 genes.

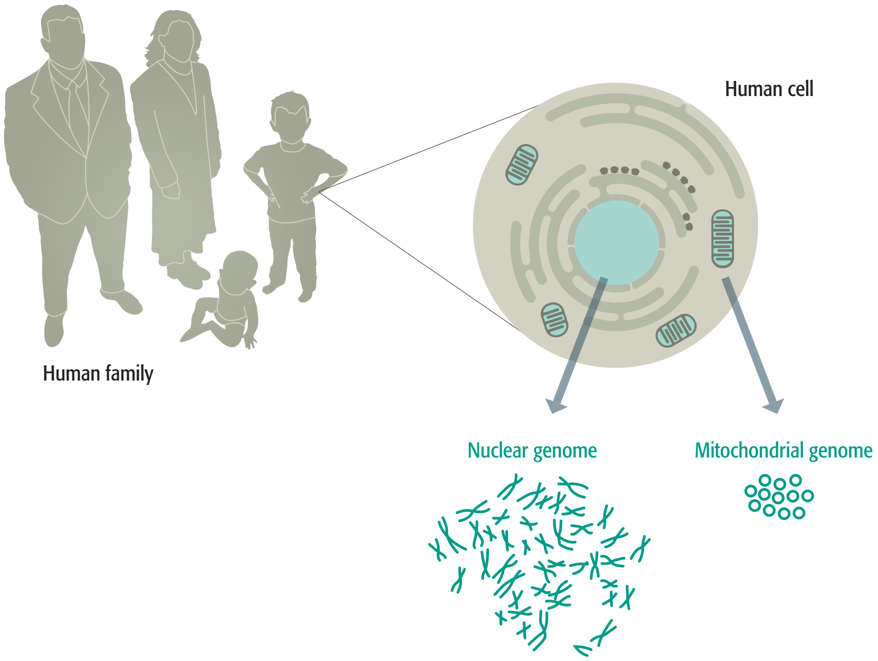


Figure 1.1 The nuclear and mitochondrial components of the human genome.

Each of the approximately 10^{13} cells in the adult human body has its own copy or copies of the genome, the only exceptions being those few cell types, such as red blood cells, that lack a nucleus in their fully differentiated state. The vast majority of cells are **diploid** and so have two copies of each autosome, plus two sex chromosomes, XX for females or XY for males—46 chromosomes in all. These are called **somatic cells**, in contrast to **sex cells,** or **gametes**, which are **haploid** and have just 23 chromosomes, comprising one of each autosome and one sex chromosome. Both types of cell have about 8000 copies of the mitochondrial genome, 10 or so in each mitochondrion.

The genome is a store of biological information but on its own it is unable to release that information to the cell. Utilization of the biological information contained in the genome requires the coordinated activity of enzymes and other proteins, which participate in a complex series of biochemical reactions referred to as **genome expression** (Figure 1.2). The initial product of genome expression is the **transcriptome**, a collection of RNA molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time. The transcriptome is maintained by the process called

transcription, in which individual genes are copied into RNA molecules. The second product of genome expression is the **proteome**, the cell's repertoire of **proteins**, which specifies the nature of the biochemical reactions that the cell is able to carry out. The proteins that make up the proteome are synthesized by **translation** of the individual RNA molecules present in the transcriptome.

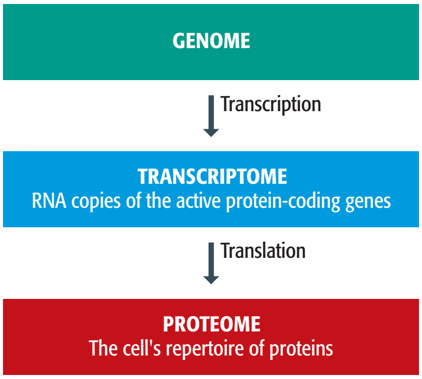

GENOME
Transcription
TRANSCRIPTOME
RNA copies of the active protein-coding genes
Translation
PROTEOME
The cell's repertoire of proteins

Figure 1.2 The genome, transcriptome, and proteome.

This book is about genomes and genome expression. It explains how genomes are studied (Part 1), how they are organized (Part 2), how they function (Part 3), and how they replicate and evolve (Part 4). It was not possible to write this book until very recently. Since the 1950s, molecular biologists have studied individual genes or small groups of genes, and from these studies have built up a wealth of knowledge about how genes work. But only during the last 10 years have techniques been available that make it possible to examine entire genomes. Individual genes are still intensively studied, but information about individual genes is now interpreted within the context of the genome as a whole. This new, broader emphasis applies not just to genomes but to all of biochemistry and cell biology. No longer is it sufficient to understand individual biochemical pathways or subcellular processes. The challenge now is provided by **systems biology**, which attempts to link together these pathways and processes into networks that describe the overall functioning of living cells and living organisms.

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

1.1 DNA

DNA was discovered in 1869 by Johann Friedrich Miescher, a Swiss biochemist working in Tübingen, Germany. The first extracts that Miescher made from human white blood cells were crude mixtures of DNA and chromosomal proteins, but the following year he moved to Basel, Switzerland (where the research institute named after him is now located), and prepared a pure sample of **nucleic acid** from salmon sperm. Miescher's chemical tests showed that DNA is acidic and rich in phosphorus, and also suggested that the individual molecules are very large, although it was not until the 1930s, when biophysical techniques were applied to DNA, that the huge lengths of the polymeric chains were fully appreciated.

1.1.1 Genes are made of DNA

The fact that genes are made of DNA is so well known today that it can be difficult to appreciate that for the first 75 years after its discovery the true role of DNA was unsuspected. As early as 1903, W.S. Sutton had realized that the inheritance patterns of genes parallel the behavior of chromosomes during cell division, an observation that led to the **chromosome theory**, the proposal that genes are located in chromosomes. Examination of cells by **cytochemistry**, after staining with dyes that bind specifically to just one type of biochemical, showed that chromosomes are made of DNA and protein, in roughly equal amounts. Biologists at that time recognized that billions of different genes must exist and the genetic material must therefore be able to take many different forms. But this requirement appeared not to be satisfied by DNA, because in the early part of the twentieth century it was thought that all DNA molecules were the same. On the other hand, it was known, correctly, that proteins are highly variable, polymeric molecules, each one made up of a different combination of 20 chemically distinct amino-acid monomers (Section 1.3.1). Hence genes simply had to be made of protein, not DNA.

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

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

Although from our perspective these two experiments provide the key results that tell us that genes are made of DNA, biologists at the time were not so easily convinced. Both experiments have limitations that leave room for sceptics to argue that protein could still be the genetic material. For example, there were worries about the specificity of the **deoxyribonuclease** enzyme that Avery and colleagues used to inactivate the transforming principle. This result, a central part of the evidence for the transforming principle being DNA, would be invalid if, as seemed possible, the enzyme contained trace amounts of a contaminating **protease** and hence was also able to degrade protein. Neither is the bacteriophage experiment conclusive, as Hershey and Chase stressed when they published their results: "Our experiments show clearly that a physical separation of phage T2 into genetic and nongenetic parts is possible…The chemical identification of the genetic part must wait, however, until some questions…have been answered." In retrospect, these two experiments are important not because of what they tell us but because they alerted biologists to the fact that DNA *might* be the genetic material and was therefore worth studying. It was this that influenced Watson and Crick to work on DNA and, as we will see below, it was their discovery of the **double helix** structure, which solved the puzzling question of how genes can replicate, which really convinced the scientific world that genes are made of DNA.

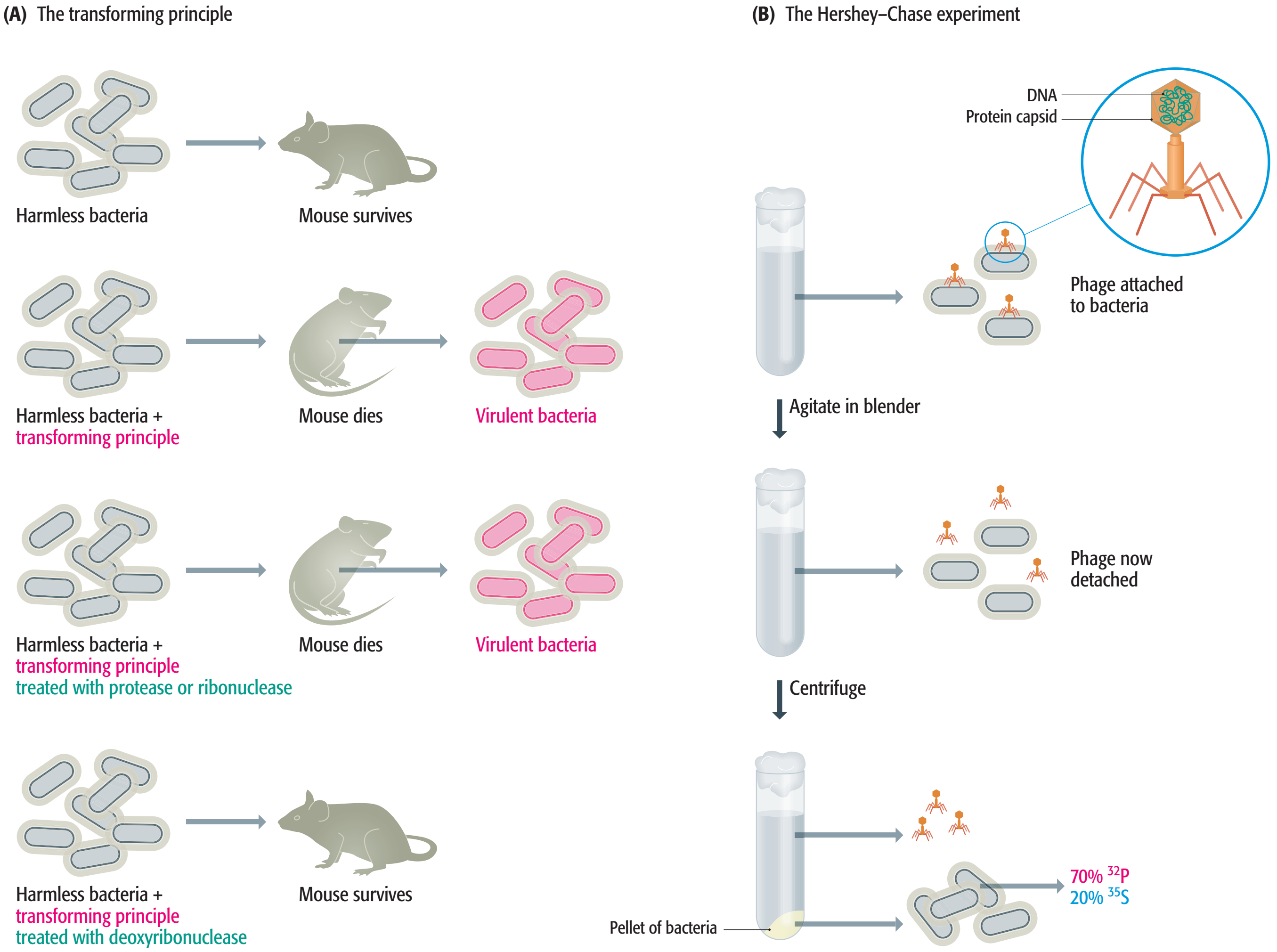


Figure 1.3 The two experiments that suggested that genes are made of DNA.

