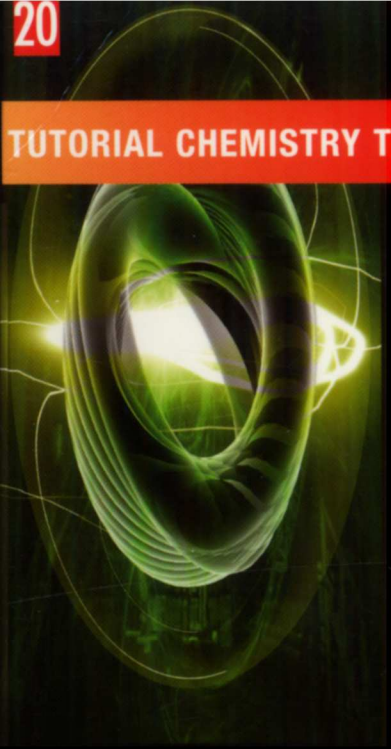
20

TUTORIAL CHEMISTRY TEXTS



RS•C

Nucleic Acids

by SHAWN DOONAN

TUTORIAL CHEMISTRY TEXTS

20
Nucleic Acids

SHAWN DOONAN
University of East London

RS•C
ROYAL SOCIETY OF CHEMISTRY

Cover images © Murray Robertson/visual elements 1998–99, taken from the 109 Visual Elements Periodic Table, available at www.chemsoc.org/viselements

ISBN 0-85404-481-7

A catalogue record for this book is available from the British Library

Published by The Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge CB4 0WF, UK
Registered Charity No. 207890
For further information see our web site at www.rsc.org

Typeset in Great Britain by Alden Bookset, Northampton
Printed and bound by Italy by Rotolito Lombarda

Preface

The nucleic acids are, in the main, informational macromolecules. DNA encodes the instructions that are passed on from parents to progeny, so that the offspring of two human beings is another human being, rather than any other life form. The related molecule RNA serves the same function in some viruses, such as the human immunodeficiency virus (HIV). The nucleic acids are, therefore, responsible for the continuation of all forms of life on this Earth, and as such are the subject of enormous interest to biologists.

Why should the student of chemistry be interested in them? The answer is that the biological roles of the nucleic acids can only be understood in terms of their chemical structures, and chemists have played a central part in establishing those structures and the understanding of how they function. The main objectives of this book are to explain what those structures are, how they were determined, and how function can be understood in terms of structure.

Although the nucleic acids are interesting substances in their own right, the real fascination of their study comes from the interplay between their chemistry and biology. Indeed, it is not really possible to disentangle the two, since some of the modern methods for the determination of the structures of nucleic acids make use of their biological properties. Nucleic acid chemistry cannot, therefore, be fully understood without some knowledge of the underlying biology. Consequently, some of the more important aspects of the fields of molecular biology and genetics are covered in Chapter 1, and in boxed material throughout the rest of the book for the benefit of those readers who lack the necessary background knowledge.

There is also a substantial amount of material in this book about the history of the subject. Study of the structure and function of nucleic acids has occupied the attention of a large number of enormously talented scientists over the last 130 years. This is reflected in the very large number of Nobel Prizes that have been awarded in the field. A proper understanding of the present state of our knowledge of the nucleic acids is, I believe, greatly illuminated by knowing something of how we arrived at it. It is, by any criterion, a remarkable story of scientific endeavour.

Shawn Doonan
London

TUTORIAL CHEMISTRY TEXTS

This series of books consists of short, single-topic or modular texts, concentrating on the fundamental areas of chemistry taught in undergraduate science courses. Each book provides a concise account of the basic principles underlying a given subject, embodying an independent-learning philosophy and including worked examples. The one topic, one book approach ensures that the series is adaptable to chemistry courses across a variety of institutions.

TITLES IN THE SERIES

Stereochemistry *D G Morris*
Reactions and Characterization of Solids *S E Dann*
Main Group Chemistry *W Henderson*
d- and f-Block Chemistry *C J Jones*
Structure and Bonding *J Barrett*
Functional Group Chemistry *J R Hanson*
Organotransition Metal Chemistry *A F Hill*
Heterocyclic Chemistry *M Sainsbury*
Atomic Structure and Periodicity *J Barrett*
Thermodynamics and Statistical Mechanics *J M Seddon and J D Gale*
Basic Atomic and Molecular Spectroscopy *J M Hollas*
Organic Synthetic Methods *J R Hanson*
Aromatic Chemistry *J D Hepworth, D R Waring and M J Waring*
Quantum Mechanics for Chemists *D O Hayward*
Peptides and Proteins *S Doonan*
Biophysical Chemistry *A Cooper*
Natural Products: The Secondary Metabolites *J R Hanson*
Maths for Chemists, Volume I, Numbers, Functions and Calculus *M Cockett and G Doggett*
Maths for Chemists, Volume II, Power Series, Complex Numbers and Linear Algebra *M Cockett and G Doggett*
Nucleic Acids *S Doonan*

TITLES IN THE SERIES

Inorganic Chemistry in Aqueous Solution *J Barrett*

FORTHCOMING TITLES

Mechanisms in Organic Reactions
Organic Spectroscopic Analysis

Further information about this series is available at www.rsc.org/tct

Order and enquiries should be sent to:
Sales and Customer Care, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge CB4 0WF, UK

Tel: +44 1223 432360; Fax: +44 1223 426017; Email: sales@rsc.org

Contents

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

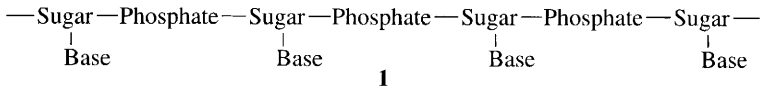
Molecular biology is a somewhat loose term. Logically it should mean the study of all life phenomena at the molecular level. It is more commonly taken to mean the study of the nucleic acids and protein synthesis.

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.



1

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 know 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.

1.3 DNA as the Carrier of Genetic Information

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**.

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 ^{35}S (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.

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**.

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!

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 F_1 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 F_1 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 F_2 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 F_1 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 F_1 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 F_2 plants? The F_1 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.

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.

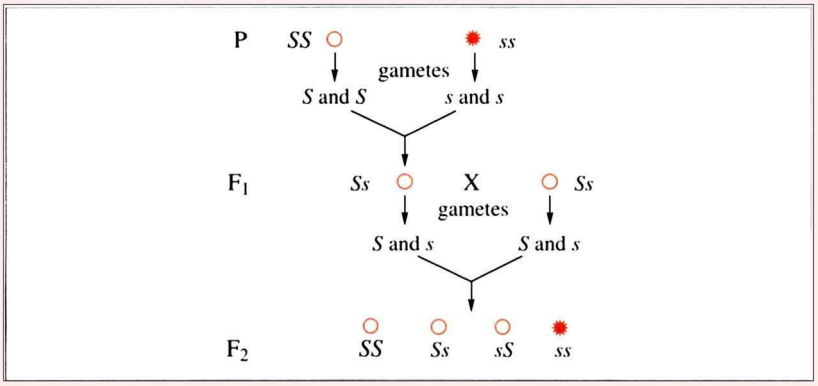


Figure 1.1 Mendelian inheritance of seed shape in peas. The seeds are either smooth (●) or wrinkled (○). Parental plants with smooth seeds have the genotype *SS* and produce gametes with the *S* allele. Parental plants with wrinkled seeds have the genotype *ss* and produce gametes with the *s* allele. 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

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*? Mandel 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

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.

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.

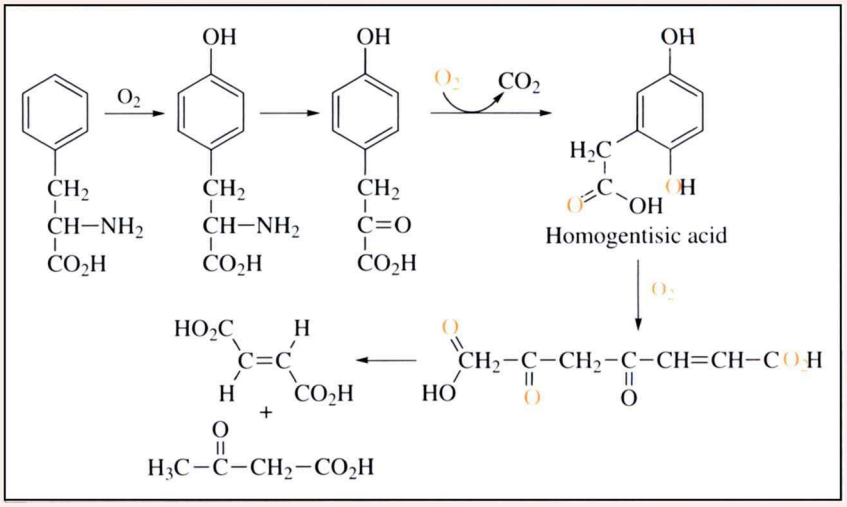
The other characteristic is that the genetic information must be **expressed**. That is, the information that the DNA contains must be interpreted in some way so that the cell in which it is contained does the right things. The information in DNA is largely a set of instructions for making **proteins**, 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 alkaptouria 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 heteozygotes, 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.

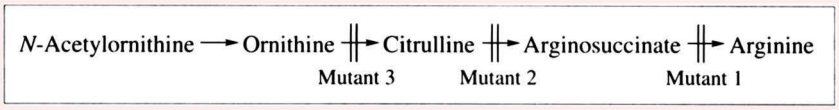
Important as they were, the significance of Garrod's results was not fully realized until considerably later, when further evidence had accumulated that genes code for proteins. Much of that evidence was obtained by George Beadle and Edward Tatum working in the USA. Their most significant studies were 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

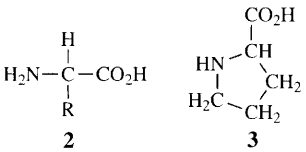
All of the fungal strains grew on a medium containing arginine, but only the wild-type strain grew on a minimal medium not containing arginine. Mutant 1 did not grow when the medium was supplemented with any of the three intermediates arginosuccinate, citrulline or ornithine, and so lacked the final 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.



2 **3**

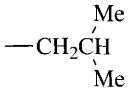
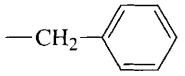
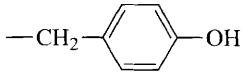
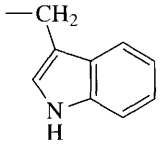
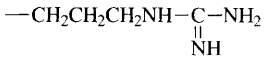
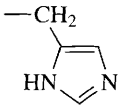
Proteins are polymers made from 19 α-amino acids (**2**) and the imino acid proline (**3**). What distinguishes one amino acid from another is the nature of the side chain (the group R in structure **2**). Table 1.2 gives a list of the 19 amino acids that are specified by the genetic code along with 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	—CH(Me)(Me)	Val	V

(continued)

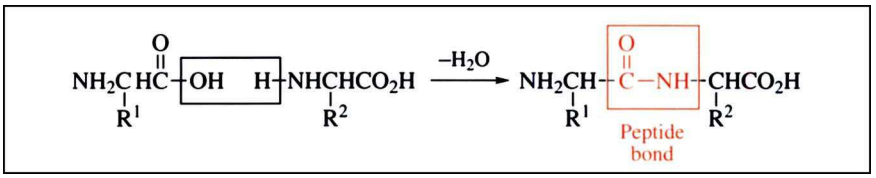
Table 1.2 *continued*

Type of side chain	*Name*	*Structure of side chain*	*Abbreviated name*	*One-letter abbreviation*
	Leucine		Leu	L
	Isoleucine		Ile	I
Aromatic	Phenyl-alanine		Phe	F
	Tyrosine		Tyr	Y
	Tryptophan		Trp	W
Alcohols	Serine	$-CH_2OH$	Ser	S
	Threonine		Thr	T
Thiol	Cysteine	$-CH_2SH$	Cys	C
Sulfide	Methionine	$-CH_2CH_2S-Me$	Met	M
Acids	Aspartic acid	$-CH_2CO_2H$	Asp	D
	Glutamic acid	$-CH_2CH_2CO_2H$	Glu	E
Amides	Asparagine	$-CH_2CONH_2$	Asn	N
	Glutamine	$-CH_2CH_2CONH_2$	Gln	Q
Bases	Lysine	$-CH_2CH_2CH_2CH_2NH_2$	Lys	K
	Arginine		Arg	R
	Histidine		His	H

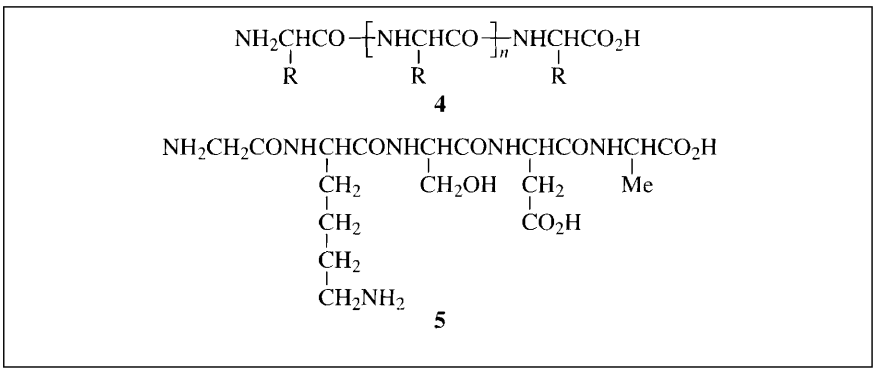
[a]The imino acid proline (structure **3**) is also a constituent of proteins. Its abbreviated name and one-letter abbreviation are Pro and P respectively.

The amino acids in proteins are linked together by an amide linkage (Scheme 1.3) which is given the special name of the **peptide bond**. Hence a

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



4

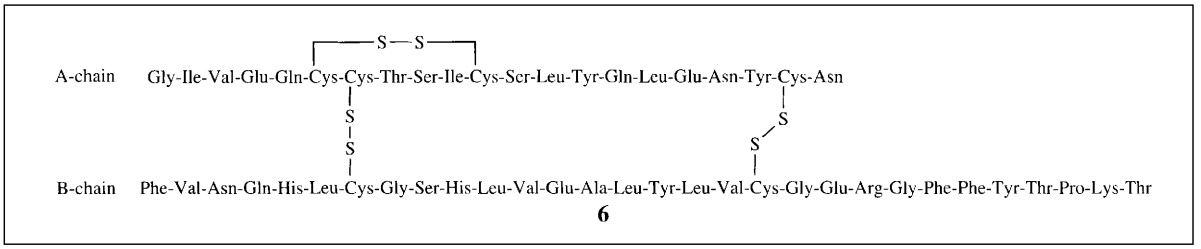
5

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.

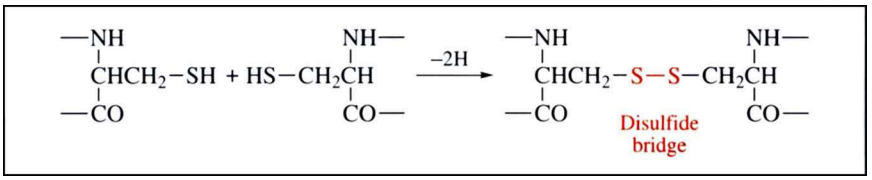
Note that there are 120 possible peptides with the same amino acid composition as **5**. There are 5 ways of choosing the N-terminal residue. For each of these there are 4 ways of choosing the second residue, on so on. So the number of unique sequences is $5\times4\times3\times2\times1=120$.

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.



6



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 UUCCACAGUGUUAUAUCCGGCUGGG. 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.

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**.

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.

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 proteins 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 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:

$$\text{DNA} \rightarrow \text{RNA} \rightarrow \text{Protein}$$

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:

$$\text{DNA} \rightleftharpoons \text{RNA} \rightarrow \text{Protein}$$

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:

GAGCUCGUAAUUUCCAUAUCUACUCAUGAAAAAAUUAACGGG

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).

2
The Covalent Structures of Nucleic Acids

Aims

By the end of this chapter you should understand:

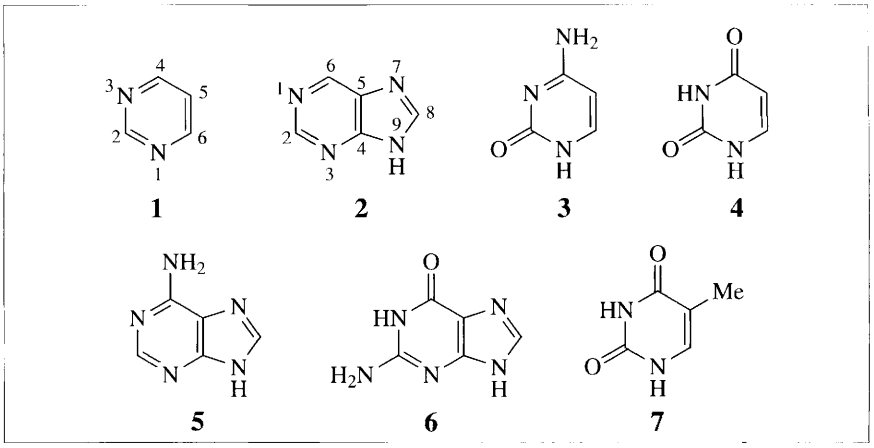
- The structures of the bases and monosaccharides from which RNA and DNA are made
- How the bases and monosaccharides are linked to form nucleosides
- That nucleotides are monophosphate esters of nucleosides
- That nucleotides are the basic structural units of the nucleic acids
- How nucleotides are linked together to form the polymeric nucleic acids
- How the structures of nucleic acids are written in shorthand form
- The sizes of representative examples of nucleic acids

2.1 The Building Bricks

As outlined in Section 1.2, both DNA and RNA are polymeric molecules with backbones consisting of alternating monosaccharide and phosphate units. Each of the monosaccharide components carries one of four possible heterocyclic bases. In both cases, two of these bases are derivatives of pyrimidine (**1**), and the other two are derivatives of purine (**2**). The pyrimidines in RNA are **cytosine** (4-amino-2-hydroxypyrimidine, **3**) and **uracil** (2,4-dihydroxypyrimidine, **4**). The purines are **adenine** (6-aminopurine, **5**) and **guanine** (2-amino-6-hydroxypurine, **6**). The same two purines occur in DNA, but one of the pyrimidines is different. Instead of uracil the fourth base in DNA is **thymine** (5-methyluracil, **7**).

