

1

The Biological Roles of the Nucleic Acids

Aims

By the end of this chapter you should understand:

- What is meant by the term genetic information
- That there are two types of nucleic acids called DNA and RNA
- That genetic information is encoded in the structure of DNA
- How the genetic information is expressed

1.1 Introduction

This book is intended mainly for students of chemistry, and so the emphasis is on the chemistry of the nucleic acids. It is, however, difficult to talk about the chemistry of these molecules without reference to their biological properties and functions. Indeed, some of the methods used to determine the structures of nucleic acids make use of those biological properties (see Chapter 5). In addition, of course, the biology of the nucleic acids is a fascinating subject because they are the molecules on which the continuation of life depends.

Most readers of this book will have studied at least a little biology at school, and will probably be aware in outline of what DNA is and what it does. Indeed, living in the modern world it is difficult to avoid hearing such terms as genes, genomes, genetic engineering and DNA fingerprinting in general usage. Nevertheless, it seems a good idea to start off with a brief account of what the nucleic acids do, and how they do it, to set the scene for what comes after. It will be easier to understand the significance of individual parts of the chemical story if the student has a broad overview of the biology. Students who are familiar with the topic may want to skip this chapter and move straight on to the more chemical material that follows. On the other hand, readers who are interested in

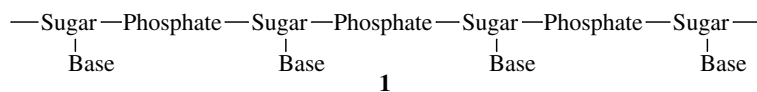
Molecular biology is a somewhat loose term. Logically it should mean the study of all life phenomena at the molecular level. It is more commonly taken to mean the study of the nucleic acids and protein synthesis.

the biology can find more extensive treatments in the books listed under Further Reading.

As well as a basic overview, this chapter (and later ones to a lesser extent) also contains a substantial amount of boxed material which essentially has to do with the early history of the development of ideas about **genetics** and **molecular biology**. This is partly to correct a common misapprehension that the study of DNA started in the last couple of decades of the 20th century. In fact it started in the middle of the 19th century, and it seems a shame not to be aware of the contributions that some of the major figures in the field made in those earlier years. In addition, it might be thought unsatisfactory to accept ideas such as DNA being the carrier of genetic information without knowing something about the evidence on which the claim is made. Study of this historical material is not essential to understanding the bulk of the text, but not to do so risks failing to appreciate the importance of the contributions made by earlier generations of scientists working in the field.

1.2 Classes of Nucleic Acids

We need to recognize from the start that there are two classes of nucleic acids. Both are polymers having a backbone of alternating monosaccharide and phosphate units, each of the sugars carrying one of four possible heterocyclic bases (shown schematically in **1**). The main difference is that in one class the polymer backbone contains 2-deoxyribose, and the molecule is referred to as **deoxyribonucleic acid** or **DNA**. In the other class the deoxyribose is replaced by ribose, and correspondingly the molecule is referred to as **ribonucleic acid** or **RNA**. We will use these abbreviations throughout the rest of the book. There is also a difference in that DNA and RNA have three of the heterocyclic bases in common, but one is different. We will return to this in Chapter 2, which deals with the details of the covalent structures of the nucleic acids.



Box 1.1 The Discovery of DNA

DNA was discovered in 1869 by Friedrich Miescher. Miescher was born in Basel, Switzerland, and trained as a physician, but decided

to make his career in scientific research. He was accepted to work in the laboratory of Felix Hoppe-Seyler in Tübingen. Hoppe-Seyler was one of the first scientists to specialize in the study of physiological chemistry. Miescher initially decided to work on the proteins in lymph cells; proteins had been discovered about 30 years previously, but their biological roles were only very poorly understood. The first problem that arose was that cells could only be obtained from lymphoid tissue in very small amounts. Because of this, he turned his attention to pus cells. These could be isolated in relatively large quantities from discarded surgical bandages obtained from a nearby clinic.

The only technique that Miescher had available for separating cellular components from one another was differential solubility in salt solutions. His key observation was that after extraction of the cells with alkaline solutions followed by neutralization, a precipitate was formed which had none of the properties of proteins. He then went on to show that this new material was located in a sub-cellular structure called the nucleus and, because of this, named it **nuclein**. At that time, of course, techniques were not available to allow the structure of a complex substance like nuclein to be determined. Methods were, however, well developed for elemental analysis, and Miescher made the important discovery that nuclein contained phosphorus. The significance of this was not to be realized for many years. Miescher's view was that nuclein was probably the cellular store of phosphate, from which it was released as required for other functions! What he had, in fact, isolated was a complex of DNA and protein. The real function of nuclein (or the DNA component of it) as the carrier of genetic information had not been established by the time Miescher died in 1895.

During the second half of the 19th century, important advances were being made in understanding the role of the nucleus. In 1842, Karl Nägeli had reported rod-like structures in the nuclei of plant cells—structures that we would now refer to as **chromosomes**. Nuclear division had also been observed, and it was becoming clear that the nucleus played a central role in maintaining the life of a cell. In 1866, Ernst Haeckel took the important step of claiming that the nucleus was responsible for the transmission of hereditary characters. An important contribution to taking matters forward was made by Paul Ehrlich who, in 1870–1880, developed a series of dyes derived from coal tar that could be used to stain individual sub-cellular components (thus paving the way for what is now known as **histochemistry**). This new technique was used to great effect by

Walther Flemming. He observed intensely stained material in the nucleus, which he termed **chromatin** (from the Greek *chroma*, meaning colour), and observed that the chromatin broke up into two portions, one of which was transmitted to each of the daughter nuclei on cell division, and so might carry the genetic instructions. There was, however, still a problem in that chromatin was shown to contain both protein and nuclein (or nucleic acid, as we shall call it from now on) and it was not clear which of these two components was involved in transmission of hereditary characteristics. Popular feeling favoured protein as the active component. This was disputed by E. B. Wilson who, in 1900, concluded that nucleic acid was the active component of chromatin, but it was still another 40 years before the role of DNA as the genetic material was finally established.

The only exception to the rule that the genetic information is carried by DNA is provided by some **viruses** which have genomes consisting of RNA. These, not surprisingly, are called **RNA viruses**. Viruses are infectious agents that can only reproduce in other living cells. Generally a virus is specific for a given cell type in one particular living organism. Some viruses infect bacterial cells; these are called **bacteriophages**, or sometimes simply **phages**.