(A) Avery and colleagues showed that the transforming principle is made of DNA. The top two panels show what happens when mice are injected with harmless *Streptococcus pneumoniae* bacteria, with or without addition of the transforming principle, a cell extract obtained from a virulent strain of *S. pneumoniae*. When the transforming principle is present, the mouse dies, because the genes in the transforming principle convert the harmless bacteria into the virulent form, these virulent bacteria subsequently being recovered from the lungs of the dead mouse. The lower two panels show that treatment with protease or ribonuclease has no effect on the transforming principle, but that the transforming principle is inactivated by deoxyribonuclease.

(B) The Hershey–Chase experiment used T2 bacteriophages, each of which comprises a DNA molecule contained in a protein capsid attached to a "body" and "legs" that enable the bacteriophage to attach to the surface of a bacterium and inject its genes into the cell. The DNA of the bacteriophages was labeled with ^{32}P, and the protein with ^{35}S. A few minutes after infection, the culture was agitated to detach the empty phage particles from the cell surface. The culture was then centrifuged, which collects the bacteria plus phage genes as a pellet at the bottom of the tube, but leaves the lighter phage particles in suspension. Hershey and Chase found that the bacterial pellet contained most of the ^{32}P-labeled component of the phages (the DNA) but only 20% of the ^{35}S-labeled material (the phage protein). In a second experiment, Hershey and Chase showed that new phages produced at the end of the infection cycle contained less than 1% of the protein from the parent phages. For more details of the bacteriophage infection cycle, see Figure 2.19.

1.1.2 The structure of DNA

The names of James Watson and Francis Crick are so closely linked with DNA that it is easy to forget that, when they began their collaboration in October 1951, the detailed structure of the DNA polymer was already known. Their contribution was not to determine the structure of DNA *per se*, but to show that in living cells two DNA chains are intertwined to form the double helix. First, therefore, we should examine what Watson and Crick knew before they began their work.

Nucleotides and polynucleotides

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

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

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

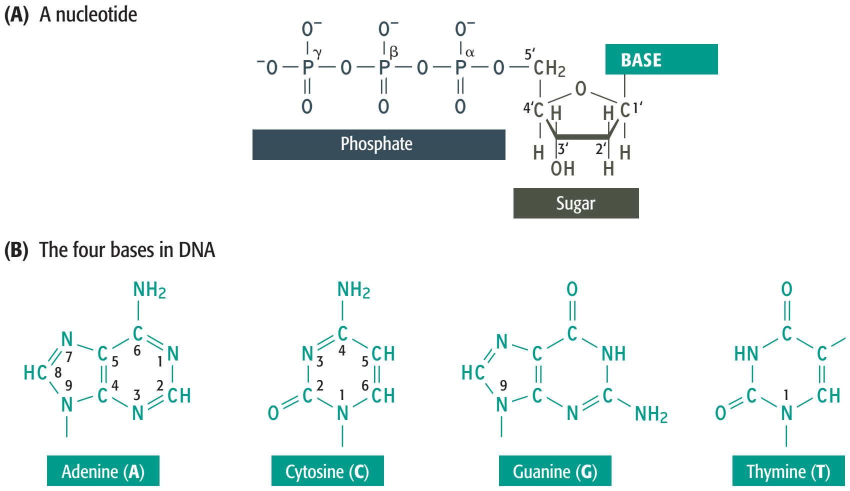


Figure 1.4 The structure of a nucleotide. (A) The general structure of a deoxyribonucleotide, the type of nucleotide found in DNA. (B) The four bases that occur in deoxyribonucleotides.

- 2′-Deoxyadenosine 5′-triphosphate.
- 2′-Deoxycytidine 5′-triphosphate.
- 2′-Deoxyguanosine 5′-triphosphate.
- 2′-Deoxythymidine 5′-triphosphate.

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

In a polynucleotide, individual nucleotides are linked together by **phosphodiester bonds** between their 5′- and 3′-carbons (Figure 1.5). From the structure of this linkage we can see that the polymerization reaction (Figure 1.6) involves removal of the two outer phosphates (the β- and γ-phosphates) from one nucleotide and replacement of the hydroxyl group attached to the 3′-carbon of the second nucleotide. Note that the two ends of the polynucleotide are chemically distinct, one having an unreacted triphosphate group attached to the 5′-carbon (the **5′** or **5′-P terminus**), and the other having an unreacted hydroxyl attached to the 3′-carbon (the **3′** or **3′-OH terminus**). This means that the polynucleotide has a chemical direction, expressed as either 5′→3′ (down in Figure 1.5) or 3′→5′ (up in Figure 1.5). An important consequence of the polarity of the phosphodiester bond is that the chemical reaction needed to extend a DNA polymer in the 5′→3′ direction is different to that needed to make a 3′→5′ extension. All natural **DNA polymerase** enzymes are only able to carry out 5′→3′ synthesis, which adds significant complications to the process by which double-stranded DNA is replicated (Section 15.2).

The evidence that led to the double helix

In the years before 1950, various lines of evidence had shown that cellular DNA molecules are comprised of two or more polynucleotides assembled together in some way. The possibility that unraveling the nature of this assembly might provide insights into how genes work prompted Watson and Crick, among others, to try to solve the structure. According to Watson in his book *The Double Helix*, their work was a desperate race against the famous American biochemist, Linus Pauling, who initially proposed an incorrect triple helix model, giving Watson and Crick the time they needed to complete

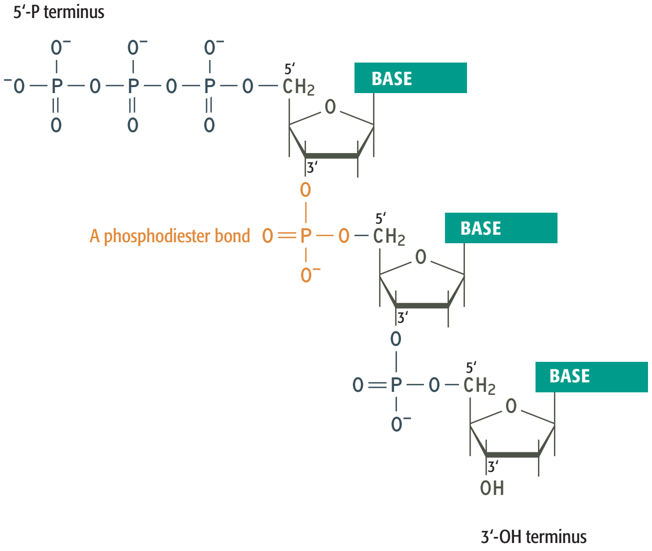


Figure 1.5 A short DNA polynucleotide showing the structure of the phosphodiester bond. Note that the two ends of the polynucleotide are chemically distinct.

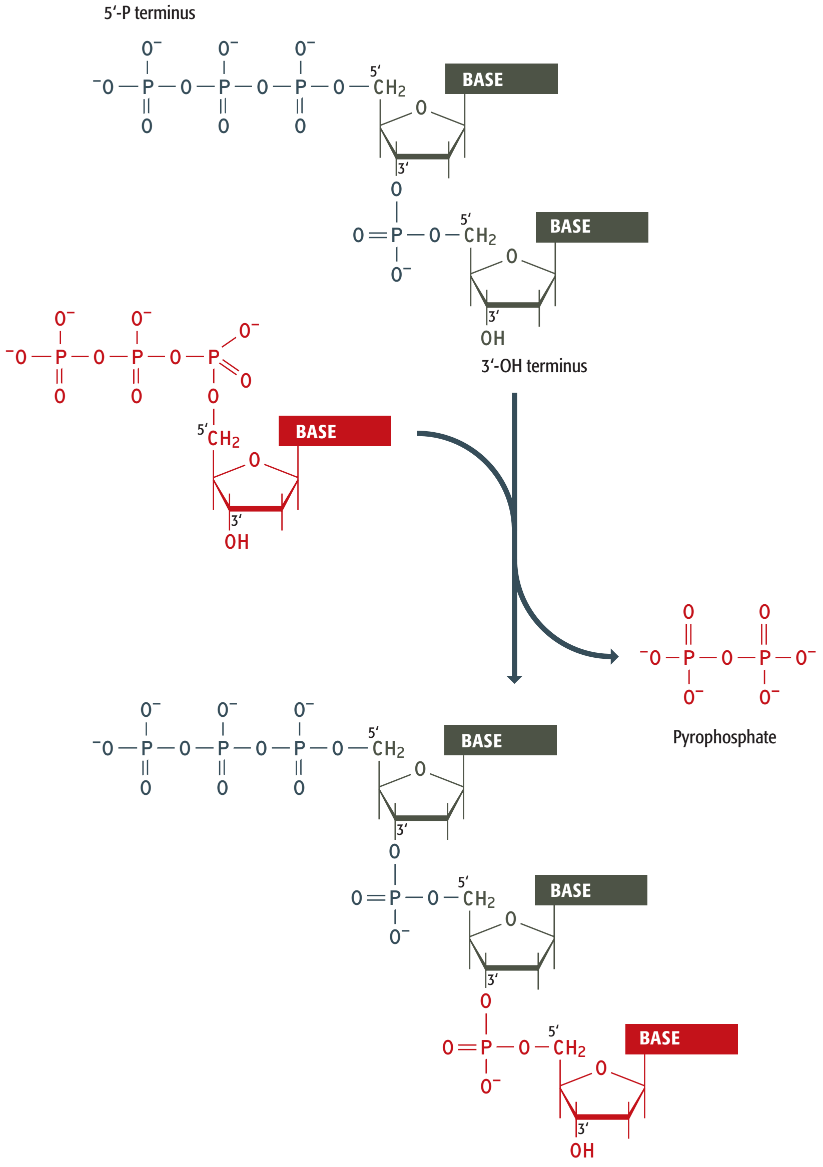


Figure 1.6 The polymerization reaction that results in synthesis of a DNA polynucleotide. Synthesis occurs in the 5′→3′ direction, with the new nucleotide being added to the 3′-carbon at the end of the existing polynucleotide. The β- and γ-phosphates of the nucleotide are removed as a pyrophosphate molecule.

the double helix structure. It is now difficult to separate fact from fiction, especially regarding the part played by Rosalind Franklin, whose **X-ray diffraction** studies provided the bulk of the experimental data in support of the double helix and who was herself very close to solving the structure. The one thing that is clear is that the double helix, discovered by Watson and Crick on Saturday 7 March 1953, was the single most important breakthrough in biology during the twentieth century.

Watson and Crick used four types of information to deduce the double helix structure:

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

- **X-ray diffraction patterns** (Technical Note 11.1), most of which were produced by Rosalind Franklin and which revealed the helical nature of the structure and indicated some of the key dimensions within the helix.
- The **base ratios**, which had been discovered by Erwin Chargaff of Columbia University, New York. Chargaff carried out a lengthy series of chromatographic studies of DNA samples from various sources and showed that, although the values are different in different organisms, the amount of adenine is always the same as the amount of thymine, and the amount of guanine equals the amount of cytosine (Figure 1.7). These base ratios led to the **base-pairing** rules, which were the key to the discovery of the double helix structure.
- Model building, which was the only major technique that Watson and Crick performed themselves. Scale models of possible DNA structures enabled the relative positioning of the various atoms to be checked, to ensure that pairs that formed bonds were not too far apart, and that other atoms were not so close together as to interfere with one another.