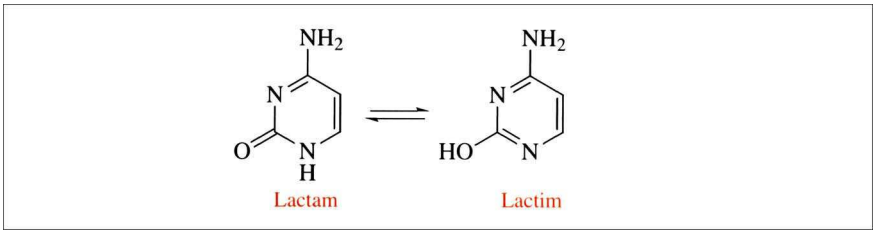
The DNA molecules in certain bacteriophages have unusual base compositions. For example, in phage PBS, which infects the bacterium *Bacillus subtilis*, uracil replaces thymine. Some coliphages (phages that infect *Escherichia coli*) contain 5-(hydroxymethyl)cytosine instead of cytosine.

Note the different convention used for numbering the ring in pyrimidine, compared with that used for the same structure as part of the purine molecule.



Tautomers are forms of a molecule that differ in the arrangements of hydrogen atoms and of double bonds, and are interconvertible. For example, 2-oxopropanoic acid (CH_3COCO_2H) and 2-hydroxyprop-2-enoic acid ($CH_2=C(OH)CO_2H$) are tautomers. But-1-ene ($CH_2=CHCH_2CH_3$) and but-2-ene ($CH_3CH=CHCH_3$) are not: they are not interconvertible.

It might not be immediately obvious why, for example, cytosine is 4-amino-2-hydroxypyrimidine. The point here is that the hydroxyl-containing purines and pyrimidines show **tautomerism**. So cytosine can exist as either of the two forms shown in Scheme 2.1, and the naming is that of the lactim form. As we shall see latter, it is the lactams that are found in the nucleic acids.

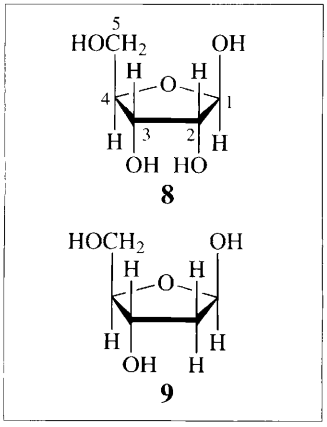


Scheme 2.1

Worked Problem 2.1

Q Name hypoxanthine (see Scheme 2.3) as a derivative of purine.

A Hypoxanthine has a tautomeric form with a hydroxyl group on C_6. It can be named, therefore, as 6-hydroxypurine.



The monosaccharide in RNA is **β-D-ribofuranose**. This is a five-carbon sugar (a **pentose**) and its structure is shown in **8**, with the numbering of the carbon atoms indicated. The structure in **8** is only one of the forms that ribose can adopt (see Box 2.1), but it is the one invariably found in RNA, and in this context it is common practice to refer to it simply as **ribose**. Note that, in most of what follows, the hydrogen atoms on carbons 1–4 will be omitted for clarity.

The monosaccharide in DNA is the closely related **β-2-deoxy-D-ribofuranose**, usually known simply as **deoxyribose**, in which the

hydroxyl group on C-2 of β-D-ribofuranose is replaced by a hydrogen atom (**9**). Although this is a seemingly small difference between the two, its effects on the properties of the nucleic acids are profound (see, for example, Box 2.4). The base and sugar components of the nucleic acids are summarized in Table 2.1.

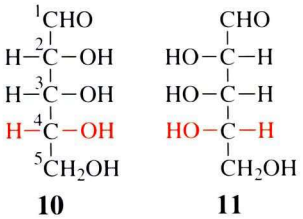
Table 2.1 Components of the nucleic acids. The differences between DNA and RNA are shown in red

	DNA	*RNA*
Bases	Adenine	Adenine
	Guanine	Guanine
	Cytosine	Cytosine
	Thymine	**Uracil**
Monosaccharide	**β-2-Deoxy-D-ribofuranose**	**β-D-Ribofuranose**

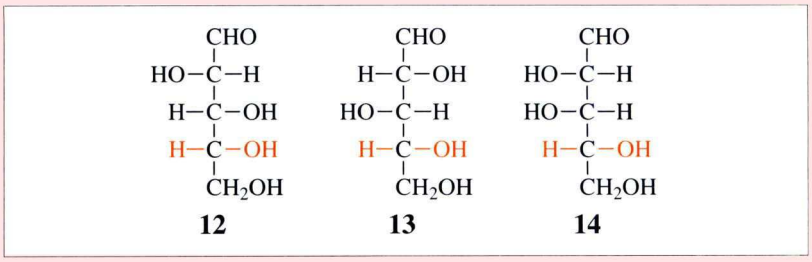
Box 2.1 The Stereochemistry and Ring Forms of Pentoses

Ribose is an **aldopentose**, and one way to represent its structure is as the straight-chain **Fischer projection** shown in **10**. In a Fischer projection, the chain of carbon atoms is drawn vertically with that of highest oxidation state at the top. For each stereogenic centre (that is, a carbon atom with four different groups attached), the vertical bonds lie behind the plane of the paper and the horizontal bonds come forward out of the plane. The molecule in **10** has the D-configuration because of the orientation of the groups shown in red on C-4; this is the naturally occurring isomer. Its mirror image (**11**) is L-ribose. It is the orientation of the groups on C-2 and C-3 which makes the molecule ribose, rather than any of the other three possible diastereoisomers (see Morris[1] for a discussion of the stereochemistry of molecules of this sort). The diastereoisomers of ribose are arabinose (**12**), xylose (**13**) and lyxose (**14**), all shown here in the D-configuration.

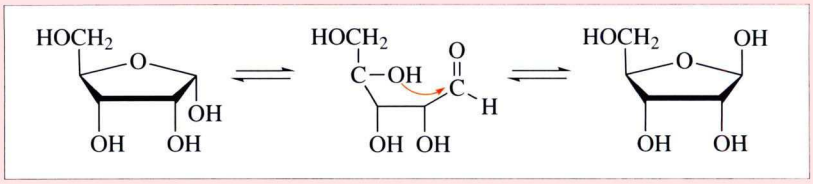
There are other types of monosaccharides in which the carbonyl function is on C-2 rather than C-1. These are the **ketoses**. Ketoses with five carbons are called **ketopentoses**.



10 **11**



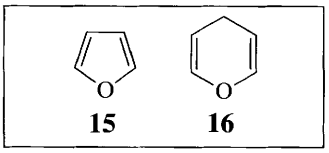
Monosaccharides such as ribose form cyclic structures by a carbonyl addition reaction involving internal attack of one of the chain hydroxyl groups on the aldehyde function on C-1. This is shown in Scheme 2.2, with attack occurring by the hydroxyl group on C-4. There are two possible products, depending on the side of the planar carbonyl group at which attack occurs. With the carbonyl group oriented as shown in Scheme 2.2, then the product is that on right-hand side of the scheme; this compound is β-D-ribofuranose. If, on the other hand, the carbonyl group had rotated through 180° before attack occurred, with the carbonyl oxygen pointing downwards, then the product would be as on the left of Scheme 2.2; this is **α-D-ribofuranose**. These two forms of the monosaccharide are called **anomers**, and C-1 is referred to as the **anomeric carbon atom**. Note that the ring forms in Scheme 2.2 are shown as so-called **Haworth projections**, in which the ring is drawn flat with bonds to it vertical, and do not represent the true geometry of the molecules.



Scheme 2.2

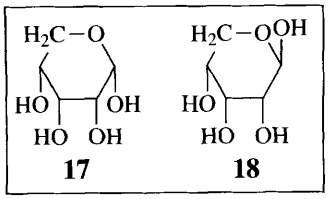
The cyclization reactions shown in Scheme 2.2 are reversible and so, in solution, ribose consists of a mixture of all three forms. The ring forms predominate in the equilibrium mixture. If, however, the hydroxyl group on C-1 is derivatized, as it is in nucleosides (see Section 2.2), then the ribose in locked into the ring form. The anomer found in these compounds is β-D-ribofuranose.

Why are the ring forms referred to as furanoses? This is by analogy with the five-membered ring system in the heterocyclic compound furan (**15**), and is a terminology used to specify the ring size. This is necessary because in principle ribose could cyclize by



attack of the hydroxyl group on C-5 onto the carbonyl group to produce a six-membered ring. The naming of such compounds is now by analogy with the six-membered ring system of pyran (**16**).

Again, two anomers would be produced, and these would be called α-D-ribopyranose (**17**) and β-D-ribopyranose (**18**). In the case of the pentoses, however, the five-membered ring system is much more stable than is the six-membered ring, and very little of the pyranose forms are found in a solution of ribose.

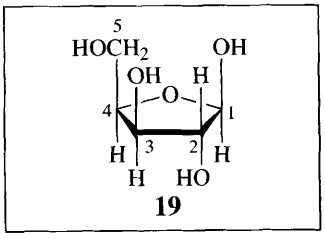


17 **18**

Worked Problem 2.2

Q Draw the structure of β-D-xylofuranose as a Howarth projection and number the carbon atoms.

A The only difference between ribose (**10**) and xylose (**13**) is in the orientation of the hydroxyl group on C-3. Hence the Howarth projection of β-D-xylofuranose is as shown in **19** with the OH on C-3 up instead of down. An instructive way to do this problem is to build a model of the straight-chain form of **13** and then convert it into a ring structure as shown in Scheme 2.2, ensuring that the carbonyl oxygen is pointing upwards when the ring is formed.

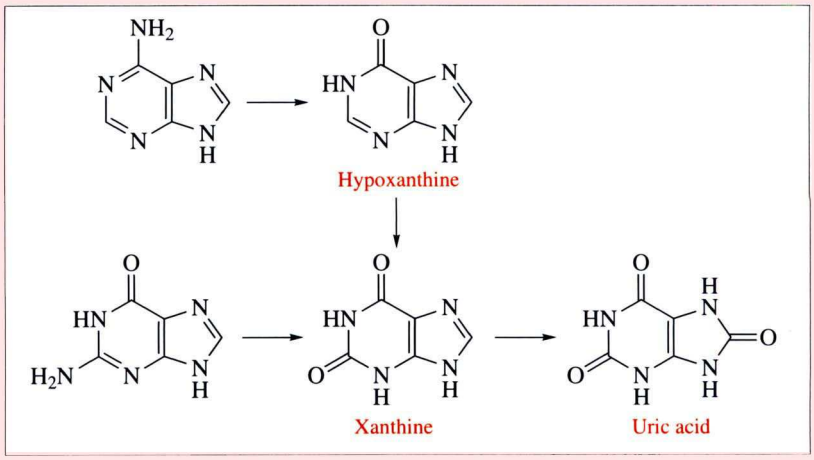


19

Box 2.2 The Discovery of the Components of the Nucleic Acids

Soon after his initial isolation of nuclein, Miescher showed that the substance he had isolated contained phosphate (see Box 1.1). He also obtained evidence that nuclein was a source of the so-called xanthine bases (**xanthine** and **hypoxanthine**), but he did not appreciate the full significance of the latter findings. This was done by the towering figure of nucleic acid chemistry in the 19th century, Albrecht Kossel. Like Miescher, Kossel also worked in Hoppe-Seyler's laboratory, where Miescher's original work had been followed up by isolation of nuclein from a variety of other sources, including yeast. In his earliest experiments, Kossel confirmed that hydrolysis of yeast nuclein yielded not only phosphate, but also xanthine and hypoxanthine.[2] Over the next few years he showed that these two substances were not themselves

components of nuclein, but rather they arose as degradation products of adenine and guanine, and that it was these latter substances that were the true components. The relationship between these compounds, and their conversion to uric acid, is shown in Scheme 2.3.



Scheme 2.3

Subsequently, Kossel, with a co-worker, A. Neumann, turned his attention to the nuclein from a different source, namely thymus tissue from calf. From this they isolated a previously unknown substance which they called thymine, in recognition of the tissue from which it was obtained. They then went on to discover yet a further new compound, cytosine, from the same source,[3] and showed that these substances were not breakdown products of adenine or guanine, but were themselves components of nuclein.

It is significant that Kossel and Neumann had used thymus tissue as their source of nuclein, because this tissue is rich in DNA, but is not a good source of RNA (the distinction between the two was not, of course, known at that time). A different result was obtained as a result of work carried out in 1900, when yeast was used as a source of nuclein. In this case, as well as adenine, guanine and cytosine, the nuclein was found to yield not thymine, but another new compound which was called uracil. The reason for this result is that yeast is a good source of RNA which, as we now know, contains uracil instead of thymine. Thus, by the beginning of the 20th century, Kossel and his co-workers had identified the complete set of bases which we now know to be constituents of the nucleic acids. Kossel was awarded the Nobel Prize in Physiology or

Medicine in 1910 "in recognition of the contributions to our knowledge of cell chemistry made through his work on proteins, including the nucleic substances".

Although Kossel first recognized the presence of carbohydrate in nucleic acids in 1893, the credit for determining the structures of the monosaccharides belongs to a Russian scientist, Phoebus Levene. The identification of D-ribose as the monosaccharide in nucleic acid from yeast (RNA) was achieved in 1909. Identification of the residue in thymus nucleic acid (DNA) took much longer. The problem was that several investigators had found that degradation of this form of nucleic acid produced 4-oxopentanoic acid ($CH_3COCH_2CH_2CO_2H$), which was thought to originate from a hexose. The intact monosaccharide from DNA proved to be very difficult to isolate because it was unstable under the relatively harsh methods originally used to degrade nucleic acids. It was not until Levene developed methods of hydrolysis of the nucleic acid using enzymes to produce nucleotides (see Section 2.2), followed by very mild acid hydrolysis to liberate the sugar component, that he managed to obtain the sugar in crystalline form. He then showed that it was the previously unknown 2-deoxy-D-ribose.[4]

By 1930, then, the most significant difference between the DNA and RNA, that is, the difference in the nature of the sugar residue, was understood. Thymus nucleic acid became known as DNA, and yeast nucleic acid as RNA. It was also established around this time that both forms of nucleic acid occur in all living organisms. The original designation of DNA as being of animal origin and of RNA as being of plant origin was simply a reflection of the relative abundance of DNA in thymus and of RNA in yeast.

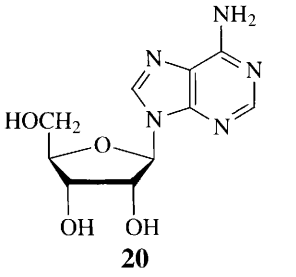
The wording of the citation emphasizes the fact that, in 1910, it was still believed that nuclein consisted, at least in part, of protein!

In the 1800s, living organisms were classified as being either plants or animals, and yeast was included among the plants. Modern classification puts the yeasts in the kingdom of the fungi.

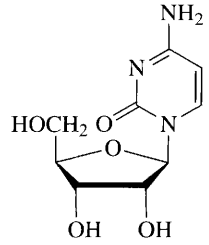
2.2 Nucleosides and Nucleotides

The individual components of the nucleic acids had been discovered by complete degradation, followed by analysis of the molecules liberated. The next problem was how these components are linked together in the intact molecule. Solution to this problem came from study of the products of partial hydrolysis of the nucleic acids. These are of two types, **nucleosides** and **nucleotides**.

Ribonucleosides are molecules containing a base linked to C-1 of ribose. Typical examples are **adenosine** (9-β-D-ribofuranosyladenine, **20**) and **cytidine** (1-β-D-ribofuranosylcytosine, **21**). Note that the linkage is to N-1 of the pyrimidine ring and to N-9 of the purine ring.



20



21

There is an entirely analogous set of molecules called **deoxyribonucleosides**, where the monosaccharide is deoxyribose. The molecule deoxyadenosine, for example, has the same structure as **20**, but with deoxyribose replacing ribose. Table 2.2 lists the names of the nucleosides formed by the bases commonly found in RNA and DNA.

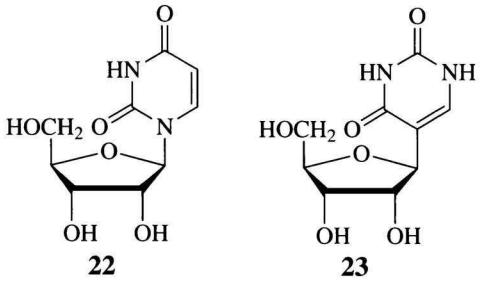
Table 2.2 Nucleosides and nucleotides derived from RNA and DNA

Sugar	*Base*	*Name of nucleoside*	*Name of nucleotide*
Ribose	Adenine	Adenosine	Adenylic acid
	Guanine	Guanosine	Guanylic acid
	Cytosine	Cytidine	Cytidylic acid
	Uracil	Uridine	Uridylic acid
Deoxyribose	Adenine	Deoxyadenosine	Deoxyadenylic acid
	Guanine	Deoxyguanosine	Deoxyguanylic acid
	Cytosine	Deoxycytidine	Deoxycytidylic acid
	Thymine	Deoxythymidine	Deoxythymidylic acid

Worked Problem 2.3

Q Transfer RNA molecules contain small amounts of unusual nucleosides. One of these is 5-β-D-ribofuranosyluracil (also known as pseudouridine, or ψ). Give the structures of both uridine and pseudouridine.

A In uridine, as with all the commonly occurring nucleosides, the linkage to ribose is *via* N-1 (**22**). In pseudouridine it is *via* C-5 so the structure is as shown in **23**.

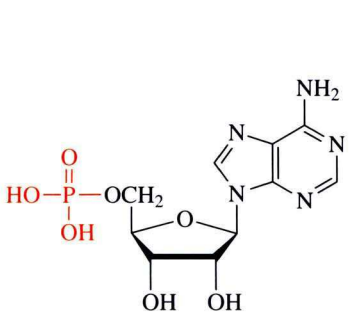


22 **23**

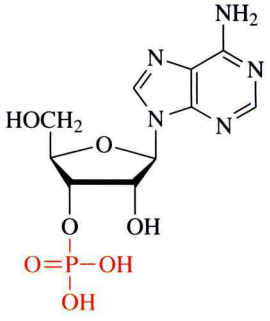
Carboxylic acid esters are familiar in organic chemistry. For example, methanol condenses with ethanoic acid to yield methyl ethanoate (**24**). Phosphoric acid forms esters in an analogous way, but in this case there are three

Nucleotides are **phosphate monoesters** . Taking adenosine as an example, there are three possible hydroxyl groups that could be esterified, and all three compounds are known. Their structures are shown in **28**–**30**. Compound **28** is adenosine 5′-monophosphate, **29** is adenosine 3′-monophosphate, and **30** is adenosine 2′-monophosphate.