1.3 DNA as the Carrier of Genetic Information

The offspring of two human beings is also a human being, with all the essential characteristics of that species. This is because both the egg and the sperm carry a set of instructions for making a new human being. Similarly, when a cell divides, the result is two identical cells each with the same **genetic information**. This genetic information is encoded in DNA molecules.

Box 1.2 The Discovery that Genetic Information is Carried by DNA

By the early 1900s it was known that the genetic information was carried on the chromosomes, and that the latter were composed of protein and DNA. It was, however, widely believed that the genetic information was a property of the protein component of the chromosomes. This was partly because proteins were already known to have complex structures and a variety of chemical activities, whereas the nucleic acids appeared to be simpler in structure and to be chemically unreactive. It was not until 1944 that the true situation was established by studies on **bacterial transformation**.

A key observation was made in the 1920s by an English physician, Frederick Griffith. He was studying the bacterium *Streptococcus pneumoniae*, which is the causative agent of pneumonia.

Griffith observed that one strain of the bacterium, when grown in culture, produced colonies of smooth cells (the S-strain), whereas another (the R-strain) produced colonies of cells with a rough appearance. The difference between the two is now known to be that the S-strain has a polysaccharide coat, but the R-strain does not. Of central importance was the finding that when the S-strain was injected into mice, it caused disease and the mice died within a day, whereas the R-strain did not cause disease. The essential difference is that the polysaccharide coat of the S-strain protects it from the immune defences of the animal, and allows the bacterium to proliferate.

Griffith then tried to produce a vaccine against the S-strain by first killing the bacteria by heating them, and then injecting the killed bacteria into animals. The heat-killed bacteria did not cause disease. The astonishing observation was, however, that if he injected killed S-strain cells along with the R-strain, then the animals developed pneumonia. Moreover, the blood of the animals contained living bacteria with the appearance of the S-strain. It appeared that a transformation of the R-strain to the S-strain had occurred.

The story was taken up by Oswald Avery and his group in the USA. The first important breakthrough was the demonstration that transformation could be carried out in bacterial cultures. This allowed the phenomenon to be studied under carefully controlled conditions. The next step was to rupture cells of the S-strain, extract the transforming material, and find out to which class of molecule it belonged. Avery and his colleagues treated samples of the transforming material with agents known to degrade proteins, nucleic acids, polysaccharides and lipids. The result was that if the DNA was destroyed, the transforming activity was lost. No other component of the extract was required. These results showed that the DNA alone was the transforming factor.

Although this groundbreaking work was published in 1944,¹ the conclusions were by no means universally accepted. Many scientists believed that the nucleic acid preparation used in the transformation experiments contained trace amounts of protein, and that it was the protein which was the active component. The question was finally settled by experiments reported by Hershey and Chase in 1952.² The experiments involved the use of a bacteriophage called T2. This is a simple virus that consists of a strand of DNA packed into a protein coat. The question was: when the virus infects a bacterium, which of these two components enters the bacterial cell? The answer was provided by producing one batch of phage particles in which the protein was labelled with the radioactive isotope ³⁵S (sulphur is not

present in DNA), and another batch where the DNA was labelled with ^{32}P (phosphorus does not occur in the viral proteins). Bacteria were incubated with the ^{35}S -containing phage for a short time, after which the part of the phage that had not entered the bacteria was stripped away, and the radioactivity associated with the bacterial cells was measured. Very little was found. On the other hand, when the experiment was repeated with ^{32}P -containing phage, most of the radioactivity remained in the bacterial cells. This suggested that the DNA had entered the bacteria. More compelling, when the experiments were continued for longer periods so that progeny phage was produced, the progeny were found to contain ^{32}P but not ^{35}S . This provided compelling evidence that the genetic material of the phage was DNA, not protein. These experiments finally convinced even the most sceptical scientists that DNA was indeed the carrier of genetic information.

Humans, for example, have 46 chromosomes. In females these consist of 23 pairs, one of each pair being inherited from the mother and the other from the father. Both members of a pair contain essentially the same genetic information. In males there are 22 pairs. The two unpaired chromosomes, which are called X and Y, are the **sex chromosomes**, and it is the possession of the Y chromosome that confers maleness. Females have a pair of X sex chromosomes. The germ cells, or **gametes** (egg and sperm), contain only one set of chromosomes and in the case of the male, half of the sperm cells contain an X chromosome and the other half a Y. The egg cells, on the other hand, all contain X chromosomes. Fertilization with a Y-containing sperm produces a male offspring whereas fertilization with an X-containing sperm produces a female. Henry VIII was wrong to blame his wives for not producing male heirs for him!

The DNA in cells of higher organisms is contained in structures called **chromosomes**, each of which is composed of a (very large) molecule of DNA and many copies of several different proteins (see Section 3.4). The total DNA of an organism is referred to as its **genome** and the individual units of information in the genome are called **genes**. In bacteria, most of the genetic information is contained in a single chromosome, but many bacteria also contain extra genetic information in small DNA molecules called **plasmids**.

Box 1.3 The Discovery of Genes

Modern ideas about genetics are inextricably linked with the name of Gregor Mendel. He was born in 1822 in Moravia (then part of Austria). As a young man he studied natural and agricultural sciences, but when his family could not afford to support him further, he entered an Augustinian monastery in Brno and became a priest in 1847. He did not, however, abandon his interest in science. He spent the years 1851–1853 in further studies at the University of Vienna and then returned to Brno where, over the next 10 years, he carried out his classic experiments on pea breeding.

What Mendel did was to study the results of cross fertilization of pea plants and to observe some of their heritable **characters** and **traits**. He obtained parental strains that were true breeding for each

of the traits studied (that is, they produced only that trait over many generations) and then transferred pollen from one strain onto the stigmas of the other strain. These plants were referred to as the **parental generation (P)**. When seeds developed in the parental strain, they were collected and planted to produce the **first filial generation (F₁)**. Mendel examined each F₁ plant and recorded the traits that it expressed.

Taking seed shape as an example, Mendel crossed plants with smooth seeds and plants with wrinkled seeds. He found that all the seeds of the F₁ plants were smooth; the wrinkled characteristic seemed to have disappeared. The next year, he grew plants from each of these seeds and allowed them to self-pollinate to produce a **second filial generation (F₂)**. He then examined the F₂ seeds and found that, of about 7500 seeds produced, almost exactly three-quarters were smooth and one-quarter wrinkled. He obtained the same result with a variety of other characters.