The key features of the double helix

The double helix is right-handed, which means that if it were a spiral staircase and you were climbing upwards then the rail on the outside of the staircase would be on your right-hand side. The two strands run in opposite directions (Figure 1.8A). The helix is stabilized by two types of chemical interaction:

- **Base pairing** between the two strands involves the formation of **hydrogen bonds** between an adenine on one strand and a thymine on the other strand, or between a cytosine and a guanine (Figure 1.8B). Hydrogen bonds are weak electrostatic attractions between an electronegative atom (such as oxygen or nitrogen) and a hydrogen atom attached to a second electronegative atom. Hydrogen bonds are longer than covalent bonds and are much weaker, typical bond energies being 1–10 kcal mol^{-1} at 25°C, compared with up to 90 kcal mol^{-1} for a covalent bond. As well as their role in the DNA double helix, hydrogen bonds stabilize protein secondary structures. The two base-pair combinations—A base-paired with T, and G base-paired with C—explain the base ratios discovered by Chargaff. These are the only pairs that are permissible, partly because of the geometries of the nucleotide bases and the relative positions of the atoms that are able to participate in hydrogen bonds, and partly because the pair must be between a purine and a pyrimidine: a purine–purine pair would be too big to fit within the helix, and a pyrimidine–pyrimidine pair would be too small.
- **Base stacking**, sometimes called **π–π interactions**, involves hydrophobic interactions between adjacent base pairs and adds stability to the double helix once the strands have been brought together by base pairing. These hydrophobic interactions arise because the hydrogen-bonded structure of water forces hydrophobic groups into the internal parts of a molecule.

Both base pairing and base stacking are important in holding the two polynucleotides together, but base pairing has added significance because of its biological implications. The limitation that A can only base-pair with T, and G can only base-pair with C, means that DNA replication can result in perfect copies of a parent molecule through the simple expedient of using the sequences of the preexisting strands to dictate the sequences of the new strands. This is **template-dependent DNA synthesis** and it is the system used by all cellular DNA polymerases (Section 15.2.2). Base pairing therefore enables DNA molecules to

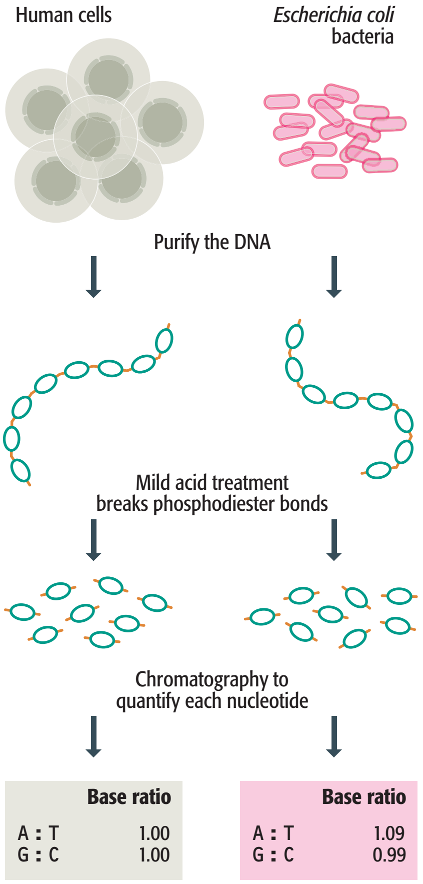

Human cells
Escherichia coli bacteria
Purify the DNA
Mild acid treatment breaks phosphodiester bonds
Chromatography to quantify each nucleotide
Base ratio
A : T 1.00
G : C 1.00
Base ratio
A : T 1.09
G : C 0.99

Figure 1.7 **The base ratio experiments performed by Chargaff.** DNA was extracted from various organisms and treated with acid to hydrolyze the phosphodiester bonds and release the individual nucleotides. Each nucleotide was then quantified by chromatography. The data show some of the actual results obtained by Chargaff. These indicate that, within experimental error, the amount of adenine is the same as that of thymine, and the amount of guanine is the same as that of cytosine.

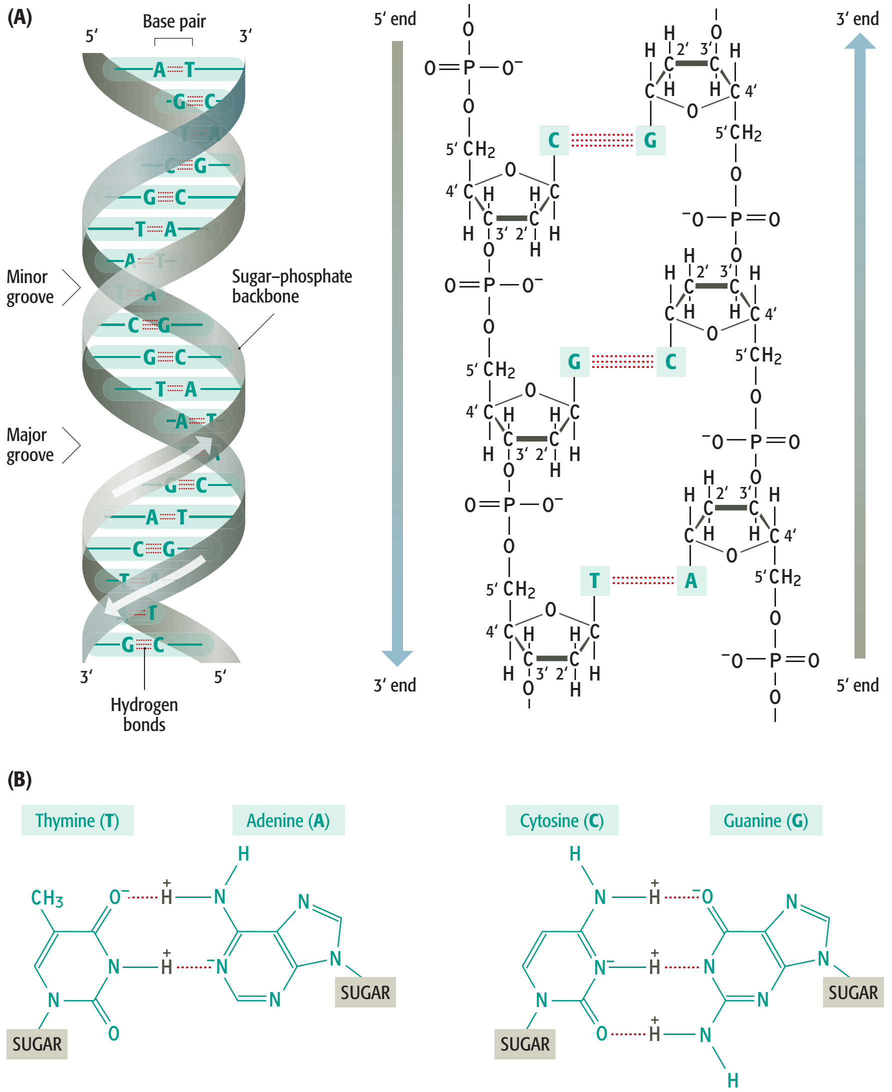


Figure 1.8 The double helix structure of DNA. (A) Two representations of the double helix. On the left the structure is shown with the sugar–phosphate "backbones" of each polynucleotide drawn as a gray ribbon with the base pairs in green. On the right the chemical structure for three base pairs is given. (B) A base-pairs with T, and G base-pairs with C. The bases are drawn in outline, with the hydrogen bonding indicated by dotted lines. Note that a G–C base pair has three hydrogen bonds whereas an A–T base pair has just two.

be replicated by a system that is so simple and elegant that as soon as the double helix structure was publicized by Watson and Crick, every biologist became convinced that genes really are made of DNA.

The double helix has structural flexibility

The double helix described by Watson and Crick, and shown in Figure 1.8A, is called the B-form of DNA. Its characteristic features lie in its dimensions: a helical diameter of 2.37 nm, a rise of 0.34 nm per base pair, and a pitch (i.e., distance taken up by a complete turn of the helix) of 3.4 nm, this corresponding to ten base pairs per turn. The DNA in living cells is thought to be predominantly in this B-form, but it is now clear that genomic DNA molecules are not entirely uniform in structure. This is mainly because each nucleotide in the helix has the flexibility to take up slightly different molecular shapes. To adopt these different conformations, the relative positions of the atoms in the nucleotide must change slightly. There are a number of possibilities but the most important conformational changes involve rotation around the β-*N*-glycosidic bond, changing the orientation of the base relative to the sugar, and rotation around the bond between the 3′- and 4′-carbons. Both rotations have a significant effect on the double helix: changing the base orientation influences the

Table 1.1 Features of different conformations of the DNA double helix

Feature	B-DNA	A-DNA	Z-DNA
Type of helix	Right-handed	Right-handed	Left-handed
Helical diameter (nm)	2.37	2.55	1.84
Rise per base pair (nm)	0.34	0.29	0.37
Distance per complete turn (pitch) (nm)	3.4	3.2	4.5
Number of base pairs per complete turn	10	11	12
Topology of major groove	Wide, deep	Narrow, deep	Flat
Topology of minor groove	Narrow, shallow	Broad, shallow	Narrow, deep

relative positioning of the two polynucleotides, and rotation around the 3′–4′ bond affects the conformation of the sugar–phosphate backbone.

Rotations within individual nucleotides therefore lead to major changes in the overall structure of the helix. It has been recognized since the 1950s that changes in the dimensions of the double helix occur when fibers containing DNA molecules are exposed to different relative humidities. For example, the modified version of the double helix called the A-form (Figure 1.9) has a diameter of 2.55 nm, a rise of 0.29 nm per base pair, and a pitch of 3.2 nm, corresponding to 11 base pairs per turn (Table 1.1). Other variations include B′-, C-, C′-, C″-, D-, E- and T-DNAs. All these are right-handed helices like the B-form. A more drastic reorganization is also possible, leading to the left-handed Z-DNA (Figure 1.9), a slimmer version of the double helix with a diameter of only 1.84 nm.

The bare dimensions of the various forms of the double helix do not reveal what are probably the most significant differences between them. These relate not to diameter and pitch, but the extent to which the internal regions of the helix are accessible from the surface of the structure. As shown in Figures 1.8 and 1.9, the B-form of DNA does not have an entirely smooth surface: instead, two grooves spiral along the length of the helix. One of these grooves is relatively wide and deep and is called the **major groove**; the other is narrow and less deep and is called the **minor groove**. A-DNA also has two grooves (Figure 1.9), but with this conformation the major groove is even deeper, and the minor groove shallower and broader compared with B-DNA. Z-DNA is different again, with one groove virtually nonexistent but the other very narrow and deep. In each form of DNA, part of the internal surface of at least one of the grooves is formed by chemical groups attached to the nucleotide bases. In Chapter 11 we will see that expression of the biological information contained within a genome is mediated by DNA-binding proteins that attach to the double helix and regulate the activity of the genes contained within it. To carry out their function, each DNA-binding protein must attach at a specific position, near to the gene whose activity it must influence. This can be achieved, with at least some degree of accuracy, by the protein reaching down into a groove, within which the DNA sequence can be “read” without the helix being opened up by breaking the base pairs. A corollary of this is that a DNA-binding protein whose structure enables it to recognize a specific nucleotide sequence within B-DNA, for example, might not be able to recognize that sequence if the DNA has taken up a different conformation.