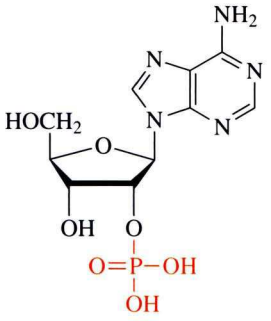
They are collectively known as **adenylic acids**. The names of the nucleotides derived from each nucleoside are given in Table 2.2.



28

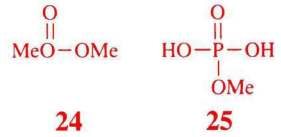


29

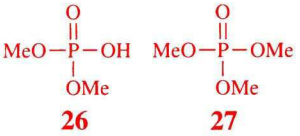


30

acidic groups. So, for example, when phosphoric acid reacts with methanol it can form monomethyl phosphate (**25**), dimethyl phosphate (**26**) and trimethyl phosphate (**27**). Monoesters and diesters of phosphoric acid are of very wide occurrence in biology.



24 **25**



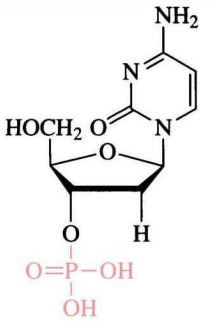
26 **27**

Note that in nucleosides and nucleotides, a prime or dash (′) is added to the numbers describing the positions in the ribose and deoxyribose rings to distinguish the atoms referred to from those in the bases. So the name of **28** is read as "adenosine five prime monophosphate" or "adenosine five dash monophosphate". Note also that in structures **28–30** the phosphate group is shown un-ionized. In solution, the phosphate group will carry zero, one or two negative charges, depending on the acidity.

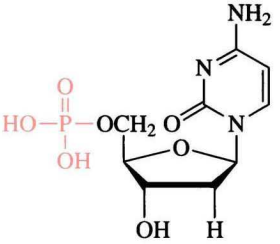
Worked Problem 2.4

Q Give the names and draw the structures of the two possible nucleotides that can be formed from deoxycytidine.

A The nucleotides are deoxycytidine 3′-phosphate (**31**) and deoxycytidine 5′-phosphate (**32**).



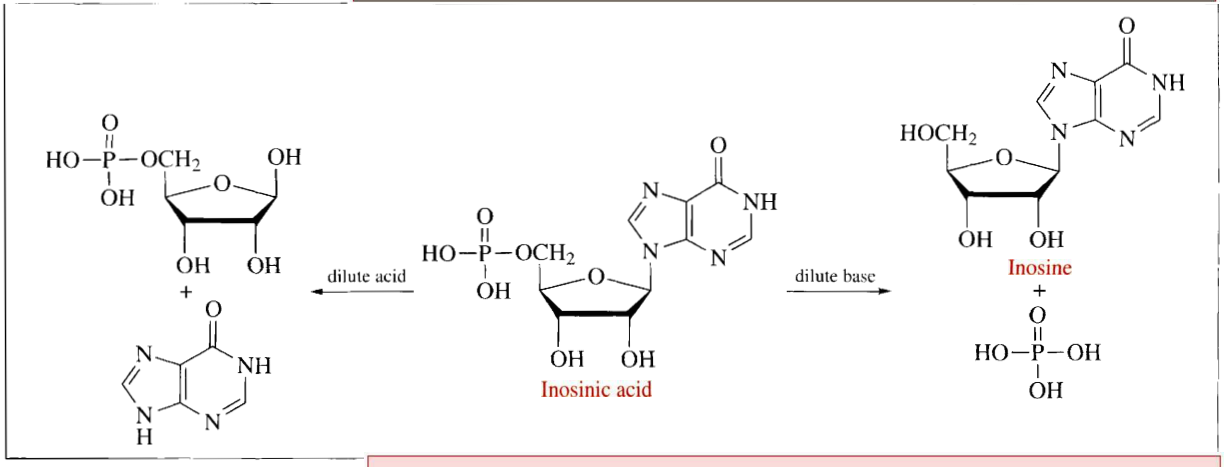
31



32

Box 2.3 Discovery of the Structures of Nucleosides and Nucleotides

The discovery of the order of linkage of the base, sugar and phosphate residues in a nucleotide was also made by Levene. The compound **inosinic acid** had been isolated from a meat extract by von Liebig in 1847. Levene and W. A. Jacobs[5] showed that treatment of inosinic acid with dilute base yielded phosphoric acid and inosine. On the other hand, treatment with dilute acid yielded ribose phosphate and the free base hypoxanthine. These results showed that the linkage order was phosphate–sugar–base (see Scheme 2.4). Levene also coined the terms nucleoside to describe substances such as inosine, and nucleotide to describe the nucleoside phosphates.



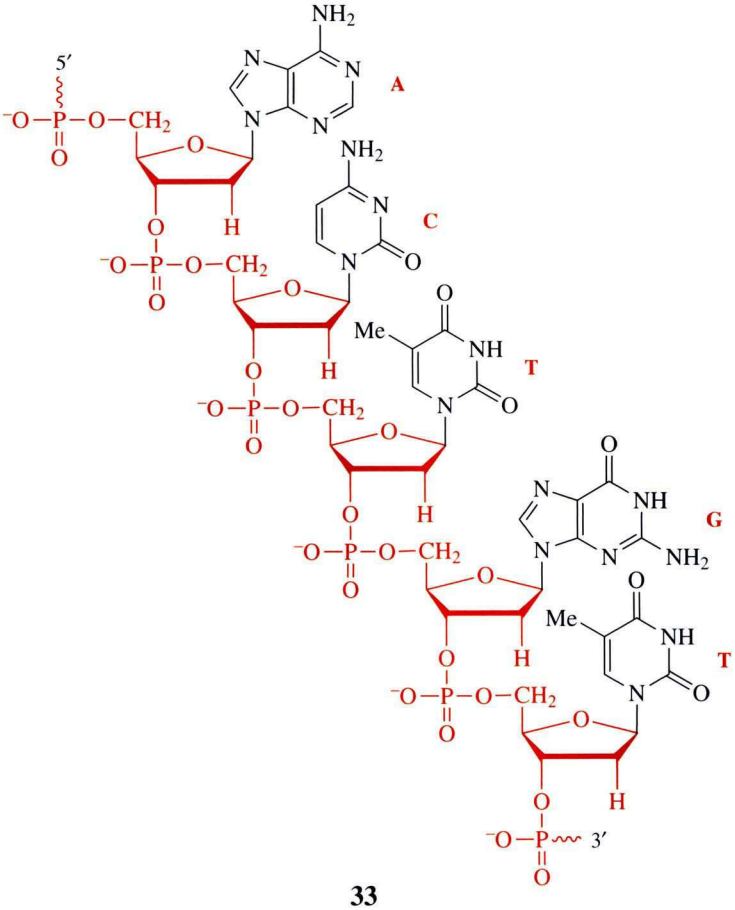
Scheme 2.4

In subsequent experiments, Levene and his co-workers showed that controlled enzymatic hydrolysis of thymus nucleic acid (*i.e.* DNA) produced mixtures of deoxynucleotides and deoxynucleosides. This work established that the nucleotide was the fundamental structural unit of nucleic acids.

2.3 The Inter-nucleotide Linkage

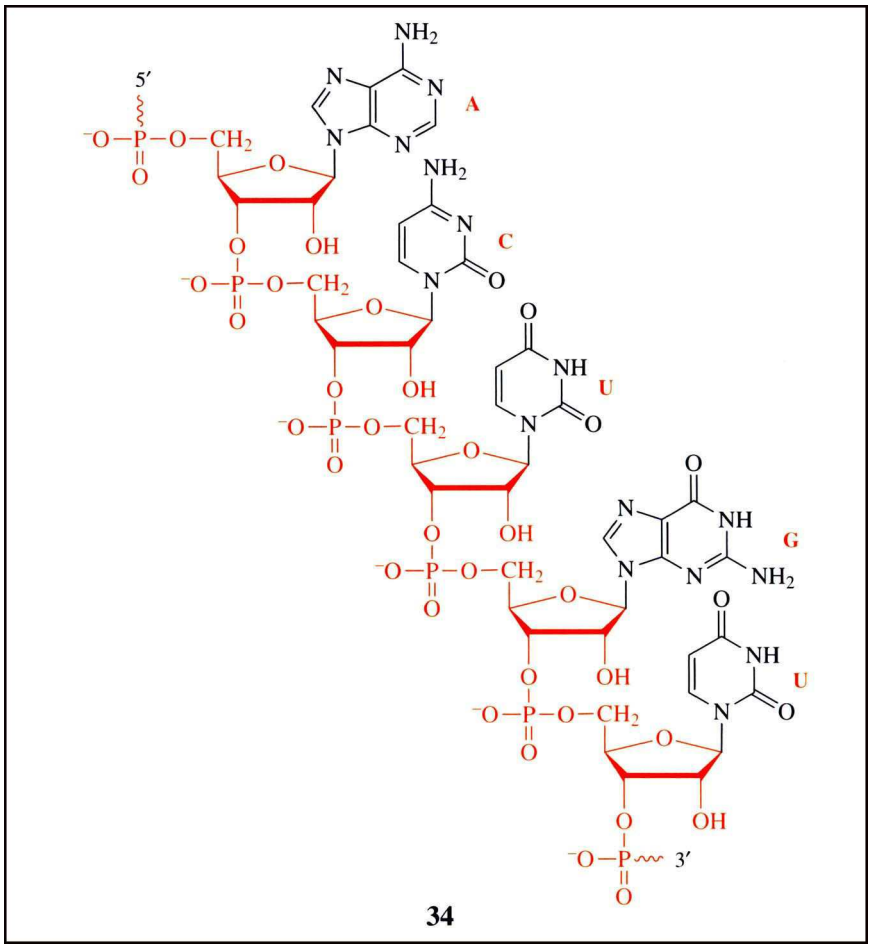
Both DNA and RNA are polymers of nucleotides in which the monomeric units are linked by phosphodiester bonds between the

5′-hydroxyl group of one sugar and the 3′-hydroxyl group of the next sugar in the chain. This is shown in **33** for a section of a DNA molecule and in **34** for a section of an RNA molecule. In both cases the sugar–phosphate backbone is shown in red, and the attached bases in black. The "squiggly bonds" at the ends indicate that the chains carry on in both directions.



33

It is important to note that the chain is directional; that is, the two ends are not the same. If the phosphodiester bonds to the next residues in the chain in both directions were broken, then the end at the top in **33** would have a free OH on the 5′- position of the deoxyribose, whereas the end at the bottom would have the OH on the 3′- position. Reading from top to bottom, the structure is said to be written **5′→3′**. This directionality is



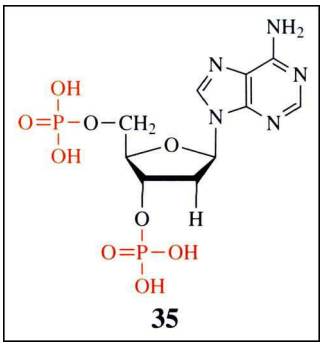
important when we use shorthand representations of the structure of a DNA chain.

The structure of RNA is essentially the same as that of DNA, except that ribose replaces deoxyribose. Here it is important to note that the 2′-OH is not involved in the linkage between the sugar units. Note that in both structures, the phosphate is shown as being negatively charged. This will always be the case except in solutions of high acidity.

Box 2.4 Establishment of the Inter-nucleotide Linkage

As well as discovering that nucleotides were the structural units of nucleic acids, Levene may also be credited with the first correct

assignment of the inter-nucleotide linkage. In 1912, he and Jacobs isolated products from the hydrolysis of thymus nucleic acid (DNA) which they identified as nucleoside diphosphates. Subsequently, the sugar unit in DNA was shown to be deoxyribose (see Box 2.2). Given that the hydroxyl group on C-1 of deoxyribose is replaced by the linkage to the base in a deoxynucleoside, it followed that the other two phosphates in deoxynucleoside diphosphates must be on C-3′ and C-5′.[6] That is, the structure is as shown in **35** taking the adenosine derivative as an example. This suggested that the nucleoside units were linked by phosphates from C-3′ on one deoxyribose to C-5′ on the next, as shown in **33**.



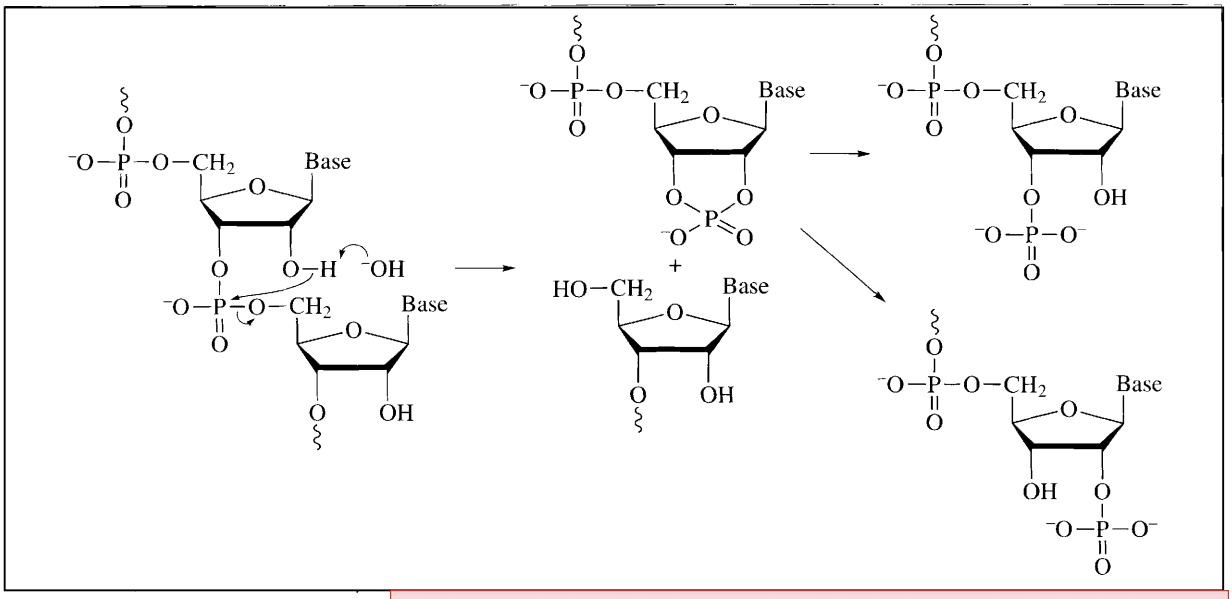
35

Levene's conclusions were questioned at the time, and the final proof of the structure of DNA and of RNA is attributed to Alexander Todd (subsequently Lord Todd of Trumpington). Todd worked in Cambridge, where he and his group did very important work on the synthesis of phosphate esters of biological importance. His main focus was on coenzymes and cofactors (see Box 2.5) but he also made very important contributions to knowledge of the structures of nucleic acids. Firstly, the synthetic methods that he and his team developed allowed him to make defined nucleoside mono- and diphosphate esters, and thus to confirm the structures of the degradation products of nucleic acids. Finally, in 1955, Todd and Michelson[7] succeeded in synthesizing a dinucleotide with a 3′–5′ linkage and demonstrated unambiguously that this synthetic product was identical to a dinucleotide isolated from DNA. Thus the covalent structure of DNA was finally established beyond question.

There was a complication in establishing the structure of RNA. It was reasonable to suppose that the inter-nucleotide linkage would involve the 5′-position of the ribose, but there were then two possible linkage points to the next ribose in the chain. This could be either to the 3′- or the 2′-position. Confusion arose because hydrolysis of RNA by base yields both 3′- and 2′-nucleoside monophosphates. Hence the possibility arose that both types of linkage, 5′→3′ and 5′→2′, occurred in RNA.

The situation was clarified in a classic paper by Brown and Todd.[8] What happens is that base-catalysed hydrolysis of RNA proceeds *via* a cyclic 2′,3′-phosphate intermediate. This then opens to yield either a nucleoside 2′-phosphate or a nucleoside 3′-phophate. The process is shown in Scheme 2.5.

Lord Todd was awarded the Nobel Prize in Chemistry in 1957 for "his work on nucleotides and nucleotide co-enzymes".



Scheme 2.5

Hydrolysis of RNA by the enzyme **ribonuclease** (RNase) also yields a mixture of 2′- and 3′-phosphates. The mechanism of action of the enzyme involves a process similar to that in Scheme 2.5.

In should be noted that a pathway to hydrolysis like that shown in Scheme 2.5 is not available for DNA because in that case there is no 2′-hydroxyl group. Indeed, whereas RNA is relatively labile in basic solution, DNA is stable under those conditions. This provides a convenient way of removing trace RNA contaminants from a DNA sample.

2.4 Shorthand Notations

It is clearly cumbersome to draw the complete covalent structures of even small sections of nucleic acids such as those shown in **33** and **34**, and rapidly becomes impossible as the sizes of the molecules increase. Shorthand notations are therefore used.

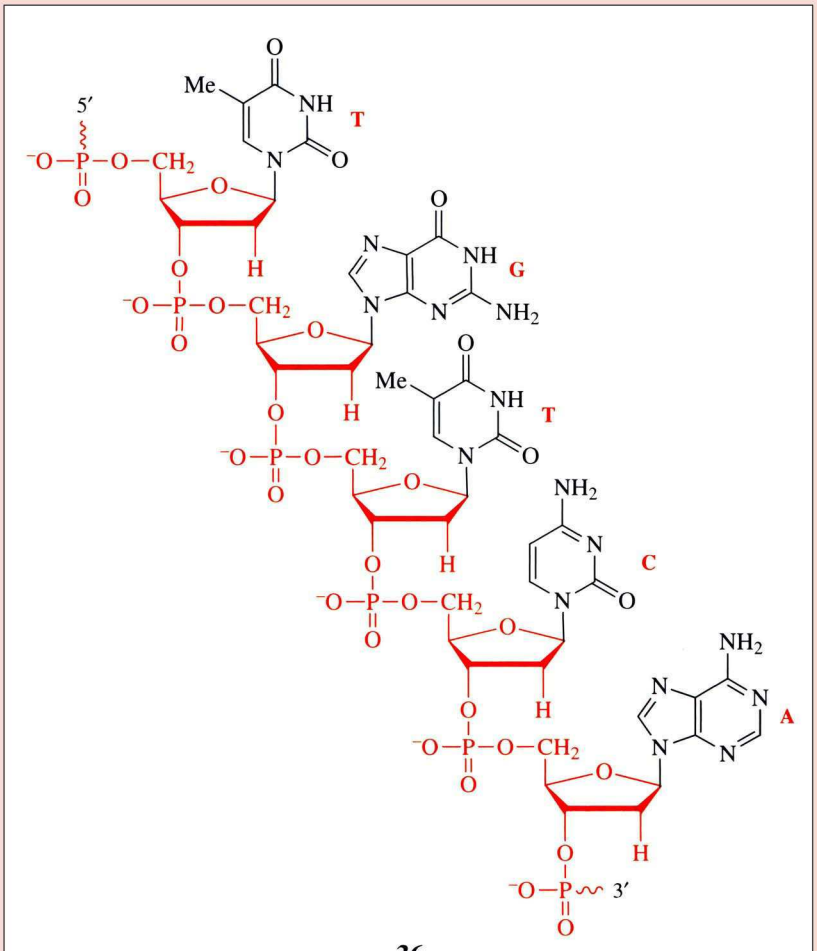
In **33**, the deoxyribonucleosides in the chain, reading from the 5′-end, are deoxyadenosine, deoxycytidine, deoxythymidine, deoxyguanosine and deoxythymidine. If we represent these by the letters A, C, T, G and T, the structure in **33** can be written as pApCpTpGpTp. Note that the convention is to write the 5′-end of the chain on the left. Although this representation is very compact, it is not really necessary to put the "p" in between each deoxyribonucleoside since the structure of DNA requires it to be there. So we can simplify further to pACTGTp. To the nucleic acid chemist, this means just the same as the structure in **33**. As well as being more compact, representing the structures of DNA molecules as simply a string of four letters is ideal for storing them electronically, as we will see in Section 5.9.