Mendel concluded from these experiments that the hereditary units responsible for any given trait exist as a pair of particles that separate from one another when the reproductive cells (gametes) are formed. These units are what we now call **genes**. Fertilization then results in a cell that contains one unit of inheritance from each of the gametes. From the results obtained with the F₁ plants he concluded that the smooth seed trait was **dominant**, and the wrinkled seed trait was **recessive**. What this means is that the wrinkled trait is only expressed if the plant has two copies of the recessive gene. If both the recessive and the dominant genes are present, then the trait expressed is that of the dominant gene; that is, smooth.

Mendel's results can be thought of as follows. Each of the cells of a plant of the parental generation with smooth seeds contain two copies of the dominant gene (let us call it *S*), whereas the cells of the plants with wrinkled seeds contain two copies of the recessive variant of the gene (which we can call *s*). These different forms of the same gene are called **alleles**.

The gametes of a plant with smooth seeds each contain a single *S*, whereas those from a plant with wrinkled seeds contain a single *s*. When the plants are cross-fertilized, the F₁ generation will have the genetic constitution *Ss*, but will have smooth seeds because the gene for smoothness is dominant.

How does this theory explain the results obtained with F₂ plants? The F₁ plants have the **genotype** *Ss*, so half of the gametes of these plants contain the *S* allele and the other half contain the *s* allele. On self-pollination, the random combination of gametes produces equal

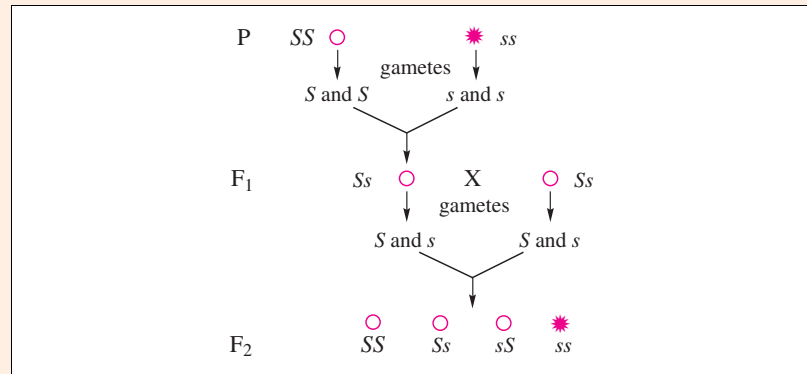
A **character** is a feature such as seed shape, and a **trait** is an example of such a character. In the case of seed shape, the traits are smooth or wrinkled.

Genotype is the term used to describe the precise genetic constitution of an individual. **Phenotype** is the term used to describe the observable properties of an individual that result from expression of the genotype under a particular set of environmental conditions.

Figure 1.1 Mendelian inheritance of seed shape in peas. The seeds are either smooth (○) or wrinkled (●). Parental plants with smooth seeds have the genotype SS and produce gametes with the S allele. Parental plants with wrinkled seeds have the genotype ss and produce gametes with the s allele. The F_1 progeny are all Ss and smooth (because S is dominant). They produce equal numbers of gametes with S and with s . Self-fertilization yields F_2 progeny with genotypes SS , Ss , sS and ss in equal numbers. The ratio of smooth to wrinkled phenotype in F_2 is 3:1

The discovery of linkage of genes on the same chromosome was largely due to the work of Thomas Hunt Morgan. He was awarded the Nobel Prize in Physiology or Medicine in 1933 "for his discoveries concerning the role played by the chromosome in heredity". Note that many references will be made in the following pages to winners of the Nobel Prize. Each prizewinner delivers a Nobel Lecture, and these lectures give fascinating insights into the discoveries that the laureates made. The lectures can be accessed in PDF format (for viewing in Adobe Acrobat Reader) at the Nobel Museum website. For a particular lecture, the address to use is <http://www.nobel.se/prize/laureates/year/surname-lecture.pdf> where *prize* will be either *medicine* or *chemistry*. So for Morgan's lecture it is <http://www.nobel.se/medicine/laureates/1933/morgan-lecture.pdf>.

numbers of SS progeny and of ss progeny, and twice that number of Ss (because Ss is the same as sS). All progeny with at least one S allele will have the smooth **phenotype**, but only the ss progeny will have the wrinkled phenotype. Hence the ratio of smooth to wrinkled seeds will be 3:1, just as Mendel observed. The SS and ss plants are said to be **homozygous**, and the Ss plants are **heterozygous** for the trait in question. A summary of the experiments is given in Figure 1.1.



Mendel went further than described above. He also showed that, for the traits he was considering, the alleles of different genes assort independently. Consider now two traits such as seed colour and flower colour, the genes for which we can call A and B . A heterozygote for both these two traits will have the genotype $AaBb$. The question is: when gametes are formed, does A always go with B , and a with b , to yield two types of gametes (AB and ab)? Or, alternatively, can four types of gametes be formed with the genotypes AB , Ab , aB and ab ? Mendel found the latter to be the case, which reinforced the idea of genes being independent entities. In fact, it is now known that **independent assortment** always occurs if the genes concerned are on separate chromosomes; that is, it is the chromosomes, each carrying many genes, which assort independently. Genes on the same chromosome usually, but not always, segregate together; they are said to be **linked**. The closer they are together on the chromosome, the more strongly they will be linked.

Remarkably, although Mendel's work was published in 1866, it was virtually ignored for over 30 years. This may be because it was published in Brno in a journal, *Proceedings of the Society of Natural Sciences*, which was not well known to other people working in the

field. It was not until the beginning of the 20th century that other scientists carried out similar experiments and rediscovered what Mendel had done. Within a few years it was recognized that the chromosomes carried the genes that Mendel had discovered, but the connection between chromosomes and DNA was still not established at that stage.

The genetic material has two essential characteristics. Firstly, it must be capable of being copied exactly. So, for example, when a bacterial cell divides, two identical copies of the DNA that it contains must be synthesized so that one copy can be passed to each of the daughter cells produced. How the structure of DNA allows for its replication is explained in Chapter 3.