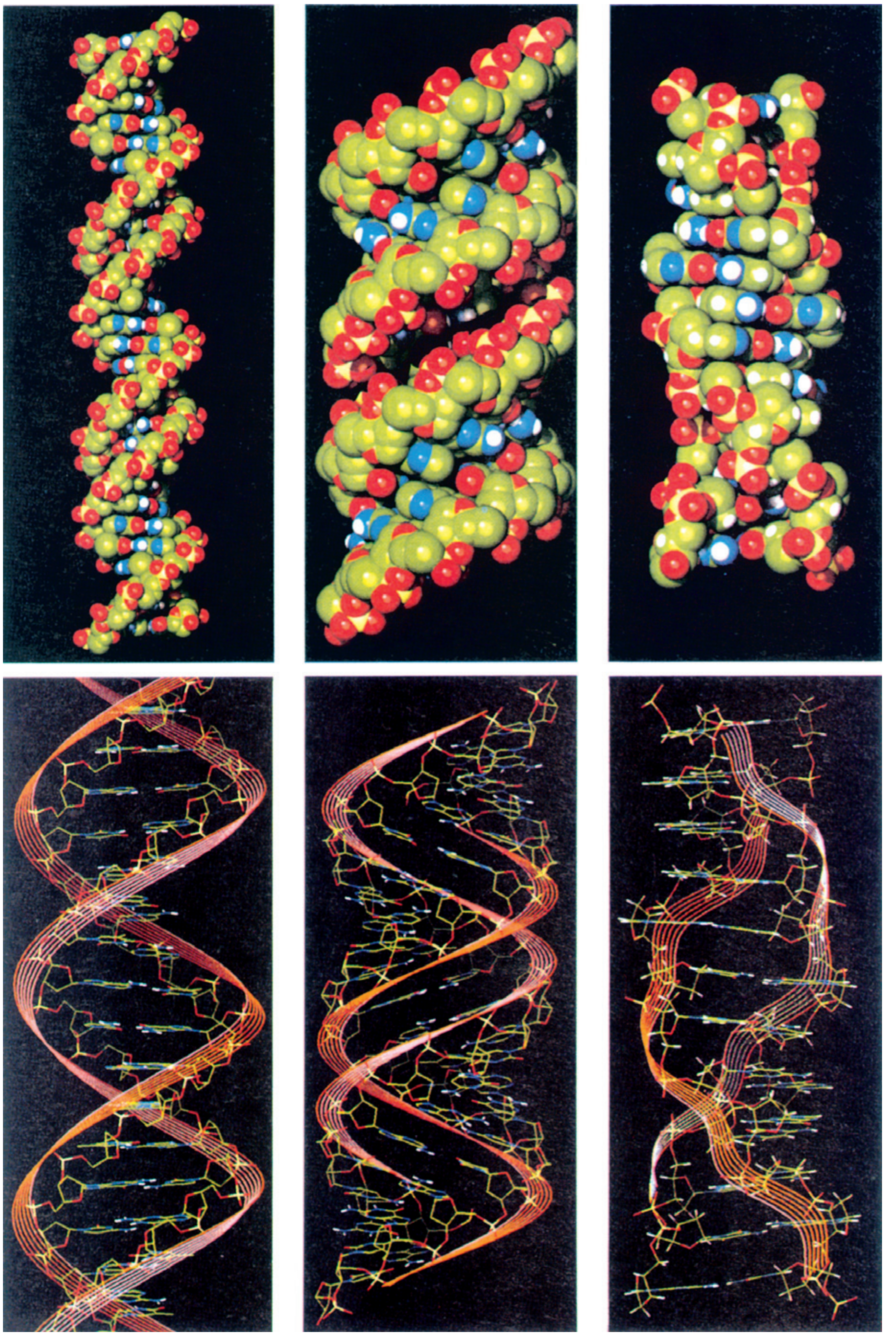


Figure 1.9 The structures of B-DNA (left), A-DNA (center) and Z-DNA (right). Space-filling models (top) and structural models (bottom) depicting different conformations of DNA molecules. Note the differences in helical diameter, number of base pairs per complete turn, and topology of the major and minor grooves between these molecules. Reprinted with permission from Kendrew, J. (Ed.), *Encyclopaedia of Molecular Biology*. © 1994 Blackwell Publishing.

As we will see in Chapter 11, conformational variations along the length of a DNA molecule, together with other structural polymorphisms caused by the nucleotide sequence, could be important in determining the specificity of the interactions between the genome and its DNA-binding proteins.

1.2 RNA and the Transcriptome

The initial product of genome expression is the transcriptome (see Figure 1.2), the collection of RNA molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time. The RNA molecules of the transcriptome, as well as many other RNAs derived from genes that do not code for proteins, are synthesized by the process called transcription. In this section we will examine the structure of RNA and then look more closely at the various types of RNA molecule that are present in living cells.

1.2.1 The structure of RNA

RNA is a polynucleotide similar to DNA but with two important chemical differences (Figure 1.10). First, the sugar in an RNA nucleotide is **ribose** and, second, RNA contains **uracil** instead of thymine. The four nucleotide substrates for synthesis of RNA are therefore:

- Adenosine 5′-triphosphate.
- Cytidine 5′-triphosphate.
- Guanosine 5′-triphosphate.
- Uridine 5′-triphosphate.

These nucleotides are abbreviated to ATP, CTP, GTP, and UTP, or A, C, G, and U, respectively.

As with DNA, RNA polynucleotides contain 3′–5′ phosphodiester bonds, but these phosphodiester bonds are less stable than those in a DNA polynucleotide because of the indirect effect of the hydroxyl group at the 2′-position of the sugar. RNA molecules are rarely more than a few thousand nucleotides in length, and although many form *intra*molecular base pairs (for example, see Figure 13.2), most are single- rather than double-stranded.

The enzymes responsible for transcription of DNA into RNA are called **DNA-dependent RNA polymerases**. The name indicates that the enzymatic reaction that they catalyze results in polymerization of RNA from ribonucleotides and occurs in a DNA-dependent manner, meaning that the sequence of nucleotides in a DNA template dictates the sequence of nucleotides in the RNA that is made (Figure 1.11). It is permissible to shorten the enzyme name to **RNA polymerase**, as the context in which the name is used means that there is rarely confusion with the **RNA-dependent RNA polymerases** that are involved in replication and expression of some virus genomes. The chemical basis of the template-dependent synthesis of RNA is equivalent to that shown for the synthesis of DNA in Figure 1.6. Ribonucleotides are added one after another to the growing 3′ end of the RNA transcript, the identity of each nucleotide being specified by the base-pairing rules: A base-pairs with T or U; G base-pairs with C. During each nucleotide addition, the β- and γ-phosphates are removed from the incoming nucleotide, and the hydroxyl group is removed from the 3′-carbon of the nucleotide at the end of the chain, precisely the same as for DNA polymerization.

1.2.2 The RNA content of the cell

A typical bacterium contains 0.05–0.10 pg of RNA, making up about 6% of its total weight. A mammalian cell, being much larger, contains more RNA, 20–30 pg in all, but this represents only 1% of the cell as a whole. The best way to understand the RNA content of a cell is to divide it into categories and subcategories depending on function. There are several ways of doing this, the most informative scheme being the one shown in Figure 1.12. The primary division is between **coding RNA** and **noncoding RNA**. The coding RNA comprises the transcriptome and is made up of just one class of molecule, the **messenger RNAs** (**mRNAs**), which are transcripts of protein-coding genes

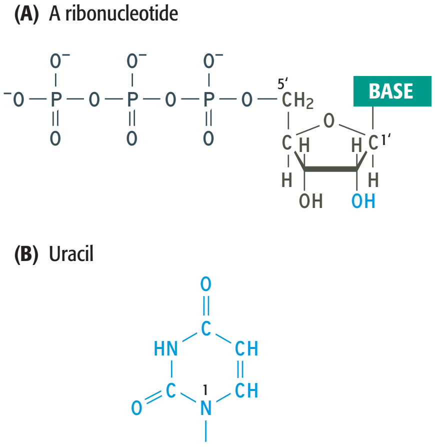


Figure 1.10 The chemical differences between DNA and RNA. (A) RNA contains ribonucleotides, in which the sugar is ribose rather than 2′-deoxyribose. The difference is that a hydroxyl group, rather than hydrogen atom, is attached to the 2′-carbon. (B) RNA contains the pyrimidine called uracil instead of thymine.

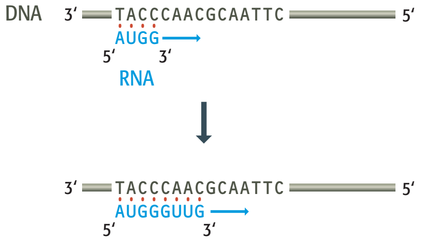


Figure 1.11 Template-dependent RNA synthesis. The RNA transcript is synthesized in the 5′→3′ direction, reading the DNA in the 3′→5′ direction, with the sequence of the transcript determined by base-pairing to the DNA template.

and hence are translated into protein in the second stage of genome expression. Messenger RNAs rarely make up more than 4% of the total RNA and are short-lived, being degraded soon after synthesis. Bacterial mRNAs have half-lives of no more than a few minutes, and in eukaryotes most mRNAs are degraded a few hours after synthesis. This rapid turnover means that the composition of the transcriptome is not fixed and can quickly be restructured by changing the rate of synthesis of individual mRNAs.

The second type of RNA is referred to as “noncoding” as these molecules are not translated into protein. However, a better name is **functional RNA**, as this emphasizes that, although not part of the transcriptome, the noncoding RNAs still have essential roles within the cell. There are several diverse types of functional RNA, the most important being as follows:

- **Ribosomal RNAs** (**rRNAs**) are present in all organisms and are usually the most abundant RNAs in the cell, making up over 80% of the total RNA in actively dividing bacteria. These molecules are components of ribosomes, the structures on which protein synthesis takes place (Section 13.2).
- **Transfer RNAs** (**tRNAs**) are small molecules that are also involved in protein synthesis and, like rRNA, are found in all organisms. The function of tRNAs is to carry amino acids to the ribosome and ensure that the amino acids are linked together in the order specified by the nucleotide sequence of the mRNA that is being translated (Section 13.1).
- **Small nuclear RNAs** (**snRNAs**; also called **U-RNAs** because these molecules are rich in uridine nucleotides) are found in the nuclei of eukaryotes. These molecules are involved in **splicing**, one of the key steps in the processing events that convert the primary transcripts of protein-coding genes into mRNAs (Section 12.2.2).
- **Small nucleolar RNAs** (**snoRNAs**) are found in the nucleolar regions of eukaryotic nuclei. They play a central role in the chemical modification of rRNA molecules by directing the enzymes that perform the modifications to the specific nucleotides where alterations, such as addition of a methyl group, must be carried out (Section 12.2.5).
- **MicroRNAs (miRNAs)** and **short interfering RNAs (siRNAs)** are small RNAs that regulate the expression of individual genes (Section 12.2.6).

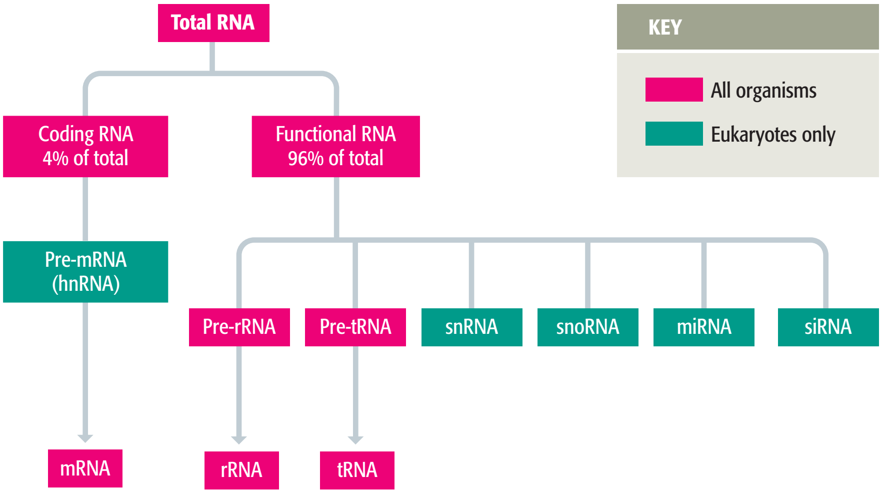


Figure 1.12 The RNA content of a cell. This scheme shows the types of RNA present in all organisms and those categories found only in eukaryotic cells.