The RNA molecule in **34** can be dealt with in just the same way. That is, its structure can be written as pApCpUpGpUp or, more simply, as pACUGUp. How do we know that this is a section of an RNA molecule rather than DNA? The answer is, of course, that RNA contains uridine, whereas DNA does not (except in very rare circumstances; see margin note on page 25).

Worked Problem 2.5

Q The partial DNA sequence in **33** is pACTGTp. Give the structure of pTGTCAp.

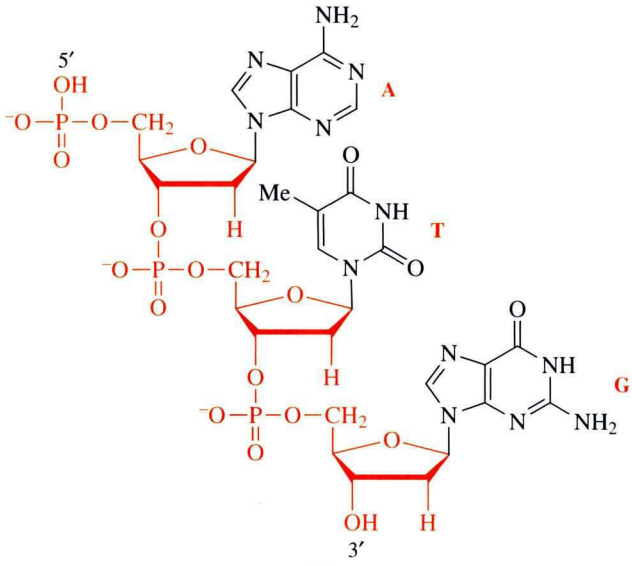
A Recall that, in the shorthand notation for nucleic acid structures, the chain is written 5′ → 3′, so the structure is as shown in **36**.



36

2.5 Oligonucleotides

The structures **33** and **34** represent short sections of DNA and RNA molecules. Structure **37** is an intact molecule consisting of three nucleosides linked together, and is referred to either as a **trinucleotide** or a **trimer**. Its structure can be represented as pATG following the convention that A is the 5′-residue and G is the 3′-residue; the "p" indicates that there is a phosphate on the 5′-OH of A. The string of letters ATG is referred to as the **base sequence** of the molecule. This is not strictly correct, because the letters really represent nucleosides rather than bases, but the terminology is universally used.



37

It is worth emphasizing that there *is* a bit of sloppiness in the terminology. Structure **37** is really a trideoxyribonucleotide. It is common practice, however, to use the short form trinucleotide provided that the context is clear; that is, when we know that we are talking about DNA rather than RNA. Similarly, the abbreviated structure should properly be written dAdTdG, but the "d"s are implied if the context is DNA. In the chapters that follow, the short terms nucleoside and nucleotide will be used unless it is necessary, for some reason, to be more specific. Similarly, the single letter abbreviations will be used for both nucleosides and bases, and only where it is necessary to make a particular point will the "d" be included to indicate a deoxyribonucleoside.

In general, molecules like **37** are referred to as **oligonucleotides** (this just means a molecule containing a few nucleotides – from the Greek *oligoi* meaning few). For example, AGTAGTCCATAG is an oligonucleotide containing 12 residues. It might also be termed, rather inelegantly, a **12mer**. We will see later on (Section 5.10) that synthetic oligonucleotides are very important in nucleic acid chemistry.

2.6 Sizes of Nucleic Acids

Nucleic acids are very large molecules. Even the smallest type of RNA molecule with which we will be concerned, transfer RNA, contains nearly 100 nucleotide residues. The 16S ribosomal RNA contains about 1150 residues. The size of an mRNA molecule depends on the protein for which it codes, and can be up to several thousand residues long.

DNA molecules are generally much larger. The first DNA molecule to have its base sequence determined was the genome of the phage ϕX174 (see Box 5.7). This is a very small genome and contains only 5375 nucleotides. Moving up the scale, mammalian mitochondrial DNA consists of a single molecule of about 15,000 nucleotides (also described further in Box 5.7). Next up in size come bacterial genomes. For example, that of the bacterium *Escherichia coli*, which colonizes the human gut, contains about 4,500,000 nucleotides in a single chain. In higher organisms the DNA is organized into a number of chromosomes (see Section 3.4), each of which contains a single molecule of DNA. For example, the total genome of the mouse contains about 2,500,000,000 nucleotides divided up between 20 chromosomes. So on average, each chromosome contains 125,000,000 residues. In fact, the sizes of the DNA molecules in the mouse genome vary between 58,000,000 and 192,000,000 residues. How the structures of these enormous molecules are determined is described in Section 5.8.

Chemists usually refer to the sizes of molecules in terms of their relative molecular masses (M_r), but this is not very profitable in the case of molecules like DNA and RNA. It is, however, easy to calculate the approximate M_r values for these molecules. In DNA, the average residue M_r is about 300. So, for example, the M_r of ϕX174 DNA is about 1,600,000.

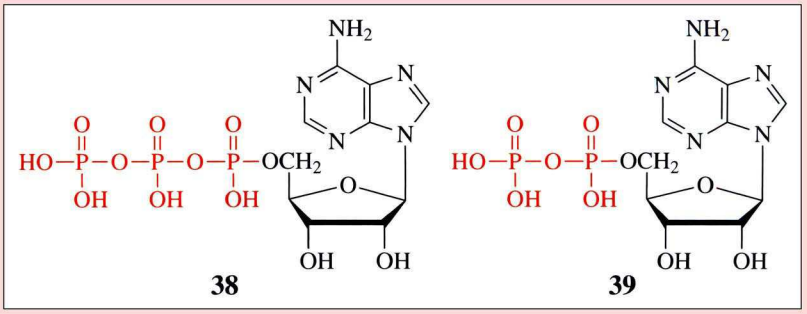
A very common method used to describe the sizes of DNA molecules is in terms of how many **thousands of bases (kb)** or **millions of bases (Mb)** that they contain. So the mammalian mitochondrial DNA has a size of about 15 kb, and the DNA molecules in the mouse genome range from 58 Mb to 192 Mb. (More properly these sizes should be given as 15 **kbp**, *etc.*, where the "p" stands for pairs. The reason for this will become apparent in the next chapter).

Box 2.5 Other Biological Roles of Nucleoside Phosphates

Phosphate esters of the nucleosides are important not only as constituents of the nucleic acids – they also play a very large number of other roles in biology. Particularly important is the 5′-triphosphate ester of adenosine, universally known as **ATP (38)**. All of the oxidative metabolism of materials that we eat is coupled to the production of ATP. For example, the process of **gylcolysis**, by which

Strictly speaking, the abbreviation ATP is ambiguous because it does not specify to which position of the ribose ring the triphosphate group is attached. Unless stated otherwise, the point of attachment is assumed to be the 5′-position. For example, the abbreviation AMP is always taken to mean adenosine 5′-monophosphate.

glucose is oxidized to 2-oxopropanoic acid (pyruvic acid), is accompanied by the production of two molecules of ATP formed by phosphorylation of adenosine diphosphate (**ADP**, **39**). The process is complex, involving about a dozen consecutive reactions, but for our purposes it can be summarized as in Scheme 2.6.



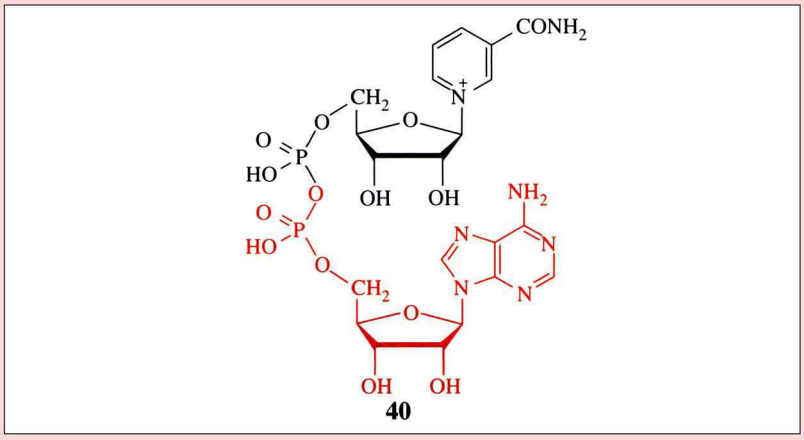
$$C_6H_{12}O_6 + 2\,ADP + 2\,P_i + 2\,NAD^+ \longrightarrow 2\,MeCOCO_2H + 2\,ATP + 2\,NADH + 2\,H^+ + 2\,H_2O$$

Glucose — 2-Oxopropanoic acid

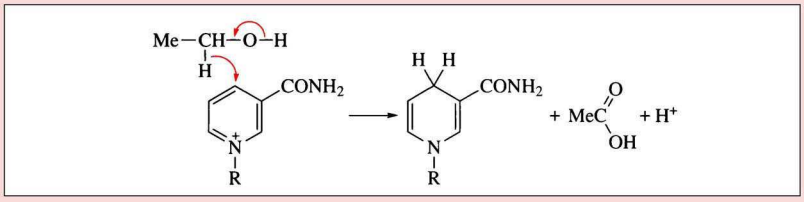
Scheme 2.6

Many enzymes require non-amino acid components in order to carry out their particular reactions. The requirement of dehydrogenases for NAD^+ is just one example of this. When, as with NAD^+, the non-amino acid molecule only attaches to the enzyme at the time that the reaction occurs, it is referred to as a **coenzyme**. In other cases, the non-amino acid component forms a permanent part of the structure of the enzyme when it is referred to as a **cofactor**. Many coenzymes and cofactors are derivatives of **vitamins**. For example, the nicotinamide part of NAD^+ is derived from nicotinic acid (vitamin B_3, niacin). Humans lack the ability to synthesize this substance and must obtain it from their diet.

There are some points that require explanation in Scheme 2.6. Firstly, the symbol P_i is used by biochemists to represent the inorganic phosphate molecule, H_3PO_4, or any of its ionized forms. Secondly, although the process of gylcolysis is an oxidation (two molecules of pyruvic acid contain four less hydrogen atoms than does glucose), it does not involve oxygen as the oxidizing agent. Rather, the oxidation is carried out by a molecule called **nicotinamide adenine dinucleotide** (NAD^+); this is one of Nature's most frequently used oxidizing agents. Its structure is shown in **40**. Note that the part in red is the same as adenosine monophosphate. The part in black is also a nucleotide (hence the dinucleotide part of the name of the compound), but in this case the base is **nicotinamide**. It is the nicotinamide part of the molecule that is involved in oxidation reactions. A typical example, the oxidation of ethanol to ethanal, is shown in Scheme 2.7. This reaction, which is the process by which we metabolize ingested alcohol, is catalysed by the enzyme **alcohol dehydrogenase**. There are many dehydrogenase enzymes, each using NAD^+ as the oxidizing agent but specific for the molecule (the **substrate**) that is oxidized.



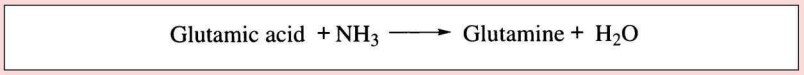
40



Scheme 2.7

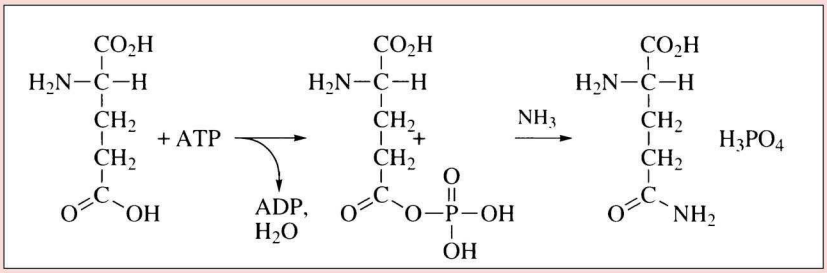
To return to ATP, the reason why it is produced in oxidative metabolism is that it is required for a vast array of biological processes. For example, it is crucially involved in the process of muscular contraction. It is also essential for many biosynthetic reactions, where its involvement is to change the apparent equilibrium position of a reaction. Consider, for example, the synthesis of glutamine from glutamic acid and ammonia (Scheme 2.8). The equilibrium for this reaction lies to the left (the ΔG° value is about 15 kJ mol^{-1}). What actually happens is the pair of coupled reactions shown in Scheme 2.9. ATP acts as a phosphorylating agent and reacts with glutamic acid to yield glutamyl-γ-phosphate. This is a reactive acid anhydride, and reacts with ammonia to yield the product glutamine. The equilibrium for the overall process lies to the right, with a ΔG° value of about -16 kJ mol^{-1}. Many biosynthetic processes proceed by coupled reactions of this sort.

The equilbrium constant, K, for a reaction is relate to ΔG° by $\Delta G^\circ = -RT\ln K$, where R is the gas constant and T is the absolute temperature.



Scheme 2.8

Scheme 2.9



There are many other examples of biological processes that involve nucleotides, and interested readers can find out about these from standard textbooks of biochemistry. The point that is of interest here is that Nature uses the same molecules to do many different things. In the examples that we have looked at here, adenosine is part of the structure of RNA, of ATP and of NAD^+, and as such is involved in processes as diverse as expression of genetic information, biosynthetic reactions and biological oxidations.

Summary of Key Points

1. Nucleic acids are polymers with a backbone consisting of alternating sugar and phosphate groups.

2. The sugar in RNA is β-D-ribofuranose and that in DNA is β-2-deoxy-D-ribofuranose. In the context of nucleic acid chemistry, the sugars are usually called simply ribose and deoxyribose.

3. Each sugar unit has attached to it a heterocyclic base. The bases in RNA are adenine and guanine (which are purines), and cytosine and uracil (which are pyrimidines). In DNA, thymine replaces uracil.

4. Nucleosides are derivatives of ribose and of deoxyribose in which the hydroxyl group on C-1 is replaced by a base.

5. Nucleotides are phosphate esters of nucleosides. It is the nucleotides that are the basic structural units of nucleic acids.

6. Nucleotides in nucleic acids are joined by a phosphodiester linkage between the 5′-hydroxyl group of one sugar and the 3′-hydroxyl group of the next residue in the chain.

7. Nucleic acid structures are represented in shorthand as a string of four letters, each of which is the initial letter of the nucleoside or base that is found at that position in the chain.

8. Small chains of nucleotide units are referred to as oligonucleotides.

9. Nucleic acids are very large molecules, ranging in size from about 100 residues for a transfer RNA molecule up to more than one hundred million residues for a chromosomal DNA molecule.

10. As well as being the component parts of nucleic acids, nucleotides play many other roles in biological systems. In particular, they are important as coenzymes and cofactors of many enzymes.

Problems

2.1. Draw the structure of α-D-lyxofuranose as a Howarth projection and number the carbon atoms.

2.2. As explained in Worked Problem 2.3, transfer RNA molecules contain small amounts of unusual nucleosides. Another of these is 5,6-dihydrouridine (known as UH_2). Draw its structure.

2.3. Draw the structures and give the names of the three monophosphate esters that can be formed from guanosine.

2.4. Draw the structures of the tetranucleotides (4mers) pTCGA and pAGCT.

2.5. Deoxyribonucleases (DNases) are enzymes that catalyse the hydrolysis of the phosphodiester bonds in DNA molecules. The enzyme from the pancreas cleaves the bond between phosphate and the 3′-hydoxyl group of deoxyribose. The enzyme from the spleen cleaves the bond between phosphate and the 5′-hyroxyl group. Tabulate the products that would be formed by both of these enzymes acting on the oligonucleotides in Problem 2.4 and state which enzyme could be used to distinguish between them.

2.6. Caffeine is the stimulant compound in coffee. It is 2,6-diketo-1,3,7-trimethylpurine. It works by prolonging the activity in the liver

of a molecule called cyclic AMP (cAMP, adenosine 5′,3′-cyclic phosphate). The result of this is to increase the level of glucose in the blood stream with resultant stimulatory effects. Give the structures of caffeine and of cAMP.

Reference

1. D. G. Morris, *Stereochemistry*, The Royal Society of Chemistry, Cambridge, 2001, p. 37.
2. A. Kossel, *Hoppe-Seyler's Z. Physiol. Chem.*, 1880, **4**, 294.
3. A. Kossel and A. Neumann, *Ber. Dtsch. Chem. Ges.*, 1894, **27**, 2221.
4. P. A. Levene, L. A. Mikeska and T. Mori, *J. Biol. Chem.*, 1930, **85**, 785.
5. P. A. Levene and W. A. Jacobs, *Ber. Dtsch. Chem. Ges.*, 1909, **42**, 1198.
6. P. A. Levene and R. S. Tipson, *J. Biol. Chem.*, 1935, **109**, 623.
7. A. M. Michelson and A. R. Todd, *J. Chem. Soc.*, 1955, 2632.
8. D. M. Brown and A. R. Todd, *J. Chem. Soc.*,1958, 52.

Further Reading

R. L. P. Adams, J. T. Knowler and D. P. Leader, *The Biochemistry of the Nucleic Acids*, 11th edn., Chapman and Hall, London, 1992.

