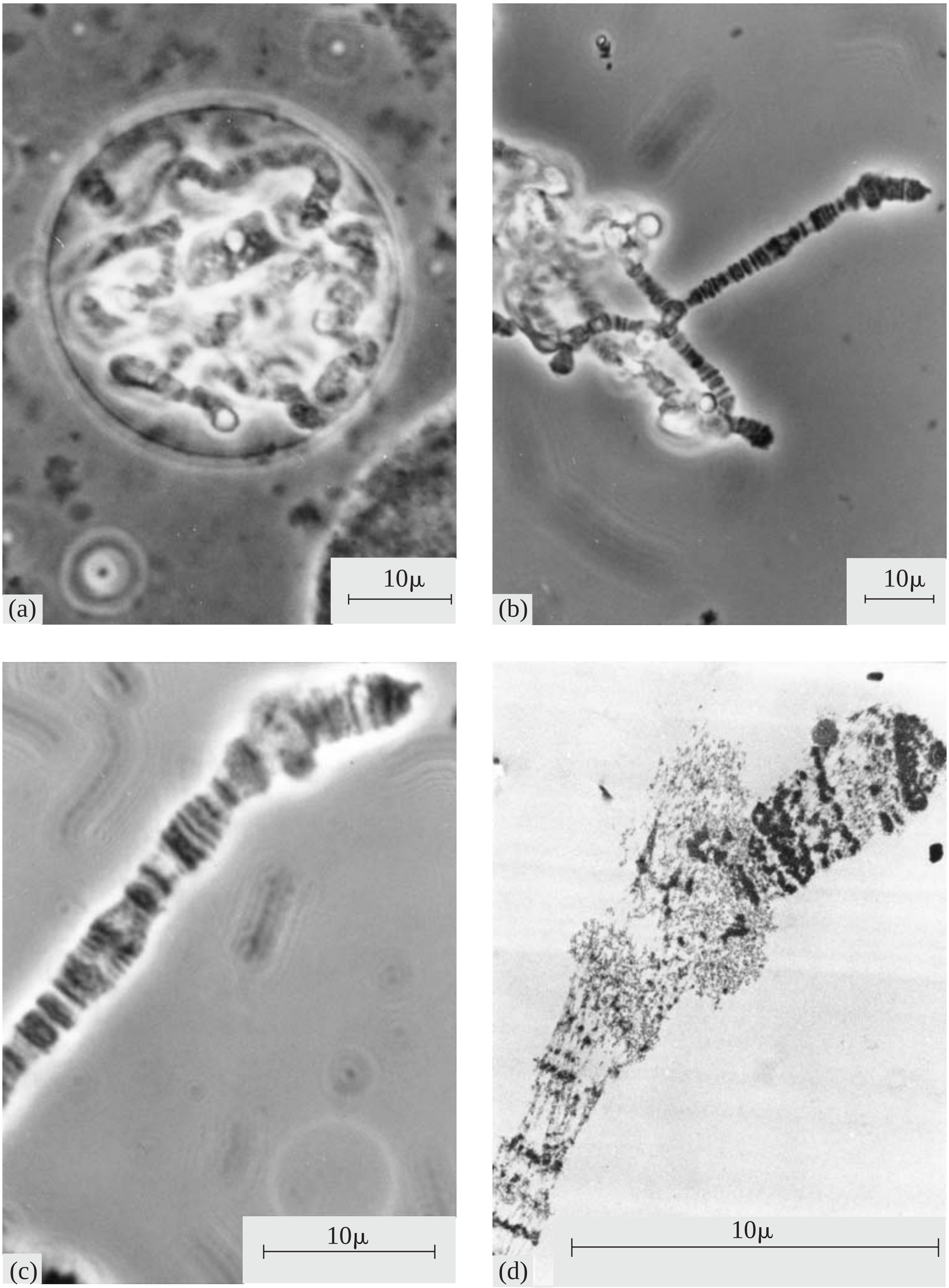

Figure 1.4 Polytene chromosomes from the salivary gland of a fruit fly, as seen in the electron microscope. Length scales: 100 000 Å or 10 μm. Courtesy of Ron Hill and Margaret Mott.