1.2.3 Processing of precursor RNA

As well as the mature RNAs described above, cells also contain precursor molecules. Many RNAs, especially in eukaryotes, are initially synthesized as precursor or **pre-RNA**, which has to be processed before it can carry out its function. The various processing events, all of which are described in Chapter 12, include the following (Figure 1.13):

- **End-modifications** occur during the synthesis of eukaryotic mRNAs, most of which have a single, unusual nucleotide called a **cap** attached at the 5′ end and a **poly(A) tail** attached to the 3′ end.
- **Splicing** is the removal of segments from within a precursor RNA. Many genes, especially in eukaryotes, contain internal segments that contain no biological information. These are called **introns** and they are copied along with the information-containing **exons** when the gene is transcribed. The introns are removed from the **pre-mRNA** by cutting and joining reactions. Unspliced pre-mRNA forms the nuclear RNA fraction called **heterogenous nuclear RNA** (**hnRNA**).
- **Cutting events** are particularly important in the processing of rRNAs and tRNAs, many of which are initially synthesized from transcription units that specify more than one molecule. The **pre-rRNAs** and **pre-tRNAs** must therefore be cut into pieces to produce the mature RNAs. This type of processing occurs in both prokaryotes and eukaryotes.
- **Chemical modifications** are made to rRNAs, tRNAs, and mRNAs. The rRNAs and tRNAs of all organisms are modified by addition of new chemical groups, these groups being added to specific nucleotides within each RNA. Chemical modification of mRNA, called **RNA editing**, occurs in many eukaryotes.

The transcriptome

Although the transcriptome makes up less than 4% of the total cell RNA, it is the most significant component because it contains the coding RNAs that are used in the next stage of genome expression. It is important to note that the transcriptome is never synthesized *de novo*. Every cell receives part of its parent's transcriptome when it is first brought into existence by cell division, and maintains a transcriptome throughout its lifetime. Even quiescent cells in bacterial spores or in the seeds of plants have a transcriptome, although translation of that transcriptome into protein may be completely switched off. Transcription of individual protein-coding genes does not, therefore, result in *synthesis* of the transcriptome but instead *maintains* the transcriptome by replacing mRNAs that have been degraded, and brings about

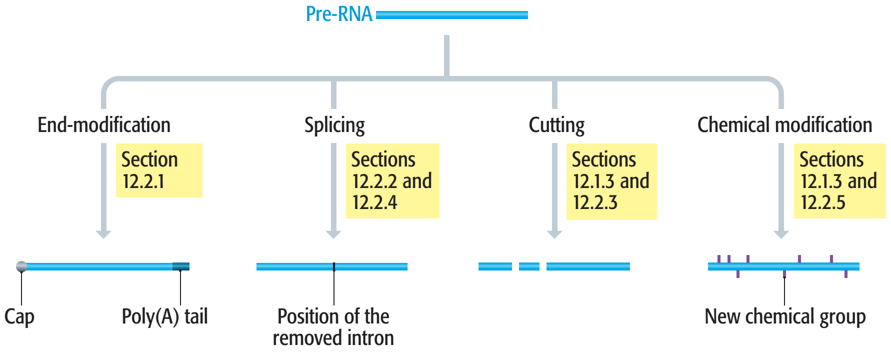

Pre-RNA
End-modification
Section 12.2.1
Splicing
Sections 12.2.2 and 12.2.4
Cutting
Sections 12.1.3 and 12.2.3
Chemical modification
Sections 12.1.3 and 12.2.5
Cap
Poly(A) tail
Position of the removed intron
New chemical group

Figure 1.13 Schematic representation of the four types of RNA processing event. Not all events occur in all organisms.

changes to the composition of the transcriptome via the switching on and off of different sets of genes.

Even in the simplest organisms, such as bacteria and yeast, many genes are active at any one time. Transcriptomes are therefore complex, containing copies of hundreds, if not thousands, of different mRNAs. Usually each mRNA makes up only a small fraction of the whole, with the most common type rarely contributing more than 1% of the total mRNA. Exceptions are those cells that have highly specialized biochemistries, reflected by transcriptomes in which one or a few mRNAs predominate. Developing wheat seeds are an example: these synthesize large amounts of the gliadin proteins, which accumulate in the dormant grain and provide a source of amino acids for the germinating seedling. Within the developing seeds, the gliadin mRNAs can make up as much as 30% of the transcriptomes of certain cells.

1.3 Proteins and the Proteome

The second product of genome expression is the proteome (see Figure 1.2), the cell's repertoire of proteins, which specifies the nature of the biochemical reactions that the cell is able to carry out. These proteins are synthesized by **translation** of the mRNA molecules that make up the transcriptome.

1.3.1 Protein structure

A protein**,** like a DNA molecule, is a linear, unbranched polymer. In proteins, the monomeric subunits are called **amino acids** (Figure 1.14) and the resulting polymers, or **polypeptides**, are rarely more than 2000 units in length. As with DNA, the key features of protein structure were determined in the first half of the twentieth century, this phase of protein biochemistry culminating in the 1940s and early 1950s with the elucidation by Pauling and Corey of the major conformations, or **secondary structures**, taken up by polypeptides. In recent years, interest has focused on how these secondary structures combine to produce the complex, three-dimensional shapes of proteins.

The four levels of protein structure

Proteins are traditionally looked upon as having four distinct levels of structure. These levels are hierarchical, the protein being built up stage-by-stage, with each level of structure depending on the one below it:

- The **primary structure** of the protein is formed by joining amino acids into a polypeptide. The amino acids are linked by **peptide bonds** that are formed by a condensation reaction between the carboxyl group of one amino acid and the amino group of a second amino acid (Figure 1.15). Note that, as with a polynucleotide, the two ends of the polypeptide are chemically distinct: one has a free amino group and is called the **amino**, **NH_2–**, or **N terminus**; the other has a free carboxyl group and is called the **carboxyl**, **COOH–**, or **C terminus**. The direction of the polypeptide can therefore be expressed as either N→C (left to right in Figure 1.15) or C→N (right to left in Figure 1.15).
- The **secondary structure** refers to the different conformations that can be taken up by the polypeptide. The two main types of secondary structure are the **α-helix** and **β-sheet** (Figure 1.16). These are stabilized mainly by hydrogen bonds that form between different amino acids in the

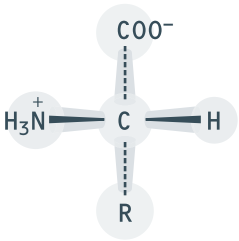


Figure 1.14 The general structure of an amino acid. All amino acids have the same general structure, comprising a central α-carbon attached to a hydrogen atom, a carboxyl group, an amino group, and an R group. The R group is different for each amino acid (see Figure 1.18).

polypeptide. Most polypeptides are long enough to be folded into a series of secondary structures, one after another along the molecule.

- The **tertiary structure** results from folding the secondary structural components of the polypeptide into a three-dimensional configuration (Figure 1.17). The tertiary structure is stabilized by various chemical forces, notably hydrogen bonding between individual amino acids, electrostatic interactions between the R groups of charged amino acids (see Figure 1.18), and hydrophobic forces, which dictate that amino acids with nonpolar ("water-hating") side-groups must be shielded from water by embedding within the internal regions of the protein. There may also be covalent linkages called **disulfide bridges** between cysteine amino acid residues at various places in the polypeptide.
- The **quaternary structure** involves the association of two or more polypeptides, each folded into its tertiary structure, into a multisubunit protein. Not all proteins form quaternary structures, but it is a feature of many proteins with complex functions, including several involved in genome expression. Some quaternary structures are held together by disulfide bridges between the different polypeptides, resulting in stable multisubunit proteins that cannot easily be broken down to the component parts. Other quaternary structures comprise looser associations of subunits stabilized by hydrogen bonding and hydrophobic effects, which means that these proteins can revert to their component polypeptides, or change their subunit composition, according to the functional requirements of the cell.

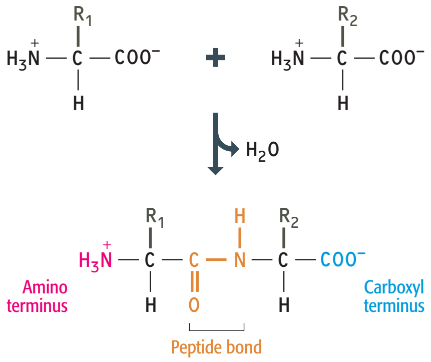


Figure 1.15 In polypeptides, amino acids are linked by peptide bonds. The drawing shows the chemical reaction that results in two amino acids becoming linked together by a peptide bond. The reaction is called a condensation because it results in elimination of water.

Amino acid diversity underlies protein diversity

Proteins are functionally diverse because the amino acids from which proteins are made are themselves chemically diverse. Different sequences of amino acids therefore result in different combinations of chemical reactivities, these combinations dictating not only the overall structure of the resulting protein but also the positioning on the surface of the structure of reactive groups that determine the chemical properties of the protein.

Amino acid diversity derives from the R group because this part is different in each amino acid and varies greatly in structure. Proteins are made up from a set of 20 amino acids (Figure 1.18; Table 1.2). Some of these have R groups that are small, relatively simple structures, such as a single hydrogen atom (in the amino acid called glycine) or a methyl group (alanine). Other R groups are large, complex aromatic side chains (phenylalanine, tryptophan, and tyrosine). Most amino acids are uncharged, but two are negatively charged (aspartic acid and glutamic acid) and three are positively charged (arginine, histidine, and lysine). Some amino acids are polar (e.g., glycine, serine, and threonine), others are nonpolar (e.g., alanine, leucine, and valine).

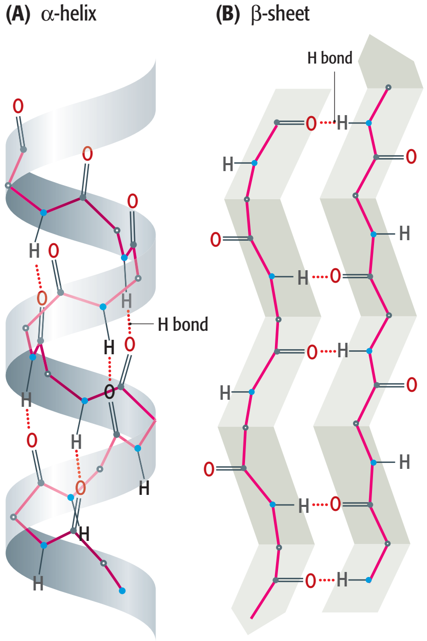


Figure 1.16 The two main secondary structural units found in proteins: (A) the α-helix, and (B) the β-sheet. The polypeptide chains are shown in outline. The R groups have been omitted for clarity. Each structure is stabilized by hydrogen (H) bonds between the C=O and N–H groups of different peptide bonds. The β-sheet conformation that is shown is antiparallel, the two chains running in opposite directions. Parallel β-sheets also occur.