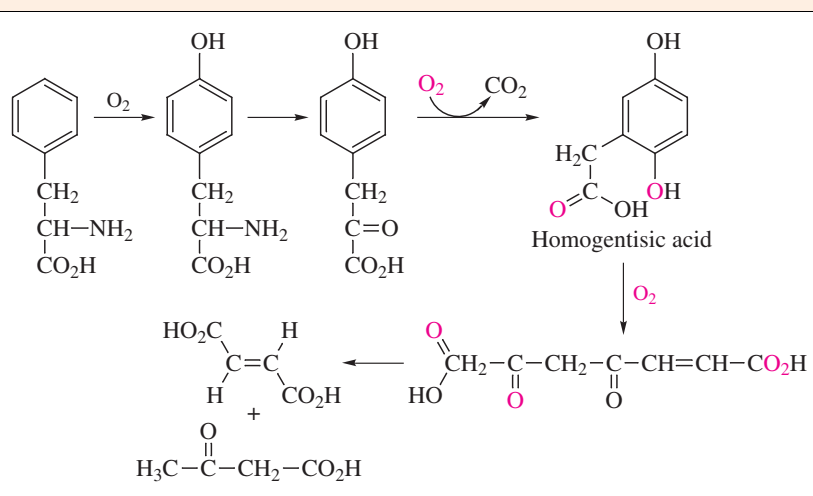
The other characteristic is that the genetic information must be **expressed**. That is, the information that the DNA contains must be interpreted in some way so that the cell in which it is contained does the right things. The information in DNA is largely a set of instructions for making **proteins**, but in addition it also codes for the structures of some RNA molecules which play a central part in protein synthesis (see Sections 1.5 and 1.7).

Box 1.4 The Discovery that Genes Code for Proteins

Although the work of Mendel and his successors established the idea of genes as the agents of transmission of inheritable characteristics from parents to progeny, there was, at the end of the 19th century, no indication of how the genotype gave rise to a particular phenotype—that is, how genetic information was expressed.

The earliest studies that eventually provided the answer to this question were carried out by Archibald Garrod, who was a physician working at St. Bartholomew's Hospital in London. Garrod was interested in a disease condition in which the urine, on exposure to air, turned black. This condition was termed **alkaptonuria**. It was found that this was due to the presence in the urine of the compound homogentisic acid (see Scheme 1.1). Garrod found that the unaffected parents of sufferers from alkaptonuria were often blood relatives, and suggested that these people were the carriers of a rare recessive gene which, when inherited homozygotically, resulted in expression of the disease. Garrod went on to propose that the disease arose from the lack of an enzyme involved in normal metabolic processes.³

This implied that some genes coded for enzymes, one of the major classes of proteins that by then were beginning to be understood.



Scheme 1.1

Conditions such as alkaptonuria are referred to as **genetic diseases**, and the result of their inheritance is called an **inborn error of metabolism**. There are many such conditions known, but each of them is relatively rare because both parents must be carriers of a recessive allele for the condition. This is more likely if the parents are blood relatives, which is why consanguineous marriage is forbidden in many countries. One of the best-known genetic diseases is phenylketonuria. In this condition, the enzyme that converts phenylalanine to tyrosine is missing (the first reaction in Scheme 1.1). If untreated, the results are severe mental defects and early death. The disease is treatable by feeding a diet low in phenylalanine. About 1 in 12,000 newborn infants have the disease, and it is common practice in many countries to test all babies for the disease at birth.

Garrod was right. The metabolic pathway by which the amino acids phenylalanine and tyrosine are degraded to a mixture of *trans*-butenedioic acid and 3-oxobutanoic acid is shown in Scheme 1.1. Each of the steps in the process is catalysed by a specific enzyme. The enzyme catalysing the conversion of homogentisic acid to 4-maleylacetoacetic acid is called homogentisate oxidase. The reaction involves molecular oxygen, as does the previous step in the pathway, and probably proceeds *via* an epoxide intermediate; the incorporated oxygen atoms are shown in red. It is the homogentisate oxidase that is missing in alkaptonuria. In heterozygotes, the effect of the defective gene is masked because sufficient of the enzyme is produced by the non-defective gene to satisfy the needs of the cell.

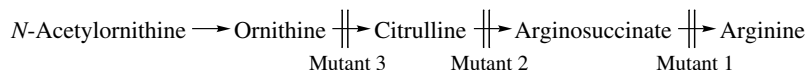
Important as they were, the significance of Garrod's results was not fully realized until considerably later, when further evidence had accumulated that genes code for proteins. Much of that evidence was obtained by George Beadle and Edward Tatum working in the USA. Their most significant studies were carried out using a fungus, *Neurospora crassa*, which is found in bakeries. What they did was to irradiate *Neurospora* cells with X-rays to generate mutants that would not grow on a minimal growth medium (that is, one containing only a carbon source and minerals), but would grow on media supplemented with materials such as vitamins or metabolic intermediates. Beadle and Tatum characterized about 100 genes, mutation of which altered the growth requirements of the fungus,

and showed that they coded for enzymes involved in the synthesis of amino acids, components of nucleic acids and vitamins.⁴

As an example of the sort of results that they obtained, we will consider mutations that affected the ability of the fungus to biosynthesize the amino acid arginine. The **wild-type** (WT) fungus can synthesize this substance, but several mutant strains were isolated that could not. The mutants could grow, however, if the growth medium was supplemented with various intermediates on the biosynthetic pathway. Typical results are shown in Table 1.1, and part of the biosynthetic pathway is shown in Scheme 1.2.

Table 1.1 Analysis of growth condition for wild type and arginine mutants of *Neurospora crassa*. The symbol ✓ indicates that growth occurred, whereas × indicates that no growth occurred

Fungal strain	Growth supplement				
	None	Arginine	Argino-succinate	Citrulline	Ornithine
Wild type	✓	✓	✓	✓	✓
Mutant 1	×	✓	×	×	×
Mutant 2	×	✓	✓	×	×
Mutant 3	×	✓	✓	✓	×



Scheme 1.2

All of the fungal strains grew on a medium containing arginine, but only the wild-type strain grew on a minimal medium not containing arginine. Mutant 1 did not grow when the medium was supplemented with any of the three intermediates arginosuccinate, citrulline or ornithine, and so lacked the final enzyme in the biosynthetic pathway (see Scheme 1.2). Mutant 2 grew when the medium was supplemented with arginosuccinate, but not when it was supplemented with citrulline or ornithine; hence it lacked the enzyme required to convert citrulline to arginosuccinate. Finally, mutant 3 grew on media supplemented with either arginosuccinate or with citrulline, but not on a medium supplemented with ornithine; hence it lacked the enzyme required to convert ornithine to citrulline. The assignment of a missing enzyme to a particular mutant was

checked by making an extract of the mutant, and checking for the presence or absence of the enzyme activity in the extract.

The results obtained by Beadle and Tatum led to the “**one gene, one enzyme**” hypothesis; that is, a gene exists for every one of the many enzymes found in any living organism. More generally, we would say that a gene exists for every polypeptide chain in the organism. Beadle and Tatum were awarded the Nobel Prize in Physiology or Medicine in 1958 for “their discovery that genes act by regulating definite chemical events”.

1.4 An Outline of Protein Structure

The chemistry of proteins has been dealt with in detail in another volume in this series.⁵ However, because of the essential connection between proteins and nucleic acids, it is necessary to give a brief outline of protein structure here.