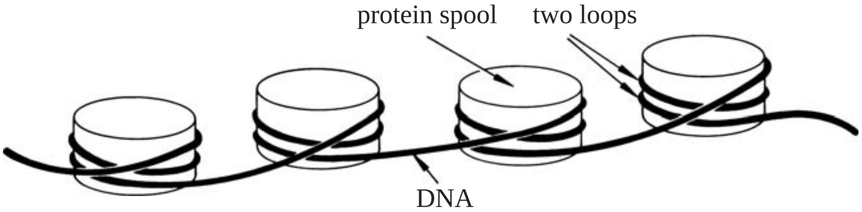

Figure 1.5 A DNA thread making two coils (left-handed) around each of a series of protein spools, as might be seen in the chromosomes of Figs 1.2, 1.3 and 1.4 if the magnification were higher. Each DNA–protein spool is about 100 Å (or 10^{-8} m) across.

are like little worms or eels surrounded by water and trapped inside a bubble. When we break open the bubble and look more closely, as in Fig. 1.4(b), (c) and (d), we can see that each chromosome is divided along its length into many clear striations of dark and light, which are known as 'bands' and 'interbands'. The bands are dense clumps of protein[1] plus DNA, whereas the interbands are sparse regions of low density. No one knows why the variations in dark and light are so sharp, and indeed so reproducible from one fly to the next. Presumably, these divisions are marked off by certain patterns in the molecules along the DNA thread. But it makes good sense in biological terms to divide a chromosome along its length: after all, computer tapes store information in the same sort of way as discrete files, each with a beginning and an end, along the length of a tape.

At a finer level, more is known. Each thread of the band and interband material in fruit-fly chromosomes (or indeed in human chromosomes) is composed of a well-defined mixture of protein and DNA. This protein is of a special kind that makes 'spools', and the DNA wraps twice around each spool, as shown schematically in Fig. 1.5, into a series of double loops. The wrapping of DNA twice about each spool reduces its overall length by a factor of about 6.

When we remove the protein spools, we are left with a very long, string-like DNA molecule. The DNA from the longest individual human chromosome, if it were enlarged by a factor of 10^6, so that it became the width of ordinary kite string, would extend for about 100 km. Imagine sitting in a train traveling from Cambridge to London, or from Los Angeles to San Diego, and looking out of the window for the whole trip at a single DNA molecule and watching the genes go by!

The basic form of this immensely long DNA fiber is shown in Fig. 1.6. It consists of two strands which coil around each other to make a 'double helix'. The term 'helix' is just another word for 'screw' or 'spiral'. The sense of wrapping of these two strands is

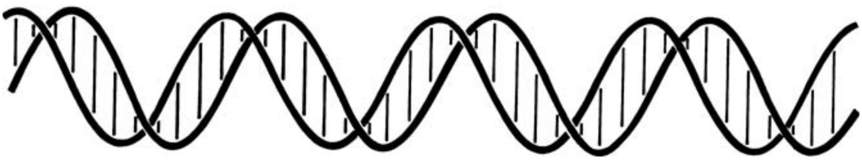

Figure 1.6 DNA double helix showing base-pairs as short vertical lines. Now the DNA thread of Fig. 1.5 is shown in its true form, as a double spiral containing two chains of DNA. The diameter of the spiral or helix is about 20 Å.

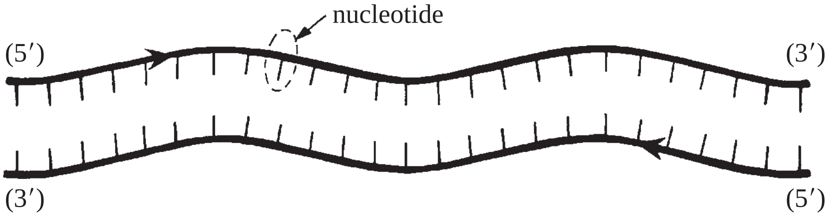

Figure 1.7 The two strands of DNA separated, showing a nucleotide. Each nucleotide is about 6 Å wide.

usually clockwise as you go forward, or right handed – the same sense as an ordinary corkscrew.

If we uncoil the two strands, as shown in Fig. 1.7, then each strand may be seen to consist of a series of units called 'nucleotides'. These are linked to one another with a certain 'directionality', known technically as '5-prime to 3-prime', in a head-to-tail sense that we shall explain below. The two strands run in opposite directions, as shown by the labels 5′ and 3′, and by the arrows.