The 20 amino acids shown in Figure 1.18 are the ones that are conventionally looked upon as being specified by the genetic code (Section 1.3.2). They are therefore the amino acids that are linked together when mRNA molecules are translated into proteins. However, these 20 amino acids do not, on their own, represent the limit of the chemical diversity of proteins. The diversity is even greater because of two factors:

- At least two additional amino acids—selenocysteine and pyrrolysine (Figure 1.19)—can be inserted into a polypeptide chain during protein synthesis, their insertion directed by a modified reading of the genetic code (Section 13.1.1).

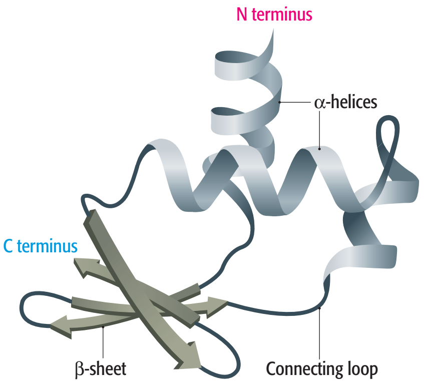


Figure 1.17 The tertiary structure of a protein. This imaginary protein structure comprises three α-helices, shown as coils, and a four-stranded β-sheet, indicated by the arrows.

- During protein processing, some amino acids are modified by the addition of new chemical groups, for example by acetylation or phosphorylation, or by attachment of large side chains made up of sugar units (Section 13.3.3).

Proteins therefore have an immense amount of chemical variability, some of this directly specified by the genome, the remainder arising by protein processing.

1.3.2 The proteome

The proteome comprises all the proteins present in a cell at a particular time. A "typical" mammalian cell, for example a liver hepatocyte, is thought to contain 10,000–20,000 different proteins, about 8×10^9 individual molecules in all, representing approximately 0.5 ng of protein or 18%–20% of the total cell weight. The copy numbers of individual proteins vary enormously, from less than 20,000 molecules per cell for the rarest types to 100 million copies for the commonest ones. Any protein that is present at a copy number of greater than 50,000 per cell is considered to be relatively abundant, and in the average mammalian cell some 2000 proteins fall into this category. When the proteomes of different types of mammalian cell are examined, very few differences

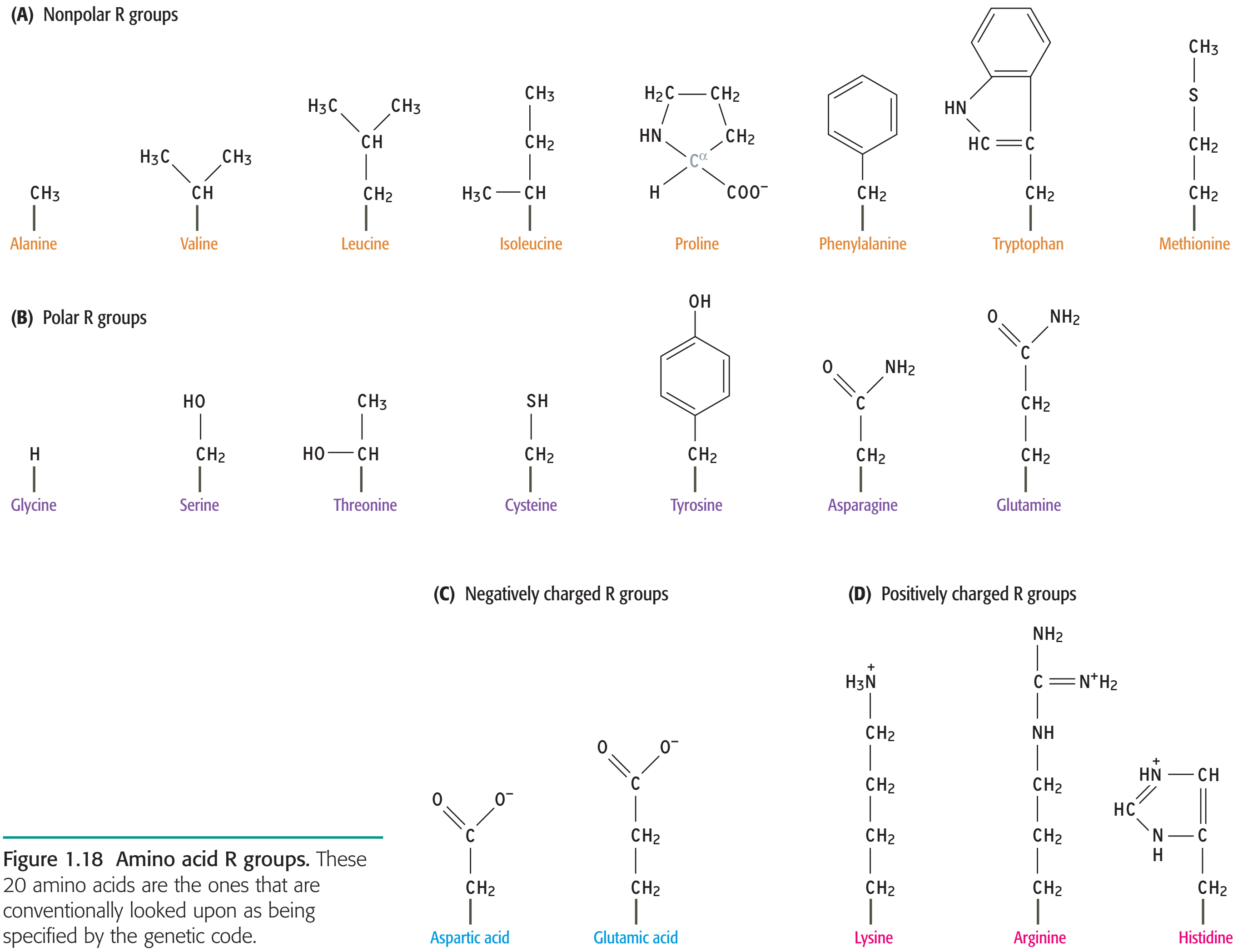


Figure 1.18 Amino acid R groups. These 20 amino acids are the ones that are conventionally looked upon as being specified by the genetic code.

Table 1.2 Amino acid abbreviations

Amino acid	Abbreviation	
	Three-letter	One-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

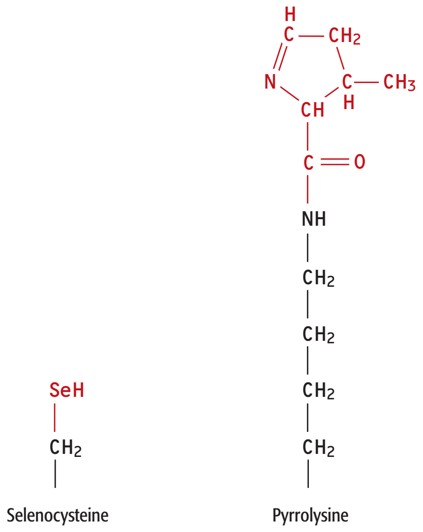


Figure 1.19 The structures of selenocysteine and pyrrolysine. The parts shown in brown indicate the differences between these amino acids and cysteine and lysine, respectively.

are seen among these abundant proteins, suggesting that most of them are **housekeeping** proteins that perform general biochemical activities that occur in all cells. The proteins that provide the cell with its specialized function are often quite rare, although there are exceptions, such as the vast amounts of hemoglobin that are present only in red blood cells.

The link between the transcriptome and the proteome

The flow of information from DNA to RNA by transcription does not provide any conceptual difficulty. DNA and RNA polynucleotides have very similar structures and we can easily understand how an RNA copy of a gene can be made by template-dependent synthesis using the base-pairing rules with which we are familiar. The second phase of genome expression, during which the mRNA molecules of the transcriptome direct synthesis of proteins, is less easy to understand simply by considering the structures of the molecules that are involved. In the early 1950s, shortly after the double helix structure of DNA had been discovered, several molecular biologists attempted to devise ways in which amino acids could attach to mRNAs in an ordered fashion, but in all of these schemes at least some of the bonds had to be shorter or longer than was possible according to the laws of physical chemistry, and each idea was quietly dropped. Eventually, in 1957, Francis Crick cut a way through the confusion by predicting the existence of an adaptor molecule that would form a bridge between the mRNA and the polypeptide being synthesized. Soon afterwards it was realized that the tRNAs are these adaptor molecules, and once this fact had been established, a detailed understanding of the mechanism by which proteins are synthesized was quickly built up. We will examine this process in Section 13.1.

The other aspect of protein synthesis that interested molecular biologists in the 1950s was the **informational problem**. This refers to the second important component of the link between the transcriptome and proteome: the **genetic code**, which specifies how the nucleotide sequence of an mRNA is translated into the amino acid sequence of a protein. It was recognized in the 1950s that a triplet genetic code—one in which each codeword, or **codon,** comprises three nucleotides—is required to account for all 20 amino acids found in proteins. A two-letter code would have only $4^2 = 16$ codons, which is not enough to account for all 20 amino acids, whereas a three-letter code would give $4^3 = 64$ codons. The genetic code was worked out in the 1960s, partly by analysis of polypeptides arising from translation of artificial mRNAs of known or predictable sequence in cell-free systems, and partly by determining which amino acids associated with which RNA sequences in an assay based on purified ribosomes. When this work was completed, it was realized that the 64 codons fall into groups, the members of each group coding for the same amino acid (Figure 1.20). Only tryptophan and methionine have just a single codon each: all other amino acids are coded by two, three, four, or six codons. This feature of the code is called **degeneracy**. The code also has four **punctuation codons**, which indicate the points within an mRNA where translation of the nucleotide sequence should start and finish (Figure 1.21). The **initiation codon** is usually 5′–AUG–3′, which also specifies methionine (so most newly synthesized polypeptides start with methionine), although with a few mRNAs other codons such as 5′–GUG–3′ and 5′–UUG–3′ are used. The three **termination codons** are 5′–UAG–3′, 5′–UAA–3′, and 5′–UGA–3′.

The genetic code is not universal

It was originally thought that the genetic code must be the same in all organisms. The argument was that, once established, it would be impossible for the code to change because giving a new meaning to any single codon would result in widespread disruption of the amino acid sequences of proteins. This reasoning seems sound, so it is surprising that, in reality, the genetic code is not universal. The code shown in Figure 1.20 holds for the vast majority of genes in the vast majority of organisms, but deviations are widespread. In particular, mitochondrial genomes often use a nonstandard code (Table 1.3). This was first discovered in 1979 by Frederick Sanger's group in Cambridge, UK, who found that several human mitochondrial mRNAs contain the sequence 5′–UGA–3′, which normally codes for termination, at internal positions where protein synthesis was not expected to stop. Comparisons with the amino acid sequences of the proteins coded by these mRNAs showed that 5′–UGA–3′ is a tryptophan codon in human mitochondria, and that this is just one of four code deviations in this particular genetic system. Mitochondrial genes in other organisms also display code deviations, although at least one of these—the use of 5′–CGG–3′ as a tryptophan codon in plant mitochondria—is probably corrected by RNA editing (Section 12.2.5) before translation occurs.