Proteins are polymers made from 19 α -amino acids (**2**) and the imino acid proline (**3**). What distinguishes one amino acid from another is the nature of the side chain (the group R in structure **2**). Table 1.2 gives a list of the 19 amino acids that are specified by the genetic code along with the structures of their side chains. Also given are two abbreviations for each amino acid. The first is a three-letter abbreviation which is generally the first three letters of the name of the amino acid. The second is a single-letter code. The initial letter of the name is used for some of the amino acids (generally those that occur most commonly in proteins), but this is not always possible because there are several cases where two or more amino acids have the same initial letter. So, for example, the abbreviation for alanine is A but that for aspartic acid is D. It might seem unnecessarily confusing to use the single-letter codes but, as we will see later, they are very useful when the structures of large proteins have to be recorded either in paper form or electronically.

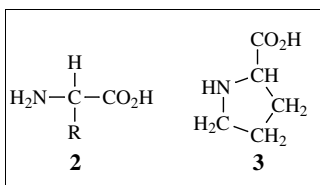


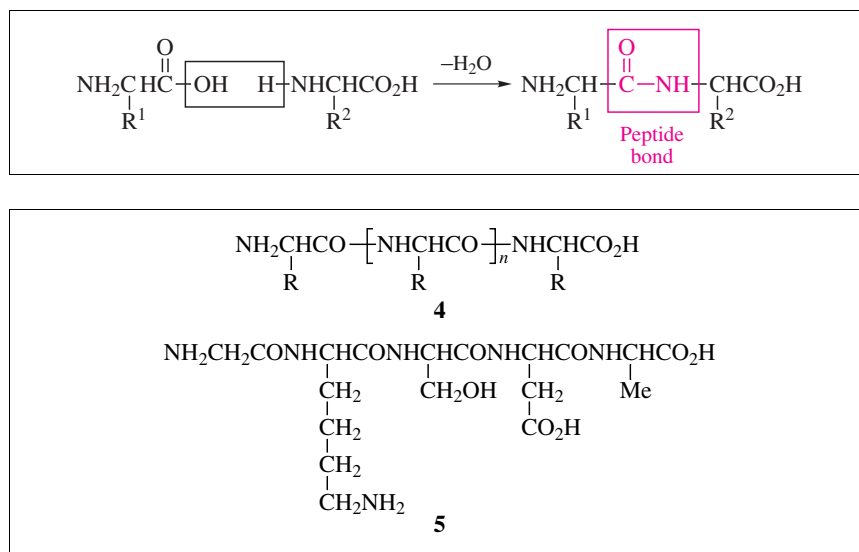
Table 1.2 The 19 α -amino acids occurring in proteins^a

Type of side chain	Name	Structure of side chain	Abbreviated name	One-letter abbreviation
Aliphatic	Glycine	—H	Gly	G
	Alanine	—Me	Ala	A
	Valine	$\begin{array}{c} \text{Me} \\ \\ \text{—CH} \\ \\ \text{Me} \end{array}$	Val	V

(continued)

protein has the general structure shown in (4); that is, it consists of a string of amino acids linked together by peptide bonds. The value of n in structure 4 can be as small as zero (when the molecule is called a **dipeptide**) or as large as several thousand. Small molecules with up to about 20 or 30 amino acids are generally referred to as **peptides** (or **polypeptides** or **oligopeptides** – these terms are interchangeable), whereas larger molecules are referred to as proteins. The point at which the nomenclature changes is not clear cut, and indeed is not very important; the important thing is that peptides are small proteins.

Scheme 1.3

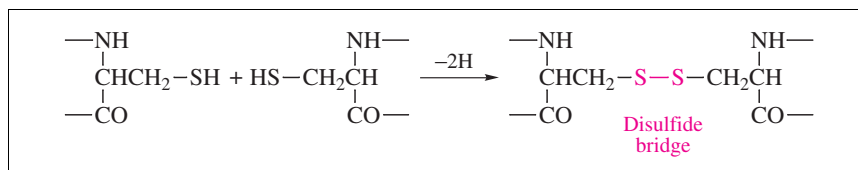
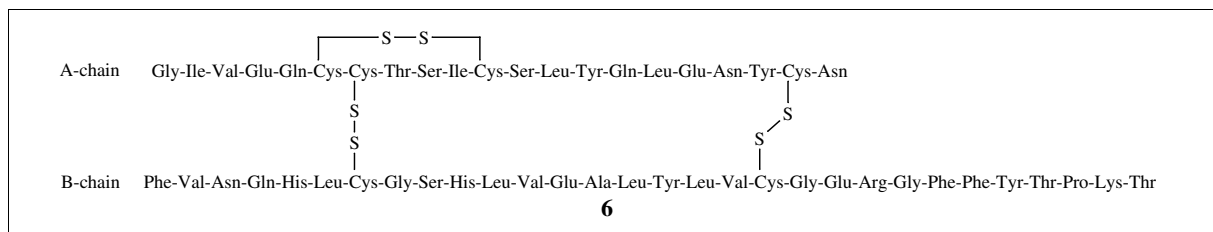


The word **residue** is used in recognition of the fact that the peptide does not strictly contain amino acids, but rather that bit of each amino acid that is left (the residue) when the peptide bonds are formed.

Note that there are 120 possible peptides with the same amino acid composition as 5. There are 5 ways of choosing the N-terminal residue. For each of these there are 4 ways of choosing the second residue, on so on. So the number of unique sequences is $5 \times 4 \times 3 \times 2 \times 1 = 120$.

An example of the structure of a pentapeptide (five amino acid residues) is shown in 5. The first important thing to note about this structure is that it has two unique ends. That at the left has a free α -amino group and is referred to as the **N-terminus**. At the other end there is a free α -carboxylic acid group; this is referred to as the **C-terminus**. Peptides and proteins are always represented this way around. This becomes important when the structures are written in shorthand form. Reference to Table 1.2 shows that the peptide in 5 contains one **residue** of each of the amino acids alanine, aspartic acid, glycine, lysine and serine; this is referred to as its **amino acid composition**. The structure can be written in shorthand as Gly-Lys-Ser-Asp-Ala or, even shorter, as GKSDA; these are two ways of writing the **amino acid sequence** or **primary structure** of the peptide. Both of these mean exactly the same to a protein chemist as the structure shown in 5. Note that it is now very important to observe the convention that the N-terminus is at the left. The peptide ADSKG has the same amino acid composition as 5 but has a different amino acid sequence; that is, it has a different structure (see Problem 1.1).