3

The Three-dimensional Structure of DNA and its Implications for Replication

Aims

By the end of this chapter you should understand:

- The three-dimensional structure of DNA and how it was determined
- How to use a computer to produce three-dimensional models of molecular structures
- Why DNA exists as double helical structure
- Conditions under which the double helical structure breaks down
- How DNA is packaged into chromatin in the cell
- How DNA is replicated
- The damage that can be done to DNA by various chemical and physical agents, and how that damage is repaired

3.1 The DNA Double Helix

3.1.1 The Structure in Outline

On 25 April 1953, a single-page paper appeared in *Nature*[1] which revolutionized the understanding of DNA and which won its authors, James Watson and Francis Crick, a Nobel Prize. The essential feature of their proposal was that the structure consists of two polynucleotide chains, each with a right-handed helical structure, coiled around the same axis and running anti-parallel to one another (Figure 3.1). That is, from top to bottom of the structure, one chain runs 5′→3′ and the other 3′→5′ (see Section 2.3). The sugar–phosphate backbones in

A copy of the paper can be obtained from *Nature's* archive of material related to the 50th anniversary of the discovery of the structure at http://www.nature.com/nature/dna50/archive.html. Also available are copies of papers by Wilkins, Stokes and Wilson, and by Franklin and Gosling, which accompanied the Watson and Crick paper, and which gave some of the X-ray diffraction data on which the structure was based.

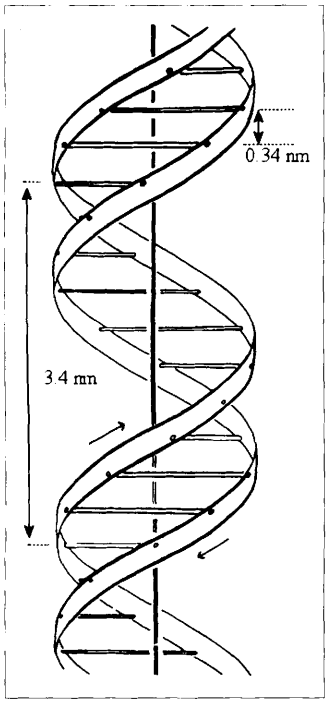
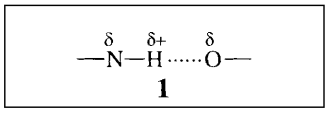


Figure 3.1 Cartoon of the double helical structure of DNA. Adapted from Watson and Crick[1]

Watson and Crick quoted the dimensions of their model in **angstrom units** (Å). This is the unit almost always used by crystallographers. The unit is named after the 19th century Swedish physicist A. J. Ångström. 1 Å is equal to 10^{-10} m or 0.1 nm. It is a convenient unit because atom sizes and bond lengths are in the range 1–2 Å, so its use avoids decimals. It is, however, frowned upon by adherents of SI units and will not be used here. You must, however, expect to see it widely used in the scientific literature. In these terms, the pitch of the double helix is 34 Å and its diameter is 20 Å.



1

Figure 3.1 are represented by ribbons. Between the ribbons, and holding the chains together, are structures shown as rods, each of which consists of a pair of bases with their planes perpendicular to the axis of the helix. There are 10 pairs of bases for each complete turn of the helix; that is, pair number 11 lies directly above pair number 1. The vertical distance travelled up the helix axis between each pair of bases is 0.34 nm, and so the pitch of the helix (the distance up the axis for one complete turn) is 3.4 nm. The diameter of the helix is about 2.0 nm.

Perhaps the most important part of the proposal put forward by Watson and Crick was that the bases are **hydrogen bonded** together (see Box 3.1), and that wherever an adenine occurs in one chain, then a thymine must occur in the other. Similarly, wherever guanine occurs in one chain, cytosine occurs in the other. Watson and Crick finished their paper with what may be the most famous throwaway remark in the scientific literature: "It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic information". That is, they had effectively solved the problem of how genetic material is replicated! We will return to this in Section 3.5.

Box 3.1 Hydrogen Bonds

When hydrogen is bonded to an electronegative atom, the bond is polarized so that there is a partial positive charge on the hydrogen. Because hydrogen has no non-bonding electrons, the resulting partial positive charge interacts strongly with other electronegative atoms, and the resulting interaction is referred to as a **hydrogen bond**. The hydrogen bond can be considered as electrostatic in nature, and is usually represented as in **1**, where nitrogen has been chosen as the donor atom and oxygen as the acceptor. It is relatively weak, with bond energies in the range 10–30 kJ mol^{-1} (that is, about 10% of the value for a normal covalent bond). It is also a rather long bond. For example, in the case shown in **1** the H–O distance will be about 0.2 nm; that is, the N–O distance will be about 0.3 nm. Hydrogen bonds are of enormous importance for the structures and activities of biological molecules. We are concerned here with the role that they play in the structure of DNA and in specific base pairing. It is, however, worth noting that hydrogen bonding plays a major role in the way in which protein molecules fold up into precisely determined three-dimensional structures.[2]

Box 3.2 How the Structure was Discovered

It is interesting to look at the experimental evidence on which Watson and Crick based their proposed structure of DNA, and how they put the pieces together. Key to the solution of the structure was the production of high-quality X-ray diffraction patterns of DNA fibres. It is easy to make fibres of DNA by taking a very concentrated solution of the molecule, touching it with a glass rod, and then withdrawing the rod. This draws out a very thin fibre of DNA. X-ray diffraction studies of such fibres were initiated in the late 1940s in King's College, London, by a research group led by Maurice Wilkins. In 1951, Rosalind Franklin joined the group and, together with Raymond Gosling, carried out a study of the effects of humidity on the diffraction patterns obtained from DNA fibres. They observed two main types of pattern. One type, obtained at low humidity, arose from an ordered, crystalline form of DNA; this had previously been observed by Wilkins and Gosling, and was called the **A-form**. At high humidity, a much simpler diffraction pattern was obtained, and the DNA giving rise to this was called the **B-form**. A diffraction pattern from B-DNA is shown in Figure 3.2.

Franklin tentatively interpreted the diffraction pattern in terms of a helical structure, with the phosphates on the outside, and consisting of two or three polynucleotide chains. Certain quantitative features of the helix could also be obtained from the pattern. For example, the very strong reflections at the top and bottom of the pattern arise from a repeating distance of 0.34 nm along the fibre axis. In addition, the X-shaped pattern at the centre can be interpreted in terms of the angle of ascent of the helix. Franklin did not, however, speculate on a detailed structure based on these results. Rather, she turned her attention to working on the A-form because it gave a more clearly defined diffraction pattern which she hoped to be able to interpret, and in so doing left the field open for Watson and Crick to solve the structure of the B-form.

Watson and Crick were working in Cambridge on projects not concerned with DNA. They were, however, very interested in the DNA problem, and set out to try to solve it. Key to their ultimate success was Watson's obtaining sight of the diffraction pattern of the B-form recorded by Franklin. Crick was well placed to interpret this pattern, because he and other colleagues in Cambridge had, around this time, worked out the theory of X-ray diffraction by helices.[3] In addition to providing them with the dimensions of the helical structure, examination of the diffraction pattern also allowed Crick

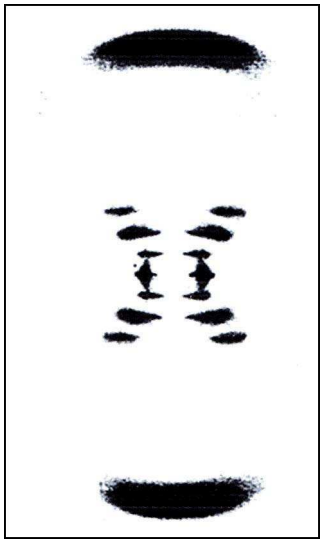


Figure 3.2 X-ray diffraction pattern of a fibre of B-DNA. Reproduced by permission of the The Nobel Foundation, 1962

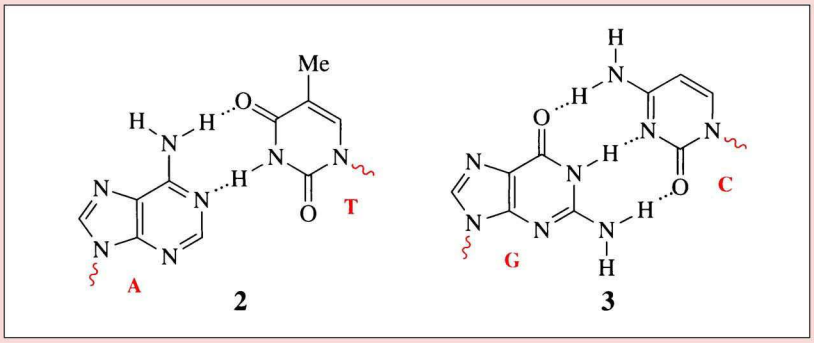
to reach the crucial conclusion that the molecule had a two-fold symmetry axis perpendicular to the fibre axis. That is, the structure remains unchanged after rotation through 180° around that axis. This was interpreted in terms of a helix consisting of two polynucleotide chains, with the chains running in opposite directions; two chains running in the same direction would not give rise to a two-fold axis of symmetry.

So, at this stage, Watson and Crick knew that they were dealing with a double helix with anti-parallel chains, and they knew its dimensions. Moreover, it was reasonable to assume that the 0.34 nm repeating distance along the fibre axis arose from stacking of the bases on top of one another. They had one other important piece of information. A Norwegian scientist, Sven Furberg, had determined the structure of a nucleoside, cytidine, using single-crystal X-ray diffraction in 1949.[4] This provided important geometrical information, assuming that the geometry of the nucleosides in DNA was the same as that in the crystal of cytidine.

Armed with these various pieces of information, Watson and Crick set out to try and build a model of the structure. Building a model of the backbone was relatively straightforward. The difficult part of the problem was how the bases fitted in. Part of the answer came from the fact that the space between the two strands of the double helix was only sufficient for a purine and a pyrimidine pair. Two purines would not fit. Two pyrimidines left a hole in the middle of the structure. There was evidence from the physicochemical properties of the nucleic acids that the bases were hydrogen bonded together, but there are many ways in which the purine and pyrimidine bases in free solution could interact. Crucially, a colleague in Cambridge, Jerry Donohue, pointed out that the bases were likely to be in the keto rather than the enol form (see Scheme 2.1). Watson finally solved the problem by making cardboard cut-outs of the bases in the keto configuration and juggling them around! He came up with the discovery that if A was paired with T, and if G was paired with C, the dimensions of the base pairs were exactly the same. Moreover they fitted snugly into the space in the centre of the double helix. The structure was solved.

The base pairs formed between adenine and thymine, and between guanine and cytosine, are shown in **2** and **3**, respectively. The red "squiggly bonds" show the point of connection between the bases and C-1′ of the deoxyribose residues. The hydrogen bonds are shown by dashed lines. Note that there are two hydrogen bonds between adenine and thymine, but three between guanine

and cytosine. In the original pairing scheme developed by Watson and Crick, only two hydrogen bonds were thought to exist between G and C.



2 3

There was, in fact, another piece of evidence available that would have made identification of the base pairs easier if Watson and Crick had realized its significance. Erwin Chargaff and his co-workers had carried out careful quantitative analyses of the base compositions of DNA from a variety of sources.[5] They found that the ratios of the content of adenine to that of guanine, and of thymine to cytosine, varied widely from one organism to another. Most importantly, however, the amount of adenine was found to equal that of thymine, and the amount of guanine to equal that of cytosine, irrespective of the source of the DNA. Watson and Crick's model required that a purine be paired with a pyrimidine. Chargaff's rules stated that the content of A was equal to that of T, and the content of G was equal to that of C. The obvious base pairs were then A/T and G/C.

It is not surprising, given the importance of the structure of DNA, that a vast amount has been written about how it was arrived at. A personal account of the discovery of the structure has been written by Watson.[6] The book gives a rather biased account of the story and does not do justice to contributions made by other workers in the field, but it is well worth reading to get a feeling for the sense of excitement that was felt by him and Crick as they raced to be first to solve what they considered to be one of the most important problems in biological sciences.

A review of the story from Wilkins' perspective is given in his Nobel Lecture (http://www.nobel.se/medicine/laureates/1962/wilkins-lecture.pdf), and an independent version of the story has been written by Olby.[7]

What we do not hear in all this is the voice of Rosalind Franklin. Unfortunately, she died form ovarian cancer in 1958 at the very

Linus Pauling, a major figure in the structural chemistry of biological macromolecules, was also hot on the trail. He had earlier predicted the two most important types of regular structure found in proteins – that is, the **α-helix** and the **β-sheet**. Hence he was used to thinking about helical structures, and Watson and Crick knew from Peter Pauling, Linus' son, who had recently arrived in Cambridge, that his father was actively pursuing the DNA structure. It was also possible that the King's College group would come up with it.

Crick, Watson and Wilkins were awarded the Nobel Prize in Medicine in 1962 for "their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material".

young age of 37. It would have been interesting to see if she would have shared in the Nobel Prize in 1962 had she survived (Nobel Prizes are not awarded posthumously) as she so richly deserved to do, since her diffraction patterns were the key to the structure. It may be wondered why, given that the King's group had the essential information to hand, they did not come up with the structure. At least part of the answer lies with the fact that personal relations between Franklin and Wilkins were very bad, and so they did not cooperate and bring their respective talents to bear on the problem (it is interesting to note that their names have now been coupled by King's College, where a new building has been named the Franklin-Wilkins Building). Watson's uncharitable view, expressed in his book, is that Franklin lacked the insight to come up with the solution. There is every reason to believe that this was not true, and indeed a manuscript was found amongst Franklin's papers, written before Watson and Crick's paper was published, giving the essentials of the structure. It was probably not submitted for publication because Franklin was cautious about putting forward a structure based on what she considered to be incomplete evidence. Interesting books have been written by Sayre[8] and by Maddox[9] charting her brief scientific life and re-evaluating her role in the DNA story.

Finally, it is worth emphasizing that Watson and Crick did not produce any of the experimental data on which the structure of DNA was based. What they did was to interpret brilliantly the results obtained by others. The key experimental result was the fibre diffraction pattern of B-DNA, and Watson and Crick have been criticized for making use of this without Franklin's permission. Unfortunately, the personalities of the main players in the drama did not allow a cooperative effort to obtain the structure.

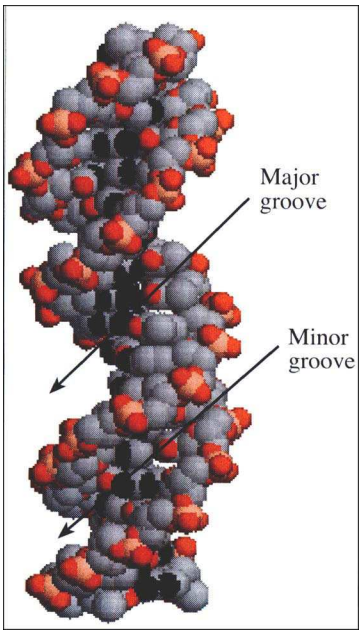


Figure 3.3 Spacefill model of a section of a B-DNA molecule

3.1.2 The Structure in Detail

Figure 3.1 gives only the bare outlines of the structure of DNA. Figure 3.3 shows a small section of a B-DNA molecule at atomic resolution. This structure was obtained by **single-crystal X-ray diffraction analysis**. We cannot go into details of how this technique is carried out, but for the interested reader a very good introduction to the subject is given in the book by Rhodes.[10]

The structure in Figure 3.3 is shown as a **spacefill** model in which each atom is represented as a sphere of radius proportional to its van der Waals radius. For clarity, hydrogen atoms are not shown; hydrogens occupy all spare valencies. The colour code used in Figure 3.3 is grey

for carbon, black for nitrogen, red for oxygen and light brown for phosphorus. The structure is orientated in the same way as is the cartoon in Figure 3.1, and so it is easy to trace the two chains through the molecule by comparison of the two Figures.

This is not the colour scheme conventionally used. Nitrogen is usually coloured blue and phosphorus yellow. Those colours are not available here.

The merit of a model of this sort is that it allows surface features of the molecule to be clearly seen. It should be immediately apparent that the phosphate groups are on the outside of the molecule and in contact with the solvent. Interaction of the negative charges on the phosphates with the dipoles of water molecules is one of the factors that stabilizes the structure of DNA (see Section 3.2). It is also clear that the bases are in the interior of the molecule, but no details of their interactions can be seen.

High-resolution structures of proteins were obtained some time before those of nucleic acids, in spite of the fact that proteins are much more complicated molecules. The reason for this is that progress with nucleic acids could not be made until methods were developed for the synthesis of oligonucleotides with defined sequences (see Section 5.10).

The other notable feature of the structure is that it has two grooves running diagonally through it. One is much wider than the other and is termed the **major groove**. The smaller one is the **minor groove**. The edges of the bases are exposed in these grooves and provide sites where other molecules can interact with the DNA. Interactions with specific proteins which are designed to fit into the grooves are very important for regulating the expression of genes.

Box 3.3 Molecular Modelling Using RasMol

Professional crystallographers also make use of physical models of molecules for these purposes. For large molecules such as nucleic acids, however, these are enormously difficult and costly to build, and are not a viable alternative for the more casual user.

Pictures of models of molecular structures such as that in Figure 3.3 are limited in their ability to provide information about important structural features. They provide a single, flat view. It is possible to construct stereo views which, with practise, allow the model to be seen in three dimensions (see Figure 3.6), but to appreciate fully what a structure looks like, and even more to obtain answers to questions about its features, there is no practical alternative to using a molecular modelling program to draw it on a computer screen. The structure can then be rotated to see it from any direction, parts of it can be selected for more detailed views, distances between features can be measured, and so on.

The molecular graphics program used to produce the images in this book is RasMol. This program was created by Roger Sayle, who made it freely available to the scientific community. It is now maintained at the University of Massachusetts. You are strongly urged to obtain a copy and install it on your computer, otherwise you will not be able to do some of the problems that follow.

First create a directory to put the program in (*e.g.* C:\RasMol), and then go to the RasMol Home Page at http://www.umass.edu/microbio/rasmol/, where you will find a description of the program

and a link to take you to the download page. You will need to look through the information in that page to decide which version of the program you need (which depends on the characteristics of your computer system), but it is then just a matter of downloading it into the directory you have created, and then installing it. Be sure to download the Help files and the RasMol Reference Manual at the same time.

You will obviously need some structures to draw with RasMol. These can be downloaded from the database of the **Research Collaboratory for Structural Bioinformatics** (RCSB) at http://www.rcsb.org/index.html (see Worked Problem 3.1).