Each nucleotide is made of about 20 atoms, such as carbon, nitrogen, and oxygen. These atoms can again be grouped into smaller parts which are connected in a particular way. The three parts of a nucleotide are its sugar, phosphate, and base; in the diagram of Fig. 1.8, they are labeled S, P, and B, respectively. For present purposes, we may draw the sugar as a five-sided ring, because the atoms in a sugar join to form such a ring. We have taken the liberty of drawing the ring in the form of an arrow, in order to indicate its directionality; and for the same reason we have put the letters P, S, and B, so that they read in this same direction from left to right. If we took out the central nucleotide in Fig. 1.8, rotated it through 180° about a vertical axis and tried to re-connect it, we should find that it would not link up.

We can draw the base as a rectangle because the atoms there join to form a flat, rigid shape. There are usually four different kinds of nucleotides in DNA, which share the same sugars and phosphates but have different bases attached to the chain. These four different

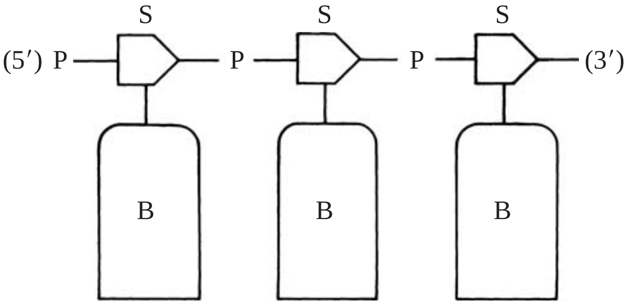

Figure 1.8 Three nucleotides, showing their sugar (S), phosphate (P), and base (B) components.

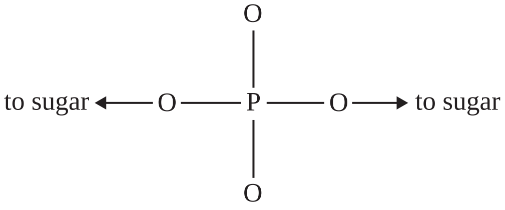

Figure 1.9 The phosphate group as part of a sugar–phosphate chain. Each atom is 1–2 Å wide; atom types are phosphorus (P) and oxygen (O).

bases form pairwise interactions that join the two strands of DNA together weakly (see Chapter 2).

The three parts of a nucleotide can also be studied in isolation, as collections of just a few atoms. For example, a phosphate contains one phosphorus and four oxygen atoms, as shown in Fig. 1.9. The atoms themselves are made of protons, neutrons, and electrons. They are connected together by the sharing of electrons; and the various kinds of chemical bond that connect atoms to each other are all aspects of this. We could go on to discuss atoms within the realm of subatomic physics, but there would be no point in this, because our knowledge of living things begins with chemistry (or biochemistry) and extends into biology. So we can stop our description of DNA with the relatively simple picture of atoms shown in Fig. 1.9, without losing anything.

If you have not studied biology before, you may be puzzled that we have put so much emphasis on the *cell*. Would it not be more sensible to start with the parts of the body, such as limbs, eyes, and lungs? The answer to this is that all of these various organs and tissues, etc. are built up from cells by essentially the same process.

You are probably familiar with the way in which a house is built. The various components – bricks, cement, tiles, timber planks, window frames, etc. – are first delivered to the building site and then assembled in accordance with the architect's plan or 'blueprint'.

This plan is a sheet of paper on which are drawings of the finished house taken from several points of view. There may well be only one copy of this plan, and it may be kept in the pocket of the foreman builder and taken out and consulted from time to time, as required. Such a scheme of construction is typical for non-living things like houses, motor cars, and gadgets.

In contrast, the scheme of construction for living things – whether they are plants or animals – is different from this in almost every way. Thus, construction of a human begins with a single cell – an egg from the mother which has been fertilized by a sperm from the father. This composite cell contains DNA from both the mother and the father; such DNA contains the complete genetic information for construction of a human being. Growth occurs by the process of cell division: each cell divides into two new cells, and these cells in turn divide, and so on. Just before any cell divides, it duplicates all of its DNA, so that every new cell contains a complete set of DNA, which again contains all the genes of the organism. (We saw some pictures of duplicated chromosomes in Fig. 1.3.)

Only a small fraction of all the genes present on this DNA are activated in any given type of cell. Thus, cells which develop into an eye use only the genes which program for the growth of eye-cells. How cells 'know' which kind of organ they belong to is a large and only partly understood area of research, *developmental biology*, which we shall not go into here.