The usefulness of these shorthand representations becomes obvious as soon as we wish to present the structures of even small proteins. For example, **6** is the structure of the protein **insulin**. It consists of two chains of amino acids. The A-chain contains 21 residues and the B-chain contains 30. An interesting feature of insulin is that it contains **disulfide bridges**. These are formed by oxidation of pairs of cysteine residues by the process shown in Scheme 1.4; the product of the reaction is called cystine. These bridges are formed after the protein has been synthesized; the presence of cystine in a protein is not specified by the genetic code. Two of the three disulfide bridges link the A and B chains together, whereas the third is internal in the A chain.



Scheme 1.4

1.5 Transcription of DNA into RNA

In eukaryotic cells, protein synthesis takes place in the cytosolic compartment, whereas the genetic material is located in the nucleus. Moreover, individual genes are located on the chromosomes, which are very large structures. So how is the information contained in the gene for a particular protein transferred from the nucleus to the cytosol? The answer is that the section of DNA coding for the protein is **transcribed** into a molecule of RNA which contains the same information as that in the gene, but in a slightly different form. This RNA can leave the nucleus through pores in the nuclear membrane, and so carries the genetic message into the cytosol. Appropriately, molecules of this sort are called **messenger RNA** or **mRNA**. Even in prokaryotic cells, which do not have a nucleus, mRNA still performs this function of acting as an intermediary between DNA as the store of genetic information and the machinery where protein synthesis occurs.

The details of the process of transcription are dealt with in Chapter 4, but a brief outline is required here as an aid to understanding the following

Eukaryotes are defined as organisms whose DNA is contained within a sub-cellular structure, bounded by a membrane, and called the **nucleus**. They also contain other membrane-bounded sub-cellular structures. All eukaryotes contain **mitochondria**, which are the site of molecular oxidation within the cell. Plants have specialized structures called **chloroplasts**, which are the site of the process of **photosynthesis** by which the radiant energy in sunlight is used to fix carbon dioxide into carbohydrates. It is of interest in the present context that both mitochondria and chloroplasts contain small, but significant, amounts of DNA. The fluid portion of the cell is referred to as the **cytosol**. Organisms in which the genetic material is not contained in a nucleus (viruses and bacteria) are referred to as **prokaryotes**.

sections. As previously stated, DNA is a linear polymer made from four different monomeric units. For the moment we will simply represent the monomers by the letters A, G, C and T and look at their structures in detail in Chapter 2. RNA similarly is a polymer made from four monomeric units. Three of these are also A, G and C, but instead of T, RNA contains U. In transcription, a **complementary** mRNA is synthesized using a section of the DNA molecule as a **template**. The process is such that:

- Wherever A occurs in the DNA, U occurs in the RNA
- Wherever T occurs in the DNA, A occurs in the RNA
- Wherever G occurs in the DNA, C occurs in the RNA
- Wherever C occurs in the DNA, G occurs in the RNA

So, for example, a section of a DNA molecule and its mRNA **transcript** might have the base sequences shown below:

DNA : CTGAAGTCGTACCTGGGAATGTTTC
 mRNA : GACUUCAGCAUGGACCCUUACAAAG

The genetic message contained in the base sequence of the DNA molecule has been transcribed into the same message encoded in the base sequence of the mRNA.

1.6 How the Message is Decoded

Just as the structure of a protein is defined by the order in which its constituent amino acid residues occur in the polypeptide chain, so the structure of an mRNA molecule is defined by the order in which its constituent units occur in the so-called **polynucleotide chain**.

The problem is, then: how does a code consisting of a string of the four letters of mRNA become **translated** into the sequence of the 20 amino acids in the primary structure of a protein? The answer is that the bases are read in non-overlapping triplets, and each triplet specifies one amino acid in the protein. These triplets are known as **codons**. Given that the combination of any three bases out of four leads to 64 possible triplets, and there are only 20 amino acids to code for (strictly, 19 amino acids plus proline), there appears to be some redundancy in the system. In fact, it turns out that some, indeed most, of the amino acids are coded by more than one triplet.

The genetic code is shown in Table 1.3. The amino acid coded by any given triplet is obtained by identifying the set of four rows corresponding to the first base shown in the column on the left of the table, then finding the column corresponding to the second base, and finally the row within that column corresponding to the third base. For example, the triplet CAU codes for His (second set of four rows, third column, and then the

first entry). There are some points of special interest in this table. Firstly, three of the triplets (UAA, UGA and UAG) do not code for amino acids. Rather, they are **stop signals**, or **termination codons**; that is, when protein synthesis reaches one of these codons, the process stops and the last amino acid incorporated before this point becomes the C-terminus of the completed protein. What is the codon for the start of synthesis? It turns out that it is AUG. This triplet always codes for methionine, but depending on the context within the mRNA, it either signals the start of synthesis of a new protein, or for the insertion of an internal Met residue. This does not mean that all proteins have methionine at the N-terminus – this residue is usually removed to leave the amino acid coded by the triplet after the **initiation codon** as the N-terminus of the completed protein.

The direction of protein synthesis is from the N-terminus to the C-terminus.

Table 1.3 The genetic code

First base	Second base				Third base
	U	C	A	G	
U	Phe	Ser	Tyr	Cys	U
	Phe	Ser	Tyr	Cys	C
	Leu	Ser	STOP	STOP	A
	Leu	Ser	STOP	Trp	G
C	Leu	Pro	His	Arg	U
	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
A	Ile	Thr	Asn	Ser	U
	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met/ START	Thr	Lys	Arg	G
G	Val	Ala	Asp	Gly	U
	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

The other notable feature is that most amino acids are coded by at least two triplets, and some by as many as six. In many cases, all four triplets with the same first two bases code for the same amino acid. For example, Pro is coded by CCU, CCC, CCA and CCG. This has important consequences. For example, it means that if the third base in such triplets is **mutated** (see Section 3.6.1), then the amino acid incorporated in the protein chain does not change. On the other hand, if the triplet CAU was to be mutated into CAA, then the result would be a change in the amino acid incorporated from His to Gln, with possible functional effects on the protein coded by the gene.

A very important concept in protein coding is that of the **reading frame**. To see what is meant by this, consider a short stretch of mRNA in the middle of a gene transcript with the base sequence UUCCACAGU-GUUAUAUCCGGCUGGG. What does it code for? If we assume that the first base in the given sequence is the first base of a codon, then we can split the sequence up as shown below and look up the corresponding amino acids in Table 1.3:

|UUC | CAC | AGU | GUU | AUA | UCC | GGC | UGG | G
Phe - His - Ser - Val - Ile - Ser - Gly - Trp

Note that the final G is the first base of the next triplet. What if we assume that the first base in the given sequence is, in fact, the last one of the previous triplet? The sequence then splits up as follows.