The atom type and its x,y,z coordinates are the raw material that RasMol uses to draw the molecule. The computer screen is treated like a piece of graph paper, with the x-axis horizontal and the y-axis vertical. The x,y coordinates of each atom are then used to determine where that atom is placed on the screen. The type of the atom (carbon, oxygen, *etc.*) determines its colour and, for spacefill models, its size. The z coordinate is effectively ignored; that is, the molecule is flattened onto the plane of the screen (except, of course, that if one atom is behind another, then it is not drawn). It becomes important, however, if the molecule is rotated, because where an atom ends up on the xy plane depends on how far it was above or below it before rotation.

To begin modelling, click on the RasMol icon that will have been set up when the program was installed. This opens a window with a black drawing surface. At the top of the window are several pull-down menus. Left clicking on *File*, then on *Open* produces a window listing whatever files are in the directory that are readable by RasMol – there should be at least one, 1D29. Clicking on that file produces a wireframe model of a DNA molecule in RasMol default colours. The molecule can be translated along the x- and y-axes by right clicking and moving the mouse left/right and up/down, respectively. Left clicking and moving the mouse left/right or up/down rotates the molecule around the y- and x-axes, respectively. Rotation around the z-axis is done by holding down the shift key whilst right clicking and moving the mouse left/right. The size of the molecule can be changed by holding down the shift key whilst left clicking and moving the mouse up/down.

Besides these basic mouse operations, much can be down with the pull-down menus, including changing the colour scheme of the molecule, and selecting between display modes (wireframe, ball and

stick, spacefill, *etc.*) However, at the bottom of the screen will be found a button labelled *RasMol Command Line*. Clicking on this opens a new window into which commands can be typed at the *RasMol>* prompt. The real power of RasMol is only obtained by making full use of this facility. The possible commands and what they do can be discovered from the user manual under the *Help* pull down, or from the reference manual. There is no substitute for studying these sources of information and trying the commands out. Some examples of commands and their use will be found in the Worked Problem 3.2.

Worked Problem 3.1

Q Download the coordinates of the DNA molecule shown in Figure 3.9 from the RCSB database.

A First log on to the RCSB database. On the home page, click on *DATABASES*, and on the next page click on *PDB*. When the PDB page opens, you will see a search box. Enter the code *1D29* and click on *Find a Structure*. After a few seconds a page will open entitled *Structure Explorer – 1D29*. This page includes a thumbnail model of the structure at top right, some information about the molecule, and a set of links to other pages. Click on *Download/Display File*. In the next page that opens, go to *Download the Structure File* and click on the "X" at the top left of the table (no compression, PDB format). A new window will open for your RasMol directory (if it opens to some other directory, then change it to RasMol). Click on *Save* to download the file.

There is a vast amount of information in the RCSB database and it is well worthwhile looking around it. You can, of course, download the structure of any other nucleic acid or protein molecule that you might be interested in (provided, of course, that its structure has been determined!). To do this, it is useful to use the customizable search facility that is available from the PDB page.

You might also like to look at what is contained in the structure file that you have downloaded. It can be opened in Word or Note Pad and read like a text file. The beginning of the file gives information about the sequence of the molecule, how the structure was determined, and who did the work. The bulk of the file consists, however, of a very boring list of the coordinates of all the heavy atoms in the molecule.

PDB stands for Protein Data Bank and it was set up to be a repository for all known three-dimensional structures of proteins.[11] When structures of nucleic acids began to become available, the scope of the database was increased to include these, but the original name has stuck.

3.1.3 The Structures of the Components of DNA

Whenever a structure has been taken from the PDB, its identification code is given along with a reference to the authors of the work.

Because DNA is such a large complicated molecule, it is difficult to see how it is built up and how the component parts interact. To make it easier, we will look at the molecule a bit at a time, starting with the nucleotide and working up. Figures 3.4–3.9 were all produced by extracting the features of interest from the structure of the 12mer CGTGAATTCACG. This is entry 1D29[12] in the RCSB database, and was the structure downloaded in Worked Problem 3.1.

Figure 3.4 shows a **ball-and-stick** model of an adenylic acid residue extracted from the structure. For small molecules or fragments such as this, the ball-and-stick model provides the clearest picture of the structure. Atoms are represented by small balls of the appropriate colour, and the bonds as sticks joining them. Where a bond joins two different atoms, the stick is half one colour and half the other.

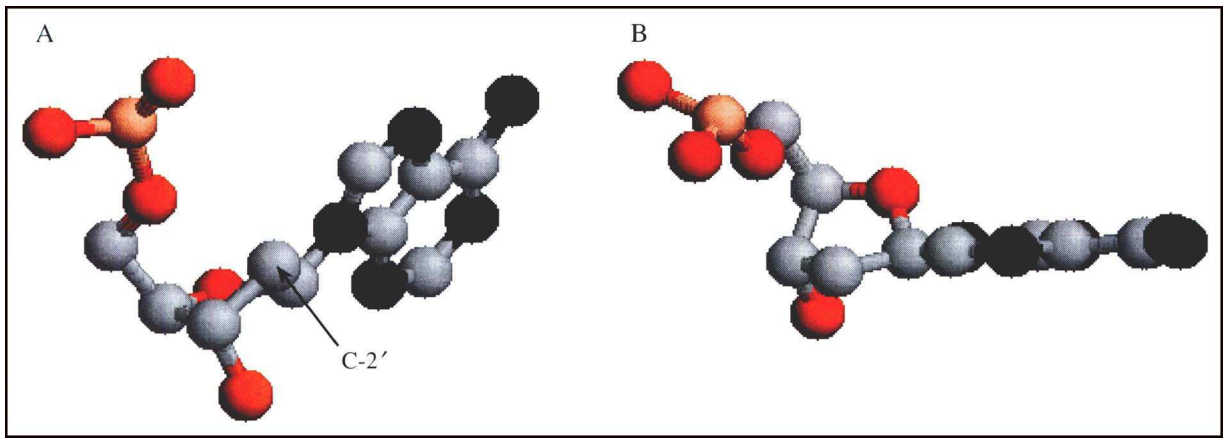


Figure 3.4 Two views of a nucleotide from a DNA molecule. View A is oriented to show the conformation of the deoxyribose ring (the C-2′ atom is indicated). In view B the molecule is rotated to show the planarity of the adenine ring

Recall that, as discussed in Section 2.2, a prime is added to the numbers of the atoms in the deoxyribose ring when it is component of a nucleoside or nucleotide.

Part A of the figure is oriented in such a way as to show the stereochemistry of the deoxyribose. The Howarth projection of deoxyribose (**9** in Chapter 2) suggests that the ring is flat, and indeed on geometric considerations alone it might be expected to be so. The angles in a regular pentagon are 108°, which is close to the preferred bond angle of a tetrahedral carbon atom (109.5°), and so a planar ring would have little bond angle strain. It would, however, have significant steric interaction between C-5 and the hydroxyl group on C-1. The steric interactions would be even greater in derivatized molecules such as nucleosides. The strain is relieved by a conformational shift in which C-2′ moves up above

the plane of the ring and C-3′ moves down, so that the ring is puckered as shown in Figure 3.4A. Again, for steric reasons, the base is oriented away from the sugar–phosphate group. With the deoxyribose ring perpendicular to the plane of the paper as in Figure 3.4A, the base is approximately parallel to that plane.

Figure 3.4B shows the adenine ring system edge-on, and makes the very important point that adenine is planar (as, indeed, are all the bases in DNA and RNA). The reason for this is that the bases are aromatic, and the formation of the π-electron bonding system requires planarity of the rings.

Moving up to the level of the dinucleotide (Figure 3.5), the beginnings of the twist in the backbone can be seen, with the planes of the deoxyribose rings inclined to one another. The essential feature here, however, is that the planes of the two bases are both perpendicular to the plane of the paper. That is, the bases are stacked on top of one another. This is possible because rotation can occur around the single bond linking C-1′ of the sugar to N-9 of the base (or the N-1 in the case of a pyrimidine). This stacking is a key feature of DNA.

The chemistry of aromatic systems has been dealt with in a previous volume in this series[13] and will not be further considered here.

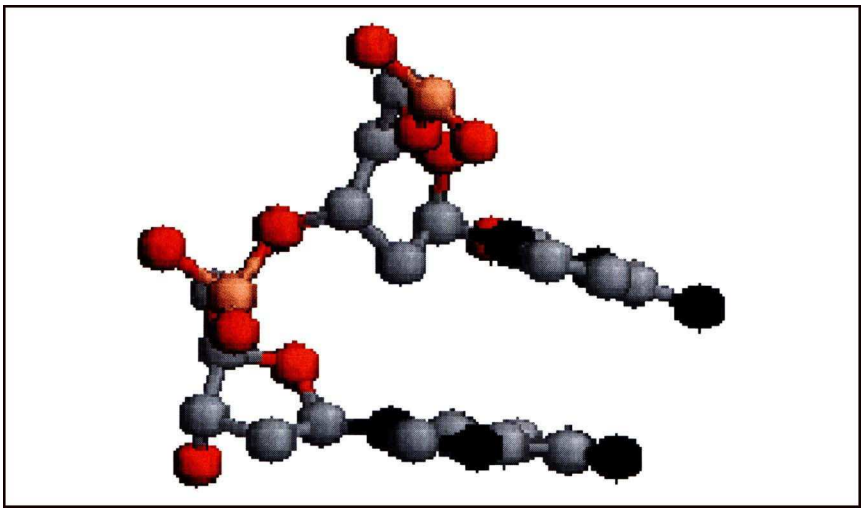


Figure 3.5 A dinucleotide from DNA

In Figure 3.6, we move up in scale to a hexanucleotide sequence; that is, one half of a turn of the helix. In this case, a **wireframe** model is used in which only the bonds joining the atoms are shown. This is because a ball-and-stick representation of a molecule as large as this would be very cluttered, making it difficult to see what is going on. We will return shortly to why there appear to be two identical copies of the molecule in this figure.

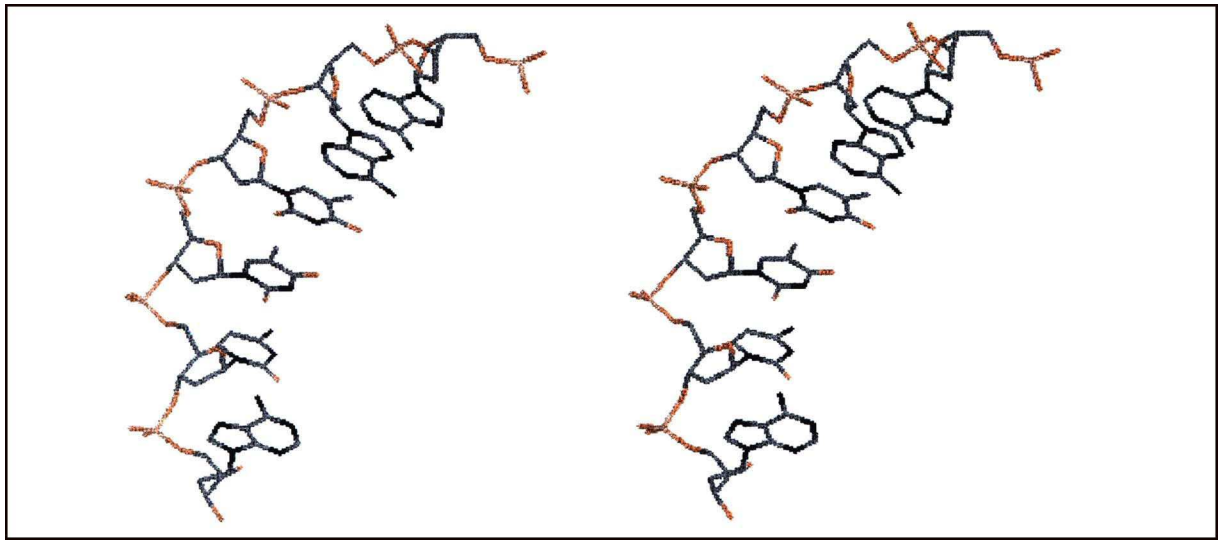


Figure 3.6 Stereo image of a hexanucleotide from DNA. This represents one half of a turn of the α-helix

Look first at the structure on the left of Figure 3.6. Its salient features are as follows. Firstly, from bottom to top of the model, the backbone goes back into the plane of the paper and then comes forward again; that is, it has a helical twist. Secondly, as a result of the twist, the adenines in the first and last positions point in opposite directions. Look, for example, at the NH_2 groups on position 6 of the adenine rings of the first and last residues in the chain; the one at the bottom of the molecule is pointing backwards, and the one at the top is pointing forwards. The reason for this is that the orientation changes by 36° from one position to the next. So in going through five nucleotides (from 1 to 6) the total rotation is 180°. The final important feature is that the bases are all stacked on top of one another.

Trying to see all of these features emphasizes the limitations of two-dimensional representations of three-dimensional objects. For example, is the NH_2 on the adenine ring at the bottom really pointing backwards, or is it pointing forwards? Are the bases really packed on top of one another? Stereo images are invaluable in answering such questions.

The two structures in Figure 3.6 are not, in fact, identically positioned. Rather, one is rotated by a few degrees with respect to the other. This allows for a technique known as **naked eye stereopsis** by which the structure to be seen in three dimensions. To do this, look at the page between the two models in Figure 3.6 and let your eyes go out of focus. You will see copies of the two images float towards one another. When they superimpose, the combined image will suddenly spring out

into three dimensions and the features described above will be obvious. It takes a little practice to get the effect first time, but after that it is easy. However, for those who cannot do this, stereo viewers are commercially available.

3.1.4 The Complete Structure

So far we have been looking at the components one chain of the double helix. We now need to put them together; that is to look at the base pairs that form between A and T, and between C and G, and how these fit into the double helical structure. Figure 3.7 shows a small region of a double helix that contains both an A/T base pair (bottom) and a G/C base pair. The hydrogen bonds are shown as dotted lines joining pairs of heavy atoms (oxygen/nitrogen or nitrogen/nitrogen). By convention, the hydrogen atoms themselves are not shown. There are three hydrogen bonds between guanine and cytosine, and two between adenine and thymine.

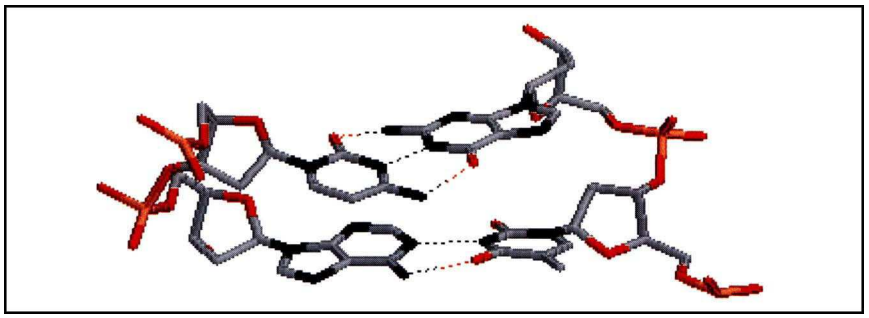


Figure 3.7 Base pairs from a section of a DNA molecule

Worked Problem 3.2

Q Produce a stereo image of the base pairs shown in Figure 3.7. The residues have been extracted from the DNA molecule downloaded in Worked Problem 3.1 (1D29). The residues on the left are C21 and A22 from chain B of the molecule. Those on the right are residues G4 and T3 from chain A.

A Open the file 1D29 in RasMol. The molecule loads such that it is viewed down the axis of the helix. Note the bases stacked in the middle and the sugar–phosphate backbone on the outside. Re-orientate it vertically by right-clicking and moving the mouse downward.

The next thing to do is to change the colour scheme. Open the *Command Line* window and type in the following commands at the *RasMol>* prompt, pressing *Enter* after each.

background white (changes the colour of the drawing canvas)

wireframe 40 (increases the thickness of the bonds; the units are internal RasMol units)

select *.n?? (selects all the nitrogen atoms in the molecule; any type of atom or group of residues can be selected like this. Once a group of atoms or residues has been selected, subsequent commands affect only that group. Another group of atoms or residues, or the entire molecule (*select all*), can be selected at any time)

color black (colours the selected atoms black. Note the spelling of *color*. Some colours such as orange, red, green and blue are pre-defined. Any others, or shades, can be chosen using the command *color [x,y,z]* where *x*, *y* and *z* are numbers between 0 and 250. The colours of the rest of the atoms are OK, but you can change them if you wish).

At this stage we have the whole of the molecule still on the screen. To choose the residues we want, enter:

restrict T3, G4, C21, A22 (restricts the display to the chosen residues)

hbonds (draws in the hydrogen bonds)

Now we have the base pairs that we want. They may not be oriented as in Figure 3.7, in which case re-orient them using the mouse. To create the stereo view, enter:

stereo

If parts of one or both of the stereo images are chopped off, either increase the size of the window by dragging on its margins, or decrease the size of the image by entering:

zoom 80 (or some such number; 100 is the default size, 50 is half size and 200 gives double size).

You should now have an image like that in Figure 3.8. It is much more difficult to do naked eye stereopsis on the screen than on paper, so it is worthwhile transferring the image to paper. Go to the *Export* pull down, and select *GIF* as the file type. Chose a name and a destination for the exported file. The file can be read into a word processor document, or used as input for a program like PaintShop Pro. Studying this image will give you a much better feel for things like how the bases are stacked than does the image in Figure 3.7.

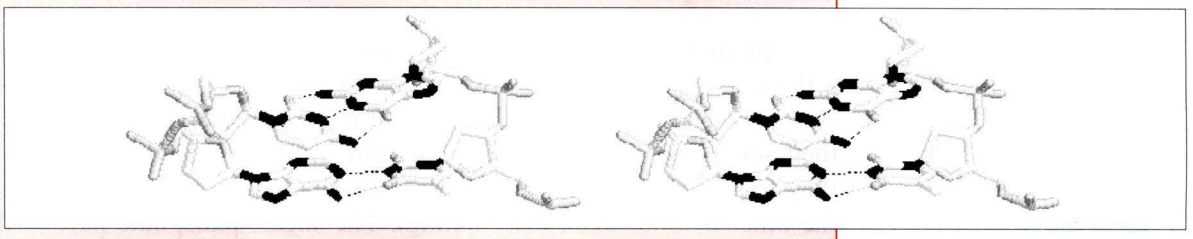


Figure 3.8 Stereo view of the base pairs shown in Figure 3.7

Finally, all of the structural elements that we have been considering are put together in the stereo diagram of the complete structure of the dodecamer in Figure 3.9. The chain starting with cytosine at top left runs 5′→3′ downwards. The chain starting at bottom right runs 5′→3′ upwards. It is instructive to trace as much of the sequences of the chains through the structure as possible.