We mention all of this because it may seem, to a non-scientist, to be enormously wasteful for Nature to provide a complete set of DNA in every one of the billions of cells of every animal or plant: would it not be more straightforward and efficient just to have a single copy of the design information, just as in the construction of a house? However, a little thought indicates that the scheme for providing every cell with a complete set of DNA is, in fact, an extremely simple way of providing this necessary information, in all places where it is required – even though, of course, the scheme requires a vast amount of repetitive copying and duplicating of DNA. The machinery for duplicating DNA is accurate enough for the entire scheme to work well, and the double-helical structure of DNA is extremely convenient, as we shall see in Chapter 2, for the purposes of duplication.

We are now ready to clarify a further point: how does DNA carry the information necessary to run the activities of a cell as a factory, as its control-center or main office? The most basic way in which DNA runs the activities of a cell is to specify the composition and structure of protein molecules. Proteins come in a wide range of shapes and sizes, and play a wide range of roles in the life of a cell. Some proteins are strong and rigid, and form the building-components for muscles,

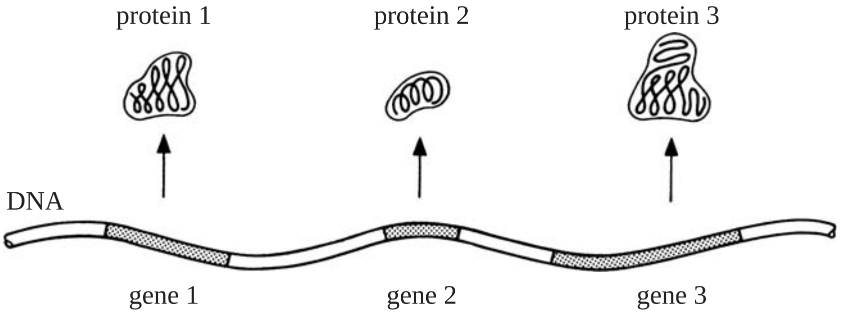

Figure 1.10 Genes in the DNA code for proteins. The DNA thread shown here represents the double helix depicted in Fig. 1.6, but here it is about 100 times longer. Both genes and proteins can vary in size, in direct proportion to one another. Each protein molecule is shown with a schematic chain-like structure.

tendons, and finger-nails. Other proteins, or 'enzymes',[1] catalyze a large number of chemical reactions, such as digestion of food or the synthesis of hormones. Other proteins carry oxygen in the blood, while still others form the protein spools around which DNA wraps in a chromosome. Where do all these proteins come from? How does a cell know which proteins to make?

It turns out that the DNA runs a very definite 'program' every time the cell needs to make a given protein molecule. The program tells the cell exactly what kind of protein to make, and approximately how much of it. This program is the well-known 'gene' of classical genetics; and each DNA molecule contains many genes along its length, as shown in Fig. 1.10. Sitting in our train traveling from Cambridge to London, we could see thousands of them as we looked out of the window watching the DNA go by!

The proteins, although very different from one another in their physical and chemical properties, share a common scheme of construction: they are all long-chain molecules, consisting of a single, unbranched chain that is made from the end-to-end joining of many small units known as 'amino acids'. Proteins are made up from amino acids, just as DNA is made up from nucleotides. For example, the sweetener in Diet Coca-Cola is a very tiny protein made by man from just two amino acids; while hemoglobin in the blood is a much larger protein, made from 600 amino acids. Altogether there are 20 different kinds of amino acid that make proteins; and this wide variety of building blocks allows for the construction of very many different proteins, with their enormous range of physical properties.

A program or gene in the DNA tells the cell in what order to assemble these amino acids and for what length of protein chain. The order of amino acids in a protein is set by the order of nucleotides in the corresponding piece of DNA. Because there are 20 possible amino acids, and only four possible nucleotides, the cell must use more than one nucleotide in the DNA to specify each amino acid in a protein. The universal rule is that *three nucleotides* specify *one* amino acid. The length of protein chain varies with the length of the DNA 'program', and is typically from about 100 to 1000 amino acids.