U | UCC | ACA | GUG | UUA | UAU | CCG | GCU | GGG
Ser - Thr - Val - Leu - Tyr - Pro - Ala - Gly

The result is a completely different protein sequence because the reading frame has changed. Obviously, there is yet a third reading frame possible, where the first two bases belong to the previous triplet (see Problem 1.3). When translation of an mRNA occurs in the cell, the selection of the initiation codon fixes the reading frame for the rest of the message.

Translating RNA sequences into protein sequences by hand using the genetic code is very tedious, and also prone to errors! It is, however, a job well suited to computers, which do not get bored and do not make mistakes. This, and other applications of computers in molecular biology, will be discussed in Section 5.9.

Box 1.5 Deciphering the Genetic Code: Part A

The first steps towards breaking the genetic code were taken by Marshall Nirenberg in the early 1960s. It had already been shown that protein synthesis could be carried out by extracts of cells supplemented with adenosine triphosphate (ATP, see Chapter 2), guanosine triphosphate (GTP) and amino acids. This is called **cell-free protein synthesis**. Nirenberg showed that if the DNA present in the extract was destroyed by addition of a specific enzyme so as to prevent synthesis of new mRNA, then protein synthesis stopped, but could be restarted by addition of RNA. The crucial part of the work was the preparation of synthetic RNA molecules containing only one base, and the use of these to direct protein synthesis in the cell-free system. The first synthetic messenger to be made was poly-U, a repeating polymer of uridylic acid (see Chapter 2). When tested in the cell-free system, it was found that translation of the synthetic messenger produced a protein that contained only phenylalanine. Hence the codon UUU codes for phenylalanine.⁶ Subsequent experiments showed that poly-A resulted in incorporation of lysine,

and poly-C incorporated proline. These results assigned the codon AAA to lysine, and CCC to proline.

Although the use of polymers containing a single base was enormously important in that it allowed the first codon assignments to be made, it had obvious limitations. At about the same time, however, H. Gobind Khorana succeeded in making polymers with defined repeating sequences of more than one base. The interpretation of the results for incorporation of amino acids into proteins with such synthetic messengers was more complicated, but the method was very powerful.

Consider, for example, a repeating polymer containing U and C (poly-UC). This messenger has two codons, UCU and CUC, irrespective of the reading frame, as shown below for a short section:

|UCU|CUC|UCU|CUC| or U|CUC|UCU|CUC|UC or UC|UCU|CUC|UCU|C

Hence it would be expected to code for the synthesis of a protein with two alternating amino acids. This is what was found. In this case the amino acids incorporated were Ser and Leu. This means that UCU codes for Ser and CUC for Leu, or *vice versa*. In fact, the first of these assignments is correct (see below).

Khorana also synthesized polymers with repeating sequences of three bases, such as poly-UUC. Such polymers produced three protein products, each containing a single amino acid (see Problem 1.4). Finally, he also made polymers with a four-base repeat. Consider, for example, poly-UAUC. It is shown below with one possible reading frame marked:

UAU|CUA|UCU|AUC|UAU|CUA|UCU|AUC|UAU|CUA|UCU|AUC

It should be clear that this synthetic messenger contains four codons, and moreover the codons are the same irrespective of the reading frame. Hence the product of cell-free synthesis with this template should be a protein with a repeating sequence of four amino acids. In fact, the product that was obtained was a protein with the repeat sequence Tyr-Leu-Ser-Ile. If this result is compared with that for poly-UC described above, it confirms that UCU does indeed code for Ser. Comparisons of the results of many different experiments of this sort allowed the assignment of most of the codons to their appropriate amino acids.

Confirmation of these assignments, and completion of the interpretation of the genetic code, was done using a method that requires some knowledge of the mechanism of protein synthesis to understand. Protein synthesis is dealt with in Section 1.7, and the second method of codon assignment is described in Box 1.6.

The "S" in expressions such as 70S ribosome or 5S RNA is the **Svedberg unit**, and is a measure of the rate of movement of a particle through solution in a unit gravitational field. It is frequently used as a measure of the size of particles which are composed of many different types of subunit, and for which it is therefore difficult to define a relative molecular mass. It was also widely used in the early days of nucleic acid chemistry as a measure of the sizes of these molecules, because sedimentation coefficients are relatively easy to measure, whereas determination of the M_r values of nucleic acids was very difficult using hydrodynamic methods. Note that sedimentation coefficients are not additive.

1.7 Protein Synthesis

The processes of protein synthesis are extremely complicated and will not be dealt with in detail until Chapter 4, but a brief summary is required to complete the overview. The site of synthesis is a particle called the **ribosome**. This is a very complex structure made of protein and RNA. In prokaryotes, the ribosome has an M_r of about 2,700,000 and a diameter of about 20 nm, so it is a very large structure indeed. Ribosomes were originally characterized by their sedimentation coefficients, and that in prokaryotes is often referred to as the **70S ribosome**.

It can be dissociated into two parts, called the **30S subunit** and the **50S subunit**, which can be further broken down into their constituent protein and RNA components. The 30S subunit consists of 21 different protein molecules and a **16S RNA** species. The larger subunit consists of 36 protein molecules and two RNA molecules, one of 23S and the other of 5S. The structures of these RNA molecules are, of course, specified by the genetic information contained in DNA, and are synthesized in just the same way as is mRNA. The eukaryotic ribosome is somewhat bigger, but the structural features are similar and the differences need not concern us.

It used to be thought that the RNA molecules, which account for about two-thirds of the mass of the ribosome, were essentially structural and that the protein components were responsible for carrying out the reactions involved in protein synthesis. This view has now changed, and it has become clear that the RNA species play central roles in those chemical events. Catalysis of reactions by RNA is the topic of Box 4.2.

Let us turn to the events of protein synthesis. We know that the base sequence of mRNA contains the information for the amino acid sequence of the protein to be produced. The question is how this information is used to put the amino acids in the correct order. Again, RNA molecules are centrally involved. There exists a set of small RNA molecules (containing between 73 and 93 monomeric units) called **transfer RNA**, or **tRNA**, which act as **adaptors** between the mRNA and the amino acids which are to be inserted in the polypeptide chain. Each tRNA can be loaded with a specific amino acid, and each has a region in its structure that recognizes the triplet codon for the amino acid that it carries (see Section 4.4). In essence, then, the mRNA combines with the ribosome and provides the template for the protein chain to be built up. Each triplet of bases is recognized by a tRNA molecule carrying the required amino acid for that position in the protein chain, and the amino acid is attached to the growing chain. Eventually a termination codon is reached and synthesis stops.