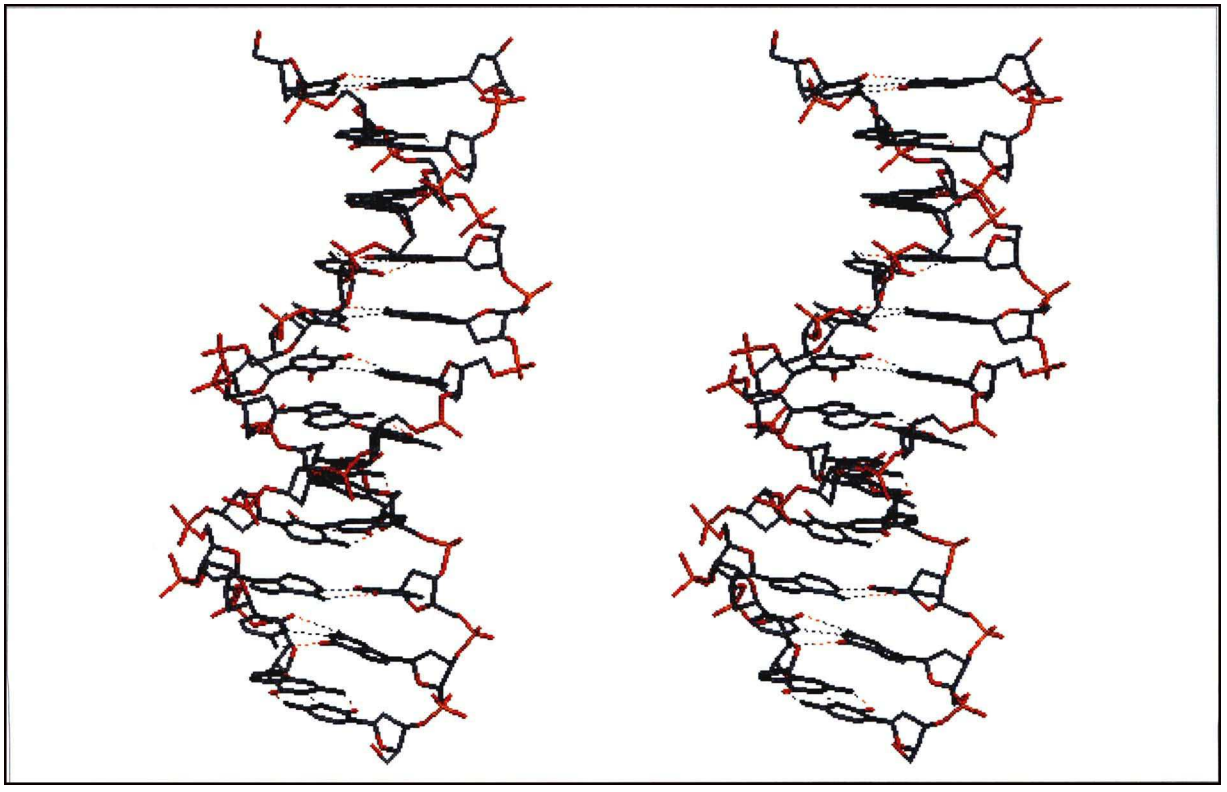


Figure 3.9 Stereo view of a complete turn of the double helix in B-DNA

3.2 Why a Double Helix?

This question can be divided into two parts. Firstly, why does DNA form a double stranded structure? The answer to this can be arrived at by considering the physical properties of the nucleotides, and in particular their solubilities in water. Nucleotides are **amphipathic** molecules; that is they consist of two parts, one of which is **hydrophilic** (water loving) and the other **hydrophobic** (water hating). The sugar–phosphate part of the molecule is polar in nature, and interacts strongly with water. The partial charges on dipolar water molecules interact electrostatically with the oxygens of the deoxyribose and with the charge on the phosphate. The bases, one the other hand, although they contain substituents capable of hydrogen bonding to water, are essentially non-polar, and indeed the free bases are insoluble in water at neutral pH.

The interior of a cell, which is the natural habitat of DNA, is an aqueous environment which is appropriate for the backbone but not for the bases. A solution to these opposing requirements is for the DNA to form a double chain with the sugar–phosphate part on the outside in contact with the surrounding water, and with the bases on the inside. This not only removes the bases from contact with the water but also allows them to form hydrogen bonds with one another, thus increasing the stability of the double stranded structure. In principle, they can also interact with each other by so-called **hydrophobic bonding**.

Box 3.4 Hydrophobicity and the Hydrophobic Bond

Hydrophobicity is quite a subtle concept, although its manifestations are often obvious. For example, if a little olive oil, which is a mixture of triglycerides, is poured into water, then the oil does not dissolve but rather forms droplets. This is explained in terms of "hydrophobic bonding" between the non-polar triglyceride molecules, and could be thought to show that the triglycerides interact favourably with one another but not with water molecules; that is, the explanation is concerned with the enthalpies of interaction. In fact this is not correct; the effect is essentially entropic. A triglyceride molecule in water is surrounded by an ordered layer of water molecules. If the triglyceride is removed from water into a lipid droplet, then the constraints on the water molecules will be removed, resulting in an increase in the entropy of the system. It is this increase in entropy that drives the formation of so-called "hydrophobic bonds". A detailed analysis of the hydrophobic effect has been given by Tanford.[14]

The second part of the question is why a helix? It might be thought that one way in which DNA could form a double stranded structure which satisfies the requirement to put the bases inside and the backbone outside would be to produce a ladder with the base pairs as the rungs. Such a hypothetical ladder is shown diagrammatically in Figure 3.10 with the bases drawn edge-on simply as rectangular objects.

The discussion here is taken from the lucid book by Calladine and Drew.[15] This book is very strongly recommended for anyone who wishes to gain an in-depth understanding of the structure of DNA.

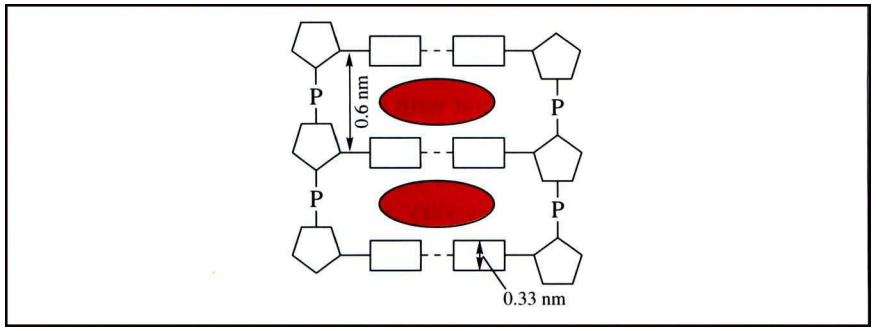


Figure 3.10 A hypothetical ladder formed from anti-parallel polynucleotide chains. The rectangles represent bases viewed edge on. The red ovals show the space that would be occupied by water in such a structure. Adapted from Calladine and Drew[15]

From known bond lengths it can be calculated that the distance between adjacent bases in such a structure would be 0.6 nm. The thickness of the bases is, however, only about 0.33 nm. So such a structure would leave a hole of about 0.27 nm between the bases as indicated by the red ovals. We clearly cannot have a hole, and equally, because of the hydrophobic nature of the bases, the hole cannot be filled with water. This structure will not do.

The question is how to fill up the hole. How this happens is shown in Figure 3.11. Here, the rectangular blocks are base pairs rather than individual bases, and the sugar–phosphate linkages are shown as dumbbells joined to the edges of the bases. In the left-hand diagram the base pairs are separated as in the DNA ladder shown in Figure 3.10. If the lower base pair is kept stationary, and the upper base pair is rotated anti-clockwise around a vertical axis through the centre, at the same time keeping the length of the dumbbells constant, then at some point the gap between them disappears. Given the dimensions shown in Figure 3.10, it can be calculated that a rotation of about 30° is required to close the gap. If such a rotation is repeated all along a ladder like that in Figure 3.10, after 13 base pairs we will have gone through 360°, and the last base pair will lie directly above the first. The ends of the dumbbells will have described a double helix just as we find in DNA. In fact, the calculations given here are not exact, and with the B-form of DNA the required angle of rotation is 36°, so there are 10 base pairs in a complete turn of the helix and the eleventh base pair lies exactly above the first.

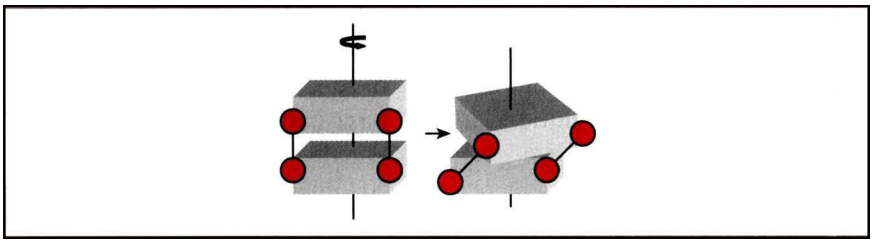


Figure 3.11 Stacking base pairs by twisting. The rectangles represent base pairs. Rotation of the upper base pair in the direction shown closes the gap between them. Adapted from Calladine and Drew[15]

In summary, DNA forms a double helix for two reasons. Firstly, the double stranded structure forms to remove the bases from contact with water and to allow them to interact with each other. Secondly, the length of the sugar–phosphate unit compared with the width of the bases requires that the double chain twists into a helical shape to close up the gap between the base pairs. It is very satisfying to be able to reach a conclusion of fundamental importance on the basis of such simple chemical and geometric arguments.

Box 3.5 Deviations from Ideal Geometry

It is easy to get the impression from a model such as that in Figure 3.1 that DNA is a very symmetric molecule, with planar base pairs strictly perpendicular to the axis of the double helix. This is not the case. Actual DNA molecules show a large number of deviations from this ideal geometry, as close inspection of Figure 3.9 will show. Look, for example, at the second base pair form the bottom in Figure 3.9. The two bases are not co-planar, but rather are twisted with respect to one another. This pair of bases has been extracted in Figure 3.12.

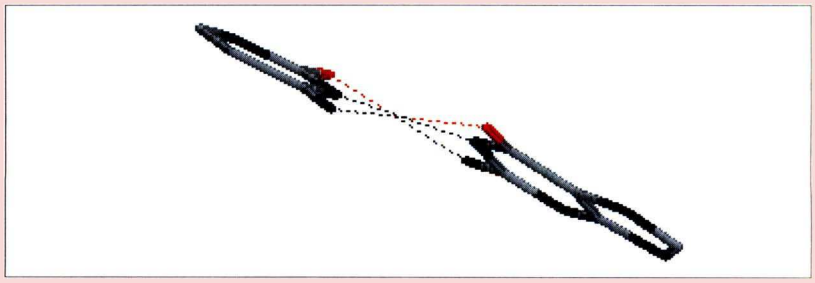


Figure 3.12 The propeller twist in a base pair from DNA. The twist results from rotation of the two bases in opposite directions

Viewed along an axis through both bases, the structure looks a bit like the two-bladed propeller on an old-fashioned aeroplane. For this reason the resulting deviation from a regular structure is called a "propeller twist". A consequence of the propeller twist is that the

hydrogen bonds between the two bases are distorted – they no longer lie directly in line with the heavy atoms. There is a small energetic cost to this, but it is outweighed by the fact that the twisting allows for greater overlap of stacked bases and consequent increase in stability of the structure.

What is not so obvious from Figure 3.9 is that there can be considerable variation in the value of the helical twist angle from the ideal value of 36°. In local regions of real DNA molecules this angle can vary from about 30° to 40°. The backbone bond angles can easily accommodate this range of twist, but it should be clear that different twist angles must be associated with movement of the bases within the structure. The details of this are complicated and we will not go into the matter here; a full treatment of the subject is given by Calladine and Drew.[15] The important point to note is that the precise structure found in any region of an actual DNA molecule depends on the base sequence in that region. That is, the local shape of a DNA molecule is sequence-dependent. This is of central importance, for example, in the way in which DNA-binding proteins recognize specific base sequences. If all DNA molecules had the idealized structure of the B-form, then the possibilities of designing proteins which recognize particular base sequences would be limited. Given, however, that the shape depends on the base sequence, then there is much more scope for Nature to design a protein that can bind to a particular region of a DNA molecule.

3.3 The Stability of the Double Helix

If solutions of double stranded DNA are heated, the chains separate and become random coils. This referred to as **denaturation**. It is easy to monitor denaturation of DNA by measuring changes in absorption of light at 260 nm, the wavelength at which the bases absorb maximally. On denaturation, the absorption of a solution of DNA increases by 20–30%; this is called the **hyperchromic** effect. The effect has its origin in the stacking of the bases in the double helical structure. The absorbance maximum at 260 nm originates from the π-electron system of the rings. When the bases are stacked, interaction between the π-electrons of the bases leads to a coupling of their transitions, and decreases the absorption. When the bases are un-stacked, this effect is removed.

The type of denaturation curve obtained is shown in Figure 3.13. The temperature at which the increase in absorbance reaches 50% of its maximum is called the **melting temperature** (T_m). For the two cases shown in Figure 3.13, the melting temperatures are about 78 °C and 89 °C.

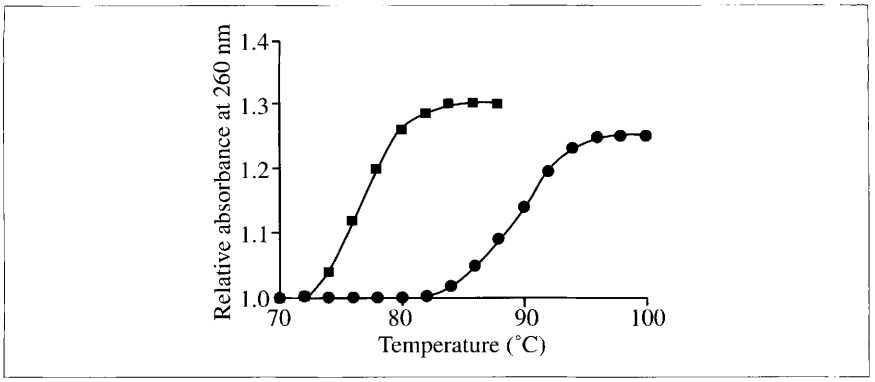


Figure 3.13 Heat denaturation curves for two DNA samples with different base compositions

One of the characteristics usually quoted for the DNA from any source is the percentage G + C. This varies over wide range from as low as 20% to as high as 80%.

The T_m value depends on the nature of the solvent and is lower at lower ionic strength. If, however, the solvent conditions are kept constant, then the most important factor governing T_m is the base composition of the DNA. The higher the content of G and C, the higher is the melting temperature. This is as expected, because the G/C base pair makes three hydrogen bonds, whereas the A/T pair makes only two. Indeed, there is an empirical quantitative relationship between the percentage G+C content and T_m as shown in equation (3.1):

$$\%(G + C) = 2.44(T_m - K) \qquad (3.1)$$

where K is a constant which depends on the solvent. Once it is known, equation (3.1) can be used to determine the %(G+C) for a newly isolated DNA molecule.

Worked Problem 3.3

Q The DNA with a melting temperature of 78 °C in Figure 3.13 has a (G+C) content of 25%. What is the %(G+C) of the molecule with $T_m = 89$ °C?

A We must first obtain the value of K in equation (3.1) by substituting in the T_m of the DNA with known %(G+C). The value obtained is $K = 68$ °C. So for the DNA with $T_m = 89$ °C:

$$\%(G + C) = 2.44(89 - 68) = 51$$

In principle, if a solution of DNA which has been heated beyond T_m is returned to a lower temperature, the double helix can reform; this process is called **renaturation** or **annealing**. It is usually very rapid for small DNA molecules, but can be very slow for large ones. The reason for

this is that with large molecules there is considerable opportunity for improper base pairing to occur between regions of similar base sequence, and these incorrect base pairings must be broken before proper annealing can occur. The optimum conditions for renaturation are to keep the temperature just below the T_m, so that incorrect base pairs can rapidly dissociate and allow the correct ones to reform. Under these conditions it is found that the time for renaturation depends on the complexity of the DNA sample. This is expressed quantitatively as the **Cot value** (more properly the $C_0t_{\frac{1}{2}}$ value), which is the product of the initial concentration of denatured DNA expressed in terms the concentration of nucleotides (C_0) and the time in seconds for 50% of the sample to renature ($t_{\frac{1}{2}}$). This type of analysis is important, for example, in distinguishing the situation where a piece of DNA contains multiple copies of the same gene. Because of its lower complexity it will have a Cot value lower than that of a piece of DNA of similar size but without repetitive elements.

We will return to the subject of melting and re-annealing DNA *in* Section 5.7, when we deal with a technique called the polymerase chain reaction.

3.4 Nucleosomes and Chromosomes

DNA molecules are enormously long compared with the size of a typical cell (see Worked Problem 3.4) and so must be folded up in some way in order to fit in. In addition, long, thin DNA molecules would be very susceptible to mechanical breakage if they were not protected in some way. Indeed, it is very difficult to isolate intact DNA molecules from cells because they are so easily broken. For these reasons it is not surprising that DNA is found in cells packaged up with proteins into compact structures. This packaged DNA is called **chromatin**.

Worked Problem 3.4

Q Assuming that the DNA from the largest chromosome of the mouse is entirely in the B-form, calculate its length when fully extended.

A In Section 2.6, the DNA in the largest mouse chromosome was stated to contain 192,000,000 residues; that is, it is 192 Mbp long. In the B-form of DNA the vertical rise for each base pair is 0.34 nm (Figure 3.1). Hence the length of the molecule if fully extended would be $192 \times 10^6 \times 0.34 = 6.53 \times 10^7$ nm. This is the same as 0.0653 m or 6.53 cm. This is enormous when compared with the size of a cell, which typically has a diameter of 10 μm.

Sizes of fragments of DNA are determined from the rates at which they migrate in a particular form of **electrophoresis**. This method is described in detail in Section 5.2.

Electron microscopy is a form of microscopy where beams of electrons are used to visualize objects rather that the light beams used in conventional microscopy. The electron beams are focused using magnets rather than glass lenses. The resolution of a microscope (that is, the size of the smallest objects that can be seen) is governed by the wavelength of the radiation used. The wavelength associated with a beam of electrons depends on the speed with which they are travelling and can be made very small indeed. For technical reasons, however, the resolving power of the electron microscope is limited to about 0.5 nm. This is not small enough to visualize individual atoms, but is sufficient to see the outline structures of large objects such as nucleosomes.