Nonstandard codes are also known for the nuclear genomes of lower eukaryotes. Often a modification is restricted to just a small group of organisms and

UUU	phe	UCU	ser	UAU	tyr	UGU	cys
UUC		UCC		UAC		UGC	
UUA	leu	UCA		UAA	stop	UGA	stop
UUG		UCG		UAG		UGG	trp
CUU	leu	CCU	pro	CAU	his	CGU	arg
CUC		CCC		CAC		CGC	
CUA		CCA		CAA	gln	CGA	
CUG		CCG		CAG		CGG	
AUU	ile	ACU	thr	AAU	asn	AGU	ser
AUC		ACC		AAC		AGC	
AUA		ACA		AAA	lys	AGA	arg
AUG	met	ACG		AAG		AGG	
GUU	val	GCU	ala	GAU	asp	GGU	gly
GUC		GCC		GAC		GGC	
GUA		GCA		GAA	glu	GGA	
GUG		GCG		GAG		GGG	

Figure 1.20 The genetic code. Amino acids are designated by the standard, three-letter abbreviations (see Table 1.2).

frequently it involves reassignment of the termination codons (Table 1.3). Modifications are less common among prokaryotes, but one example is known in *Mycoplasma* species. A more important type of code variation is **context-dependent codon reassignment**, which occurs when the protein to be synthesized contains either selenocysteine or pyrrolysine. Proteins containing pyrrolysine are rare, and are probably only present in the group of prokaryotes called the **archaea** (Chapter 8), but selenoproteins are widespread in many organisms, one example being the enzyme glutathione peroxidase, which helps protect the cells of humans and other mammals against oxidative damage. Selenocysteine is coded by 5′–UGA–3′ and pyrrolysine by 5′–UAG–3′. These codons therefore have a dual meaning because they are still used as termination codons in the organisms concerned (Table 1.3). A 5′–UGA–3′ codon that specifies selenocysteine is distinguished from true termination codons by the presence of a hairpin loop structure in the mRNA, positioned just downstream of the selenocysteine codon in prokaryotes, and in the 3′ untranslated region (i.e., the part of the mRNA after the termination codon) in eukaryotes. Recognition of the selenocysteine codon requires interaction between the hairpin structure and a special protein that is involved in translation of these mRNAs. A similar system probably operates for specifying pyrrolysine.

The link between the proteome and the biochemistry of the cell

The biological information encoded by the genome finds its final expression in a protein whose biological properties are determined by its folded structure and by the spatial arrangement of chemical groups on its surface. By specifying proteins of different types, the genome is able to construct and maintain a proteome whose overall biological properties form the underlying basis of life. The proteome can play this role because of the huge diversity of protein structures that can be formed, the diversity enabling proteins to carry out a variety of biological functions. These functions include the following:

- Biochemical catalysis is the role of the special type of proteins called enzymes. The central metabolic pathways, which provide the cell with energy, are catalyzed by enzymes, as are the biosynthetic processes that

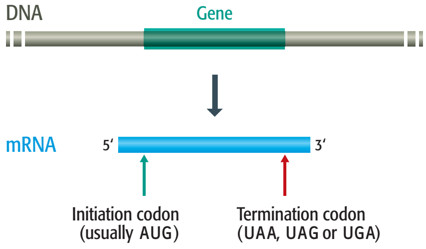

DNA
Gene
mRNA
5′
3′
Initiation codon (usually AUG)
Termination codon (UAA, UAG or UGA)

Figure 1.21 The positions of the punctuation codons in an mRNA.

Table 1.3 Examples of deviations from the standard genetic code

Organism	Codon	Should code for	Actually codes for
Mitochondrial genomes			
Mammals	UGA	Stop	Trp
	AGA, AGG	Arg	Stop
	AUA	Ile	Met
Drosophila	UGA	Stop	Trp
	AGA	Arg	Ser
	AUA	Ile	Met
Saccharomyces cerevisiae	UGA	Stop	Trp
	CUN	Leu	Thr
	AUA	Ile	Met
Fungi	UGA	Stop	Trp
Maize	CGG	Arg	Trp
Nuclear and prokaryotic genomes			
Several protozoa	UAA, UAG	Stop	Gln
Candida cylindracea	CUG	Leu	Ser
Micrococcus sp.	AGA	Arg	Stop
	AUA	Ile	Stop
Euplotes sp.	UGA	Stop	Cys
Mycoplasma sp.	UGA	Stop	Trp
	CGG	Arg	Stop
Context-dependent codon reassignments			
Various	UGA	Stop	Selenocysteine
Archaea	UAG	Stop	Pyrrolysine

Abbreviation: N, any nucleotide.

result in construction of nucleic acids, proteins, carbohydrates, and lipids. Biochemical catalysis also drives genome expression through the activities of enzymes such as RNA polymerase.

- Structure, which at the cellular level is determined by the proteins that make up the cytoskeleton, is also the primary function of some extracellular proteins. An example is collagen, which is an important component of bones and tendons.
- Movement is conferred by contractile proteins, of which actin and myosin in cytoskeletal fibers are the best-known examples.
- Transport of materials around the body is an important protein activity: for example, hemoglobin transports oxygen in the bloodstream, and serum albumin transports fatty acids.
- Regulation of cellular processes is mediated by signaling proteins such as STATs (signal transducers and activators of transcription, Section 14.1.2), and by proteins such as **activators** that bind to the genome and influence the expression levels of individual genes and groups of genes (Section 11.3). The activities of groups of cells are regulated and coordinated by

extracellular hormones and cytokines, many of which are proteins (e.g., insulin, the hormone that controls blood sugar levels, and the interleukins, a group of cytokines that regulate cell division and differentiation).

- Protection of the body and of individual cells is the function of a range of proteins, including the antibodies and those proteins involved in the blood-clotting response.
- Storage functions are performed by proteins such as ferritin, which acts as an iron store in the liver, and the gliadins, which store amino acids in dormant wheat seeds.

This multiplicity of protein function provides the proteome with its ability to convert the blueprint contained in the genome into the essential features of the life process.

Summary

The genome is the store of biological information possessed by every organism on the planet. The vast majority of genomes are made of DNA, the few exceptions being those viruses that have RNA genomes. Genome expression is the process by which the information contained in the genome is released to the cell. The first product of genome expression is the transcriptome, the collection of RNAs derived from the protein-coding genes that are active in the cell at a particular time. The second product is the proteome, the cell's repertoire of proteins that specify the nature of the biochemical reactions that the cell is able to carry out. Experimental evidence showing that genes are made of DNA was first obtained between 1945 and 1952, but it was the discovery of the double helix structure by Watson and Crick in 1953 that convinced biologists that DNA is, indeed, the genetic material. A DNA polynucleotide is an unbranched polymer made up of multiple copies of four chemically different nucleotides. In the double helix, two polynucleotides are wound around one another, with the nucleotide bases on the inside of the molecule. The polynucleotides are linked by hydrogen bonding between the bases, with A always base-paired to T, and G always base-paired to C. RNA is also a polynucleotide but the individual nucleotides have different structures compared with those found in DNA, and RNA is usually single-stranded. DNA-dependent RNA polymerases are responsible for copying genes into RNA by the process called transcription, which results in synthesis not only of the transcriptome but also of a range of functional RNA molecules, which do not code for proteins but still play vital roles in the cell. Many RNAs are initially synthesized as precursor molecules that are processed by cutting and joining reactions, and by chemical modifications, to give the mature forms. Proteins are also unbranched polymers, but in these the units are amino acids linked by peptide bonds. The amino acid sequence is the primary structure of a protein. The higher levels of structure—secondary, tertiary, and quaternary—are formed by folding the primary structure into three-dimensional conformations and by association of individual polypeptides into multiprotein structures. Proteins are functionally diverse because individual amino acids have different chemical properties that, when combined in different ways, result in proteins with a range of chemical features. Proteins are synthesized by translation of mRNAs, with the rules of the genetic code specifying which triplet of nucleotides codes for which amino acid. The genetic code is not universal, variations occurring in mitochondria and in lower eukaryotes, and some codons can have two different meanings in a single gene.

Multiple Choice Questions

*Answers to odd-numbered questions can be found in the Appendix

1.1.* Which of the following statements about an organism's genome is FALSE?

a. The genome contains the genetic information to construct and maintain a living organism.
b. The genomes of cellular organisms are composed of DNA.
c. The genome is able to express its own information without the activity of enzymes and proteins.
d. Eukaryotic genomes are composed of both nuclear and mitochondrial DNA.

1.2. Somatic cells are those that:

a. Contain a haploid set of chromosomes.
b. Give rise to the gametes.
c. Lack mitochondria.
d. Contain a diploid set of chromosomes and make up the majority of human cells.

1.3.* The flow of genetic information in cells is which of the following?

a. DNA is transcribed into RNA, which is then translated into protein.
b. DNA is translated into protein, which is then transcribed into RNA.
c. RNA is transcribed into DNA, which is then translated into protein.
d. Proteins are translated into RNA, which is then transcribed into DNA.

1.4. In the early twentieth century it was thought that proteins might carry genetic information. This reasoning was due to which of the following?

a. Chromosomes are composed of approximately equal amounts of protein and DNA.
b. Proteins were known to be composed of 20 distinct amino acids whereas DNA is composed of only 4 nucleotides.
c. Different proteins were known to have unique sequences, whereas it was thought that all DNA molecules have the same sequence.
d. All of the above.

1.5.* Which type of bonds link the individual nucleotides together in DNA?

a. Glycosidic.
b. Peptide.
c. Phosphodiester.
d. Electrostatic.

1.6. In solving the structure of DNA, Watson and Crick actively used which of the following techniques?

a. Model building of DNA molecules to ensure that the atoms were correctly positioned.
b. X-ray crystallography of DNA.
c. Chromatographic studies to determine the relative composition of nucleotides from various sources.
d. Genetic studies that demonstrated that DNA is the genetic material.

1.7.* Erwin Chargaff studied DNA from various organisms and demonstrated that:

a. DNA is the genetic material.
b. RNA is transcribed from DNA.
c. The amount of adenine in a given organism is equal to the amount of thymine (and guanine to cytosine).
d. The double helix is held together by hydrogen bonding between the bases.

1.8. The transcriptome of a cell is defined as:

a. All of the RNA molecules present in a cell.
b. The protein-coding RNA molecules present in a cell.
c. The ribosomal RNA molecules present in a cell.
d. The transfer RNA molecules present in a cell.

1.9.* How do DNA-dependent RNA polymerases carry out RNA synthesis?

a. They use DNA as a template for the polymerization of ribonucleotides.
b. They use proteins as a template for the polymerization of ribonucleotides.
c. They use RNA as a template for the polymerization of ribonucleotides.
d. They require no template for the polymerization of ribonucleotides.

1.10. Which type of functional RNA is a primary component of the structures required for protein synthesis?

a. Messenger RNA.
b. Ribosomal RNA.
c. Small nuclear RNA.
d. Transfer RNA.

1.11.* The proteome of a cell is defined as:

a. All of the proteins that a cell is capable of synthesizing.
b. All of the proteins present in a cell over the cell's lifetime.
c. All of the proteins present in a cell at a given moment.
d. All of the proteins that are actively being synthesized in a cell at a given moment.

Multiple Choice Questions (continued) *Answers to odd-numbered questions can be found in the Appendix

1.12. Which level of protein structure describes the folded conformation of a multisubunit protein?

a. Primary structure.
b. Secondary structure.
c. Tertiary structure.
d. Quaternary structure.

1.13.* Which type of covalent bond is important for linking cysteine residues located at various places in a polypeptide?

a. Disulfide bridge.
b. Hydrogen bond.
c. Peptide bond.
d. Phosphodiester bond.

1.14. Most of the abundant proteins in a cell are thought to be housekeeping proteins. What is their function?

a. They are responsible for the specific functions of individual cell types.
b. They are responsible for regulating genome expression in cells.
c. They are responsible for removing waste materials from cells.
d. They are responsible for the general biochemical activities that occur in all cells.