The cellular machinery that enables a certain set of three nucleotides in the DNA to be associated with only one of the 20 possible amino acids is extremely complex. Likewise, the conversion of the magnetized stripes on a regular computer tape into printed characters on a page of computer output must be rather complicated. But we do not have to worry about this kind of machinery in order to explain the simple code or cipher by which DNA makes proteins. Usually, this cipher is presented in the form of a table called 'The Genetic Code' (Table 1.1). This same code is used for specifying

Table 1.1 The Genetic Code

1st base	2nd base				3rd base
	T	C	A	G	
T	Phe	Ser	Tyr	Cys	T
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	T
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	T
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	T
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

Any series of three bases (or nucleotides) in the DNA prescribes for an amino acid in the protein chain, or gives a 'stop transcribing' signal. The bases are always read from left to right. The chain usually starts with ATG or methionine (Met). Abbreviations used: A, adenine; G, guanine; C, cytosine; T, thymine (or U, uracil in RNA). Ala, alanine; Arg, arginine; Asn, asparagine; Asp, aspartic acid; Cys, cysteine; Gln, glutamine; Glu, glutamic acid; Gly, glycine; His, histidine; Ile, isoleucine; Leu, leucine; Lys, lysine; Met, methionine; Phe, phenylalanine; Pro, proline; Ser, serine; Thr, threonine; Trp, tryptophan; Tyr, tyrosine; Val, valine.

proteins in all living things, whether they are bacteria, plants, or animals, with only a few minor exceptions.

The four different nucleotides in DNA are called adenine, guanine, cytosine, and thymine, or simply A, G, C and T. The 20 possible amino acids in a protein have names such as 'methionine', usually abbreviated to 'met' or 'M'. As shown in Table 1.1, each possible set of three nucleotides in the DNA specifies one amino acid. For example, 'TTT' specifies phenylalanine, while 'AAA' calls for lysine, and 'GCT' gives alanine. In each case, the letters are read from left to right. There are 4 × 4 × 4 = 64 combinations of DNA triplet, and often two of these triplets, such as 'AAG' and 'AAA', specify the same amino acid. Thus, a series of nucleotides, such as 'TTTAAAAAGGCT', specifies a portion of protein with amino-acid sequence 'Phe–Lys–Lys–Ala', as shown in Fig. 1.11. Certain triplets along the length of DNA also do the special work of telling the protein chain where to start and stop.

Finally, the completed protein chain usually folds by itself into a precisely determined shape, which depends on the exact arrangement of the amino acids in its chain. A correct three-dimensional shape is essential for the physical or chemical activity of a protein.

You should not think that all DNA does is to make proteins like a piece of computer tape. In fact, only a small fraction of the long DNA molecule in a chromosome – about 1 percent in humans – contains programs to make specific proteins. The vast majority of DNA in our bodies does things that we do not presently understand. There is plenty of room here for people to make new discoveries.

We said above that we can safely ignore the complex machinery which the cell uses in order to make proteins. This is true for the most part, but we do need to know about a few features of the process. We described above how the DNA-containing chromosomes are surrounded by a membrane in the nucleus of every cell

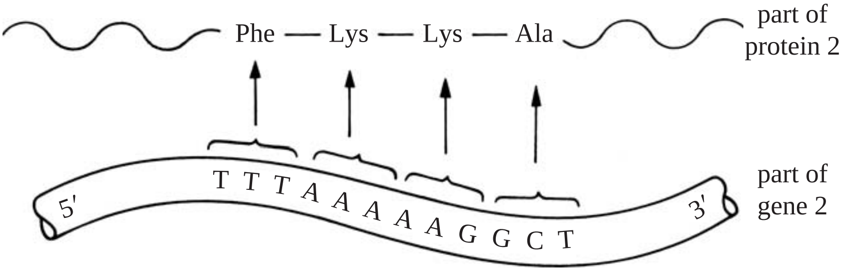

Figure 1.11 Part of the gene codes for part of the protein chain. Each set of three nucleotides (or bases) in the DNA specifies one amino acid, according to the scheme of Table 1.1.