Box 1.6 Deciphering the Genetic Code: Part B

The second approach to solution of the genetic code, again due to Nirenberg, was different in nature from that described in Box 1.5. Nirenberg discovered that synthetic trinucleotides promote binding of specific tRNA molecules to ribosomes. For example, the trinucleotide AAA, when added to ribosomes, promoted the binding of the tRNA specific for lysine. This confirmed that AAA codes for Lys. A very ingenious assay was developed to determine which tRNA was bound. Individual tRNA molecules were loaded with their specific amino acids, and then mixed together. In each mixture, one of the amino acids was radioactively labelled. The test trinucleotide was added to the ribosomes, followed by mixtures of tRNA loaded with amino acids; in each experiment, a different amino acid was labelled. It was then necessary to discover which of the tRNA molecules had bound to the ribosomes. This was done by passing the assay system through filters which retained ribosome-tRNA complexes, but allowed unattached tRNAs to pass through. The radioactivity was then measured on the filter and in the filtrate. In the test system where the trinucleotide was recognized by the tRNA carrying a labelled amino acid, the radioactivity would be retained on the filter. Otherwise, the radioactivity would be found in the filtrate. Again, use of this experimental approach allowed most of the codons to be identified. In combination with results obtained using protein synthesis in cell-free systems, the entire genetic code had been solved by 1966.

The 1968 Nobel Prize for Physiology or Medicine was awarded to Robert Holley, H. Gobind Khorana and Marshall Nirenberg for "their interpretation of the genetic code and its function in protein synthesis". Holley's contribution was concerned with the discovery and characterization of tRNA; this will be discussed in Section 4.4.

1.8 The "Central Dogma" of Molecular Biology

This is a phrase coined by Francis Crick (see Chapter 3 for an account of Crick's major contribution to the study of DNA) to emphasize the essentially unidirectional flow of information in living organisms. It can be summarized by the sequence:



That is, the information in DNA is transcribed into information in RNA, and the latter is then translated into the structure of protein. This is essentially the process that has been summarized in the discussion above. This scheme is now known to be incomplete, and should be properly be written as:



A note on the naming of enzymes might be useful. The vast majority of enzyme names end in "ase"; exceptions are enzymes like trypsin and pepsin that were discovered before systematic names were introduced. As far as possible, the name describes what the enzyme does and what it acts on. Thus homogentisate oxidase oxidizes homogentisate and reverse transcriptase reverse-transcribes DNA.

That is, in some circumstances, information can flow from RNA to DNA. This is a process restricted to certain viruses called **retroviruses**, of which the best known is, perhaps, the human immunodeficiency virus (HIV), which is thought to be the causative agent of AIDS. Organisms like this have an enzyme called **reverse transcriptase** which they use to transcribe their RNA genomes into DNA. Once this has happened, the normal machinery of the host cell transcribes the DNA to make multiple copies of the viral RNA, which can be used both as a messenger to make viral proteins, and as the genome for further copies of the virus.

The last part of the central dogma appears, however, to be sacrosanct. There is no known situation in which the information in protein structure can be translated back into the structure of a nucleic acid.

Summary of Key Points

1. There are two classes of nucleic acid, called DNA and RNA.
2. DNA is the carrier of genetic information; that is, of the instructions that are passed on from parents to progeny, and that are duplicated and passed to the daughter cells on cell division.
3. The genetic instructions are contained in individual units of inheritance called genes, and the genes are organized on structures called chromosomes.
4. The genes are mainly sets of instructions for making proteins; some genes code for RNA molecules.
5. For expression of the instructions in a gene, the DNA is first transcribed into a complementary messenger RNA.
6. The base sequence of the messenger RNA specifies the amino acid sequence of a protein.
7. The message encoded in the messenger RNA is read three letters at a time. Each group of three letters, called a codon, specifies a particular amino acid. There is a single codon that signals the start of the protein chain, and three codons that signal termination.
8. Protein synthesis occurs on complex structures called ribosomes that are complexes of RNA and proteins.

9. For incorporation into the protein chain, each amino acid is first linked to a specific molecule of transfer RNA. The transfer RNA has within its structure a region that recognizes the codon for that particular amino acid in the messenger RNA.
10. The transfer RNA, carrying its specific amino acid, binds to the messenger RNA and the amino acid that it carries is added to the growing protein chain.
11. Genetic information usually travels from DNA to RNA to protein. A class of viruses called retroviruses contain an enzyme that allows information to flow from RNA to DNA.

Problems

- 1.1. Draw the structure of the peptide ADSKG.
- 1.2. Translate the following piece of mRNA into the corresponding protein sequence written in the three-letter code, assuming that first base in the sequence is the first letter of a codon:

GAGCUCGUAAUUCUAUACUCAUGAAAAAUUAACGGG

Re-write the protein sequence in the one-letter code. Read the sequence as a sentence in English and comment on the statement that it makes! (I am indebted to Mr Malcolm Ward for this example).
- 1.3. In Section 1.6 a piece of RNA is shown translated in two possible reading frames. Give the translation in the third reading frame.
- 1.4. What protein product, or products, would you expect to be synthesized in a cell-free system programmed with poly-CAG?
- 1.5. When a cell-free system is programmed with poly-AUAG, the product formed is a tripeptide. Explain this result.

References

1. O. T. Avery, C. M. MacCleod and M. McCarty, *J. Exp. Med.*, 1944, **79**, 137.
2. A. D. Hershey and M. Chase, *J. Gen. Physiol.*, 1952, **36**, 39.
3. A. E. Garrod, *Lancet*, 1902, **2**, 1616.
4. G. W. Beadle and E. L. Tatum, *Proc. Natl. Acad. Sci. USA*, 1941, **27**, 499.
5. S. Doonan, *Peptides and Proteins*, The Royal Society of Chemistry, Cambridge, 2002.
6. M. W. Nirenberg and H. J. Matthaei, *Proc. Natl. Acad. Sci. USA*, 1961, **47**, 1589.

Further Reading

- F. H. Portugal and J. S. Cohen, *A Century of DNA*, MIT Press, Cambridge, MA, 1977.
- J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, 5th edn., Freeman, New York, 2002 (Chapters 1 and 5).
- W. K. Purves, D. Sadava, G. H. Orians and H. C. Heller, *Life*, 6th edn., Sinauer, Sunderland, MA, 2001 (Chapters 9–12).