1.15.* The degeneracy of the genetic code refers to which of the following?

a. Each codon can specify more than one amino acid.
b. Most amino acids have more than one codon.
c. There are several initiation codons.
d. The stop codons can also code for amino acids.

1.16. Which of the following is NOT a biological function of proteins?

a. Biological catalysis.
b. Regulation of cellular processes.
c. Carrying genetic information.
d. Transport of molecules in multicellular organisms.

Short Answer Questions

*Answers to odd-numbered questions can be found in Appendix 1

1.1.* Provide a time line for the discovery of DNA, the discovery that DNA is the genetic material, the discovery of the structure of DNA, and the characterization of the first genome.

1.2. Which two types of chemical interaction stabilize the double helix?

1.3.* Why does the specific base pairing between A with T, and G with C, provide a basis for the fidelity of DNA replication?

1.4. What are the two important chemical differences between RNA and DNA?

1.5.* Why do mRNA molecules have short half-lives compared to other RNA molecules?

1.6. Is the mRNA that is translated in the same form as that synthesized from the DNA template?

1.7.* Do cells ever lack a transcriptome? Explain the significance of your answer.

1.8. How do hydrogen bonds, electrostatic interactions, and hydrophobic forces play important roles in the secondary, tertiary, and quaternary structures of proteins?

1.9.* How can proteins have so many diverse structures and functions when they are all synthesized from just 20 amino acids?

1.10. In addition to the 20 amino acids, proteins have additional chemical diversity because of two factors. What are these two factors, and what is their importance?

1.11.* How can the codon 5′–UGA–3′ function as both a stop codon and as a codon for the modified amino acid selenocysteine?

1.12. How does the genome direct the biological activity of a cell?

In-depth Problems

*Guidance to odd-numbered questions can be found in the Appendix

1.1.* The text (page 10) states that Watson and Crick discovered the double helix structure of DNA on Saturday 7 March 1953. Justify this statement.

1.2. Discuss why the double helix gained immediate universal acceptance as the correct structure for DNA.

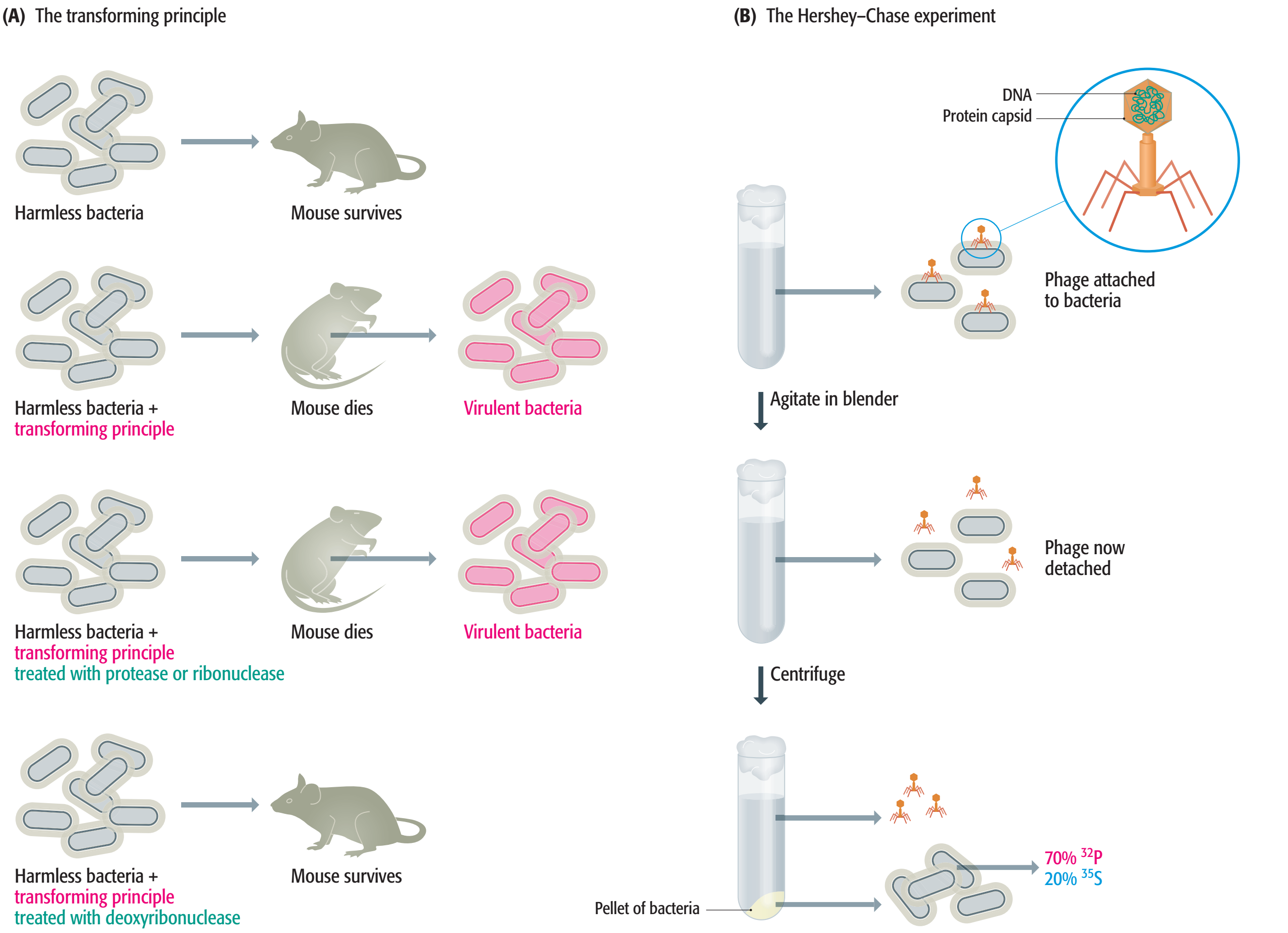
1.3.* What experiments led to elucidation of the genetic code in the 1960s?

1.4. The transcriptome and proteome are looked on as, respectively, an intermediate and the end-product of genome expression. Evaluate the strengths and limitations of these terms for our understanding of genome expression.

Figure Tests

*Answers to odd-numbered questions can be found in the Appendix

1.1.* Discuss how each of these experiments helped to demonstrate that DNA, and not proteins, contains genetic information.



(A) The transforming principle

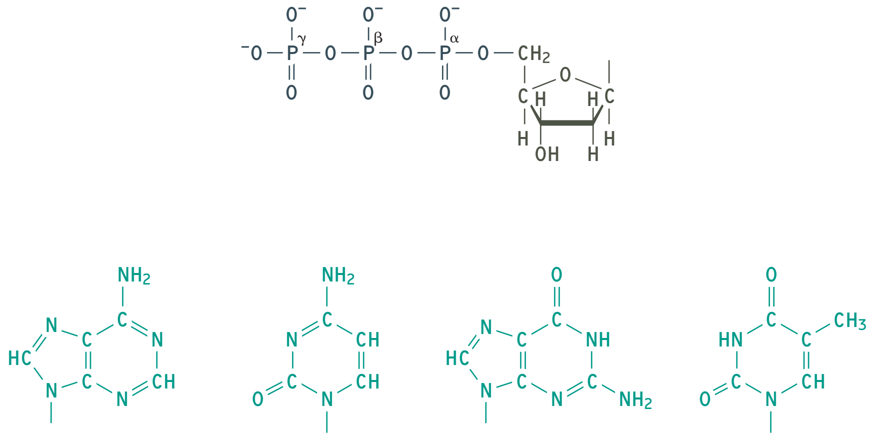
(B) The Hershey–Chase experiment

Harmless bacteria
Mouse survives
Harmless bacteria + transforming principle
Mouse dies
Virulent bacteria
Harmless bacteria + transforming principle treated with protease or ribonuclease
Mouse dies
Virulent bacteria
Harmless bacteria + transforming principle treated with deoxyribonuclease
Mouse survives
DNA
Protein capsid
Phage attached to bacteria
Agitate in blender
Phage now detached
Centrifuge
Pellet of bacteria
70% ^{32}P
20% ^{35}S

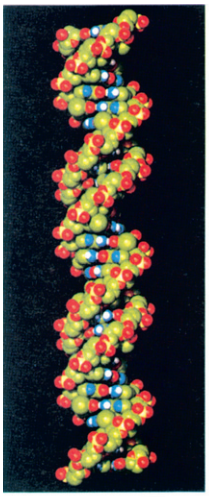
Figure Tests (continued)

*Answers to odd-numbered questions can be found in the Appendix

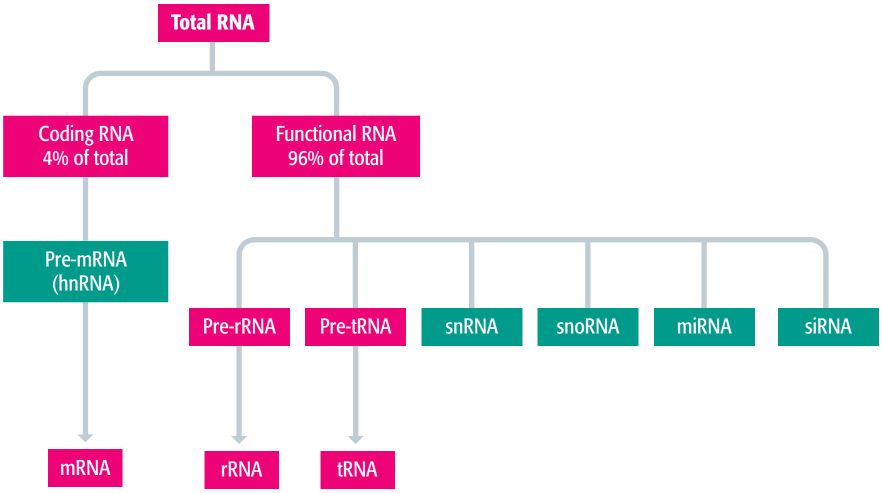
1.2. Identify the deoxyribose, phosphate groups, and different nitrogenous bases. Can you identify the 1′ through 5′ carbon atoms of the deoxyribose?



1.3.* For this space-filling model of B-DNA, describe the important structural features of the molecule.



1.4. Explain the differences in RNA between prokaryotic and eukaryotic cells.



Further Reading

Books and articles on the discovery of the double helix and other important landmarks in the study of DNA

Brock, T.D. (1990) *The Emergence of Bacterial Genetics.* Cold Spring Harbor Laboratory Press, New York. *A detailed history that puts into context the work on the transforming principle and the Hershey–Chase experiment.*

Judson, H.F. (1979) *The Eighth Day of Creation.* Jonathan Cape, London. *A highly readable account of the development of molecular biology up to the 1970s.*

Kay, L.E. (1993) *The Molecular Vision of Life.* Oxford University Press, Oxford. *Contains a particularly informative explanation of why genes were once thought to be made of protein.*

Lander, E.S. and Weinberg, R.A. (2000) Genomics: journey to the center of biology. *Science* **287:** 1777–1782. *A brief description of genetics and molecular biology from Mendel to the human genome sequence.*

Maddox, B. (2002) *Rosalind Franklin: The Dark Lady of DNA.* HarperCollins, London.

McCarty, M. (1985) *The Transforming Principle: Discovering that Genes are Made of DNA.* Norton, London.

Olby, R. (1974) *The Path to the Double Helix.* Macmillan, London *A scholarly account of the research that led to the discovery of the double helix.*

Watson, J.D. (1968) *The Double Helix.* Atheneum, London. *The most important discovery of twentieth century biology, written as a soap opera.*