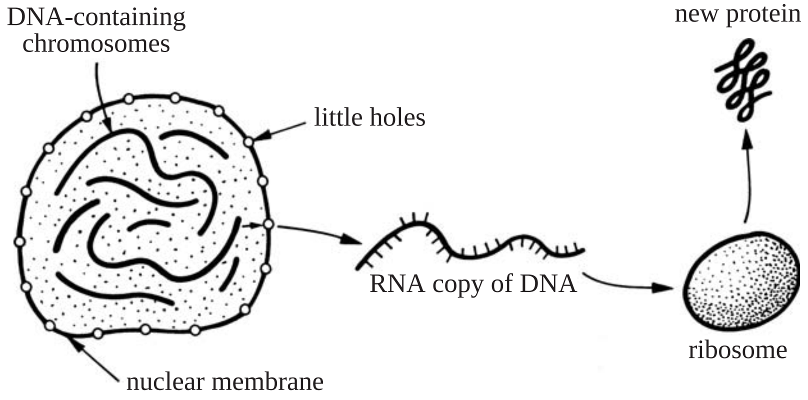

Figure 1.12 From DNA to RNA to protein. The DNA is copied into 'messenger-RNA', which then travels outside the nucleus to the ribosome, where it specifies the assembly of some particular protein, according to the series of nucleotides in its chain as in Table 1.1. For simplicity, we have not mentioned here that many genes in higher organisms contain substantial amounts of non-coding DNA, or 'introns', that are cut out at the level of RNA before the RNA copy of DNA leaves the nucleus.

(except in bacteria[1] – and bacteria-like single-cell microbes called archaea[1], where there are no nuclear membranes). It has been discovered that proteins are made in the cellular space *outside* of this membrane. What happens is that the DNA of a particular gene first makes a copy of itself that can pass through the small holes in the membrane; then this copy goes off to the protein-making machinery, which is called a 'ribosome'. This process is shown schematically in Fig. 1.12. These copies of the DNA program are made from a slightly different kind of molecule called 'RNA'. At the level of the picture shown in Fig. 1.8, RNA has an extra oxygen atom on the sugar ring as compared with DNA. A second difference is that RNA lacks a carbon and three associated hydrogen atoms on the thymine (T) base, and so this base is renamed 'uracil' (U). The RNA copy itself is called 'messenger-RNA' because it carries instructions on how to make a particular protein from the chromosome to the ribosome.

Interestingly enough, messenger-RNA makes up only a small fraction of the total RNA in any cell, about 5 percent of the total; and there is also much more RNA than DNA altogether. A lot of RNA is copied from DNA but never used to make protein. These abundant RNA molecules are different from messenger-RNA, and are mainly of two types: 'transfer-RNA' and 'ribosomal-RNA'. Transfer-RNA carries single amino acids from elsewhere in the cell to the ribosome; there the amino acids, while still bound to the transfer-RNA, can pair up with specific triplets of nucleotides on the messenger-RNA chain, and so make a protein in accordance with the cipher of Table 1.1. In other words, transfer-RNA performs the function of the vertical arrows shown in Fig. 1.11, once the DNA has been copied to messenger-RNA. We explain more about this in Chapter 2. Ribosomal-RNA makes up the bulk of the ribosome in association with a few proteins; it helps to align a series of transfer-RNA molecules along a chain of messenger-RNA, so that a series of amino acids can be joined chemically to form a long protein. It seems remarkable that so many steps in the synthesis of protein involve an RNA intermediate. This observation has led many people to speculate that life on Earth began with RNA, rather than with DNA, as the substance of genes.

Now we have presented all of the essential background information that is required for you to understand how DNA works in biology. Other books could be written on RNA or protein, but here we focus exclusively on the role of DNA. We shall work from small to large, starting with the basic chemistry of DNA and ending with the role of DNA in medicine and the new field of 'epigenetics'. Despite much recent progress, the great revelations in biology will no doubt belong to the new twenty-first century; we cannot pretend that our present knowledge is any more than a rough and incomplete foundation on which others can build.

Note

[1] The linguistic or historical derivations of words marked thus are given in Appendix 1.

Further Reading

Alberts, B.M., Johnson, A., Lewis, J., Raff, M.C., Roberts, K., and Walter, P. (2002) *Molecular Biology of the Cell* (4th Edn). Garland, New York. A good general reference volume.

Chargaff, E. (1963) *Essays on Nucleic Acids*. Elsevier, New York. An important historical document describing the state of knowledge about DNA in the late 1950s.

Crick, F.H.C. (1963) The recent excitement in the coding problem. *Progress in Nucleic Acid Research* **1**, 163–217. A good scientific review of how the Genetic Code was discovered.

Crick, F.H.C. (1988) *What Mad Pursuit*. Weidenfeld & Nicolson, London. An anecdotal history of the discovery of the structure of DNA and of the Genetic Code. Also, the source of the quotation used in the Preface.