The basic unit of packaging is the **nucleosome**. Nucleosomes were discovered in 1973 by Hewish and Burgoyne.[16] What they did was to digest chromatin with an enzyme (**deoxyribonuclease**) that hydrolyses DNA, and then to measure the sizes of the fragments that they obtained. The fragments were found to be 200 bp long, or a small multiple of that figure. Hewish and Burgoyne argued that this was because the DNA in chromatin is packaged in units of 200 bp, each of these units being associated with a protein complex. The deoxyribonuclease was able to hydrolyse the DNA in between these units, but the DNA in direct contact with the protein complex was protected from hydrolysis. Supporting evidence for this conclusion was obtained from studies on chromatin using **electron microscopy**. The chromatin had the appearance of a string of beads, the string being the DNA and the beads the protein complexes. These complexes of protein and DNA, or nucleosomes as they were called, are shown diagrammatically in Figure 3.14.

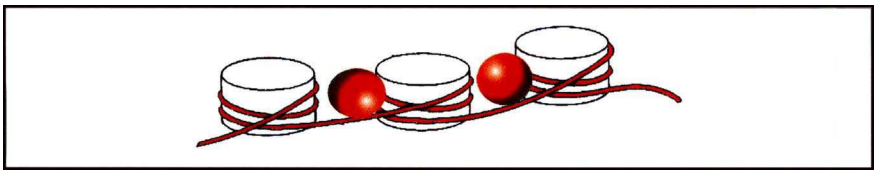


Figure 3.14 Nucleosomes. The DNA double helix is shown as a red ribbon. The cylindrical objects are complexes of two copies each of histones H2A, H2B, H3 and H4. The red spheres are molecules of histone H1

The protein components of the nucleosomes have been isolated and studied in detail. The core of the particle consists of two copies each of four proteins called **histones**. The four proteins in question are called histone H2A, H2B, H3 and H4. If these proteins are mixed in solution, then two copies of each combine to give a stable complex called the **histone octamer**. This is the structure represented by the cylinders in Figure 3.14. The DNA makes just less than two complete loops around the octamer spool. A complete loop contains eight turns of the double helix, and in total about 150 bp of the DNA is in contact with the protein spool. This leaves about 50 bp in the linking region between the octamers. This stretch of DNA can bind a ninth protein, called histone H1, shown in Figure 3.14 as a red sphere. Whether or not H1 is bound in a particular preparation of nucleosomes depends on the solution conditions and in particular on the ionic strength. High salt concentrations cause the H1 to dissociate.

One further point needs to be made about the nucleosome. The usual picture that one has of DNA is as a straight rigid cylinder, but this view has to be modified when we consider the structure of the nucleosome. The DNA is wound around the protein spool and for every base pair in the

sequence on average the helix must bend by about 4.5° (there are 80 base pairs in a complete turn, and the angle passed through is 360°). The ease with which local areas of the double helix can distort from the ideal structure depends on the base sequences in those areas. Equally, then, different regions of DNA will bind to the octamer core with different affinities; the easier it is to bend the DNA, the tighter it will bind and *vice versa*. So, for example, runs of A/T or G/C pairs which render the DNA more rigid, weaken the binding. These effects are important because for the DNA to be transcribed into RNA it must detach temporarily from the protein core. This process will start at points where the binding is weakest. Consistently, the start points of genes often contain sequences that facilitate detachment from the nucleosome.

An enormous increase in knowledge of the structure of the nucleosome occurred in 1997 as a result of the determination of the structure of the **core particle** by X-ray diffraction methods.[17] In this context, core particle means the protein octamer and the DNA that is in contact with it. We cannot go into the details of the structure here, but if you wish to look at it the structure can be found as entry 1AOI in the PDB.

Let us return to the problem posed at the beginning of this section, that is, how the DNA is packaged up into a compact structure that will fit into a cell. Formation of nucleosomes is part of the answer, but not all of it (see Problem 3.4). The poly-nucleosome fibre itself wraps up into a spiral with, on average, six nucleosomes per turn. This structure has a diameter of about 30 nm and is known as the **30 nm fibre**. It is thought the histone H1 plays a part in stabilizing its structure. There are, however, several other proteins associated with the 30 nm fibre which may play a part in its formation. The main ones belong to a group called the **high-mobility group proteins** (or HMG proteins). The name arises from the fact that the proteins migrate rapidly when subject to electrophoresis in an acidic buffer. They are present at relatively low levels (1–10% of the amount of histones) and their precise functions are still unclear.

One turn of the 30 nm fibre has a length of about 11 nm. It contains approximately 1200 bp of DNA (six nucleosomes each with 200 bp), which in the B-form would have a length of a little over 400 nm. So the net reduction in length resulting from formation of nucleosomes and then packing them into 30 nm fibres is about 40-fold.

There must be yet more compaction of the structure, but the details of this are still far from clear. What probably happens is that the 30 nm fibre forms loops with the ends anchored onto some sort of protein scaffolding, as shown schematically in Figure 3.15. It is thought that the loops contain an average of about 50 turns of the 30 nm fibre (the turns are represented by the little cells in Figure 3.15), so the length of each loop is about 250 nm or 0.25 μm. The total DNA in each loop is of the order of 60,000 bp, and the overall length of the DNA has been reduced by roughly 1000-fold. It is probably in this form that the DNA exists in the chromosomes of cells that are not dividing. The process by which the DNA in this form compacts by a factor of at least 10 again to form chromosomes that are clearly visible in the light microscope during cell division is unknown.

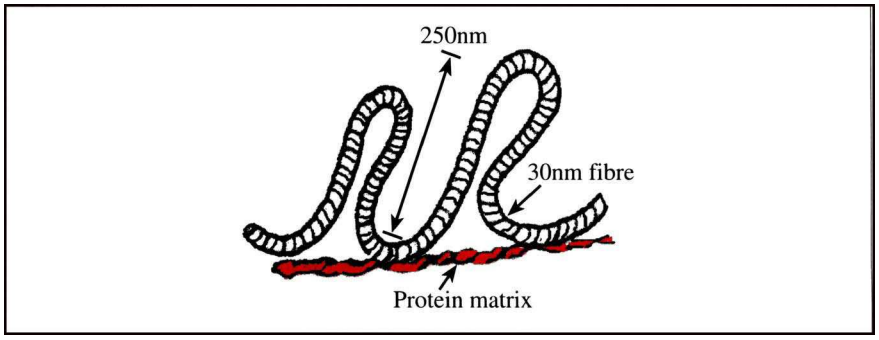


Figure 3.15 Possible arrangement of the 30 nm fibres into loop structures in chromosomes. Each segment represents a turn of the 30 nm fibre

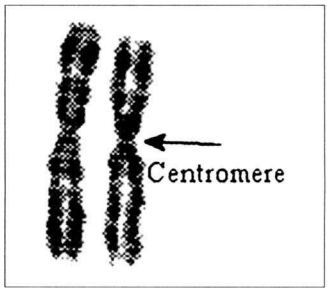


Figure 3.16 Human chromosome 3

Box 3.6 How the Packaging of DNA Varies During the Cell Cycle

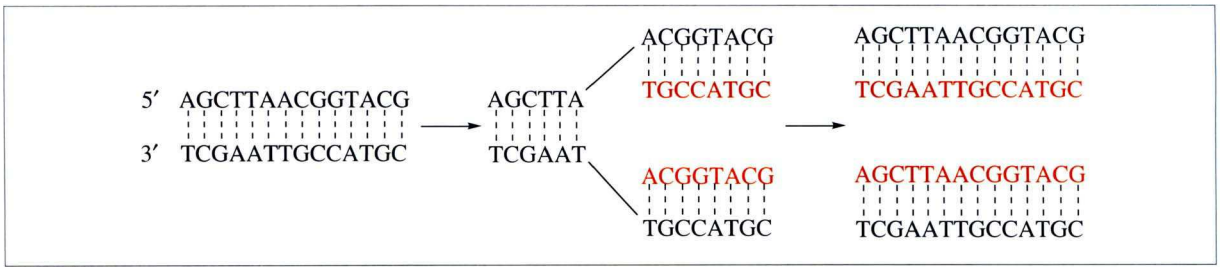
Recall that cells (except the gametes) contain two copies of each chromosome, one inherited from the mother and the other from the father. Cells that are in the period between divisions are said to be in **interphase**. In this part of the cycle the chromosomes are diffuse and not visible under the microscope; it is likely that the DNA is in the form of 30 nm fibres, looped as shown in Figure 3.15, and spread throughout the nucleus. At the end of interphase, the genetic material is replicated so that there are now four copies of the DNA from each chromosome; that is, replicate pairs of the maternal DNA and replicate pairs of the paternal DNA. These newly formed species are referred to as **chromatids**. In **prophase** the chromatids become much more compact and pair up into visible chromosomes. There will be two copies of each chromosome, one containing two chromatids derived from maternal DNA, and the other containing two chromatids derived from paternal DNA. For example, Figure 3.16 shows the appearance of one of the pairs of human chromosomes. Each pair of chromosomes has a somewhat different appearance, but they share the common feature of a region (the **centromere**) where the chromatids are closely associated together. Those in Figure 3.16 are the two copies of chromosome 3. They are about 10 μm long and easily visible in the light microscope.

In **prometaphase** the nuclear membrane breaks down, and then in **metaphase** the chromosomes, or paired chromatids, become aligned in a plane along the equator of the cell. After this, in **anaphase**, the centromeres become attached to **microtubules** (protein fibres) which pull the chromosomes apart. The two chromatids from each of the

two members of a pair of chromosomes move to opposite poles of the cell. So at this stage the genetic material has been copied, and one set of chromosomes is located at one side of the cell, and an identical set at the other. Finally, in **telophase**, new nuclear membranes are formed around the two new sets of chromosomes and cell division occurs. The chromatin then becomes diffuse again, and the cell returns to interphase.

3.5 DNA Replication

At the end of Section 3.1.1, a quotation was given from Watson and Crick's paper on the structure of DNA which indicated that they realized how their proposed structure could explain the mechanism of replication of the genetic information. Very shortly afterward, they published a second paper in which the theory was spelled out.[18] The essence of it is that the strands of the double helix separate, and that each is used as a template for the synthesis of a new chain. This is shown schematically in Scheme 3.1.



Scheme 3.1

On the left is a small section of a double stranded DNA molecule (remember that DNA is helical, but it is unnecessary to show it as a helix for the present purposes). The chains separate, and on each of them a new complementary chain is synthesized. In the middle part of the scheme, the first eight bases have been copied and the new partial chains are shown in red. At the right of the scheme, copying is complete. What we now have is two molecules of DNA, both identical to the original molecule. Each contains one original chain, shown in black, and one new chain, shown in red. Because the new molecules contain one old and one new chain, this process is referred to as **semi-conservative replication**.

Box 3.7 The Meselson–Stahl Experiment

Semi-conservative replication of DNA as shown in Scheme 3.1 is not the only way it could happen. It is in principle possible that replication could be conservative; that is, the original molecule of DNA might not come apart but simply serve as a template to produce another, entirely new, DNA molecule. In this scheme, the product would be the original molecule plus one that was entirely new. That this is not what happens was shown in a classical experiment done by Meselson and Stahl.[19]

They made use of the technique of **density gradient centrifugation**. In this method, a centrifuge tube is filled with a solution, the density of which increases from top to bottom of the tube. If a sample containing compounds of different densities is applied to the top of the tube and then centrifugation is carried out for a prolonged period, when equilibrium is reached each component in the applied mixture will have come to rest at a point in the tube where its density is the same as that of the solution. The densities of DNA molecules are about 1.7 g cm^{-3}, and so solutions of high density are required; Meselson and Stahl used gradients of CsCl to obtain the required densities. The centrifugation was carried out at 140,000 × *g* for 20 hours.

What samples did they analyse? First, they grew cells of the bacterium *Escherichia coli* in a medium containing $^{15}NH_4Cl$ for many generations, so that the cellular components, including the DNA, were essentially completely labelled with ^{15}N. They then transferred the cells to a new medium containing $^{14}NH_4Cl$ and took samples after one, two and three generations, and so on. Each sample of cells was broken open and the DNA subjected to density gradient centrifugation. The results that they obtained are shown in Figure 3.17.

With the cells that had been grown for several generations in $^{15}NH_4Cl$ (generation 0), a single band of DNA was observed. This corresponded to DNA with all its nitrogen atoms in the heavy isotopic form. The first generation of cells after transfer to $^{14}NH_4Cl$ produced a single band of DNA but at a lower density. In the second generation, two types of DNA were produced in equal amounts. One band of DNA was found at the same density as that from generation 1, but the other was less dense again. The interpretation of these results was that in generation 1, the single type of DNA consisted of molecules where one strand of the double helix contained ^{14}N and the other contained ^{15}N. In the second generation the ^{15}N-containing strand had produced a $^{15}N/^{14}N$ hybrid, but the ^{14}N-containg strand had produced a $^{14}N/^{14}N$ molecule.

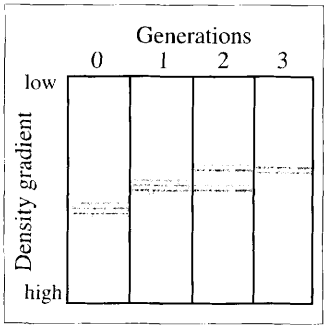


Figure 3.17 Results from the Meselson–Stahl experiment. Each column shows the position of the band, or bands, of DNA after density gradient centrifugation. The intensity of the bands is proportional to the amount of DNA present in each

Note that the charged states of the molecules in Schemes 3.2 and 3.3 are not accurately represented. For example, the charge on dGTP at neutral pH will be about −2.5 rather than −3 as shown. Similarly, inorganic phosphate will have a

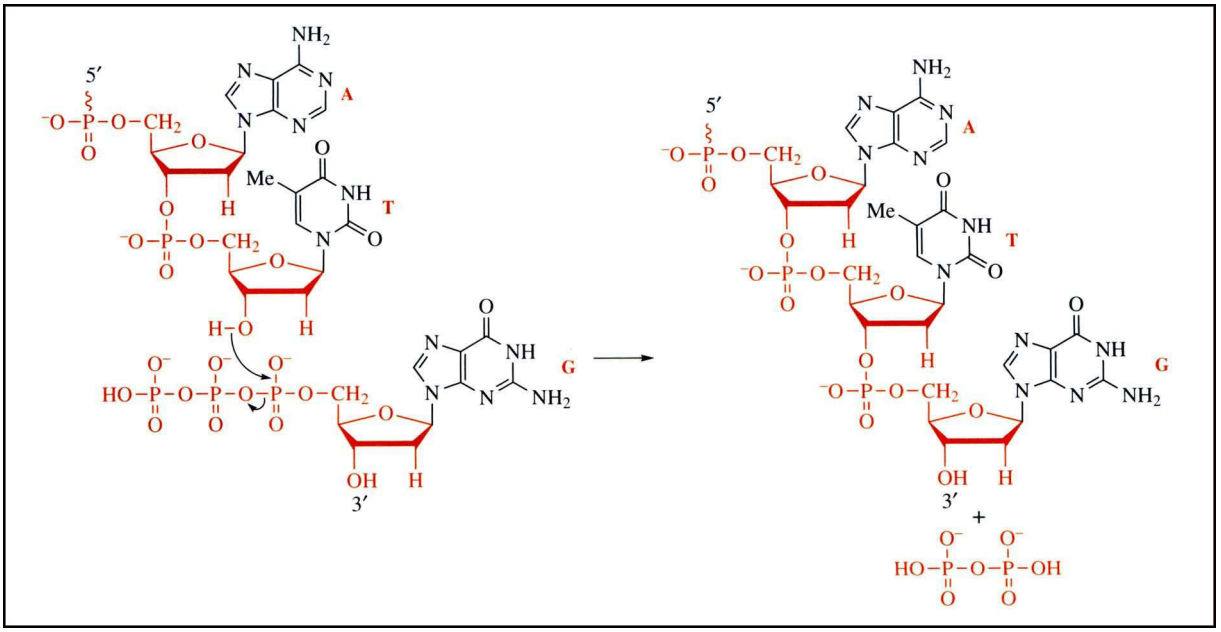
These results were only consistent with semi-conservative replication (see Problem 3.5).

The chemistry of chain elongation is shown in Scheme 3.2. The incoming residue is added to the 3′-end of the growing polynucleotide. In Scheme 3.2, a G residue is to be added (that is, there would be a C at this position in the chain being copied). The residue is donated by deoxyguanosine 5′-triphosphate (dGTP), and the reaction involves nucleophilic attack of the 3′-OH of the last residue in the chain on the α-phosphate of the dGTP, catalysed by the enzyme **DNA polymerase**. The result is formation of a new phosphodiester linkage and the liberation of pyrophosphate. It is important that the reaction goes to completion, and this is ensured by the presence of the enzyme **pyrophosphatase** which hydrolyses the pyrophosphate to two molecules of phosphate (Scheme 3.3) and in so doing pulls the equilibrium of Scheme 3.2 to the right.

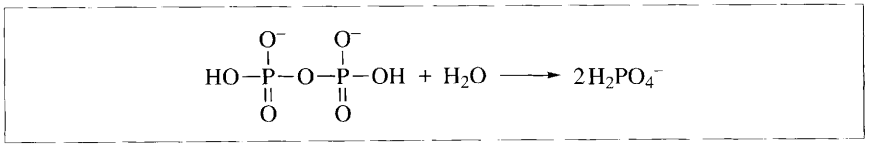
charge of about −1.5 (the pK_a values for the first two ionizations of phosphoric acid are about 2.2 and 7.2). For our present purposes the precise charge state is unimportant.

DNA polymerase was discovered by Arthur Kornberg. He shared the Nobel Prize in Medicine in 1959 with Severo Ochoa "for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid".

It is now known that most organisms have three different DNA polymerases. The one involved in DNA replication is called **DNA polymerase III**. The other two are involved in DNA repair (see Section 3.6.2).



Scheme 3.2



Scheme 3.3

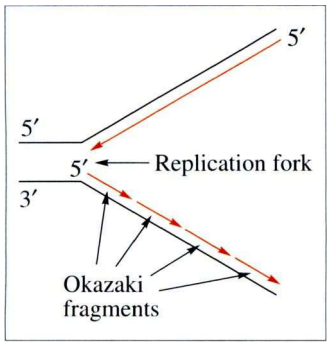
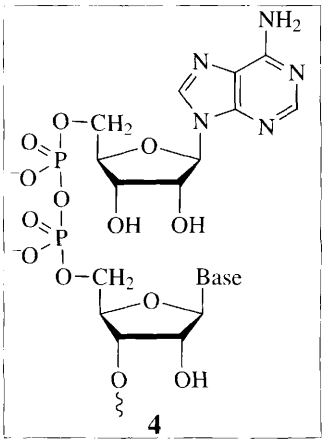


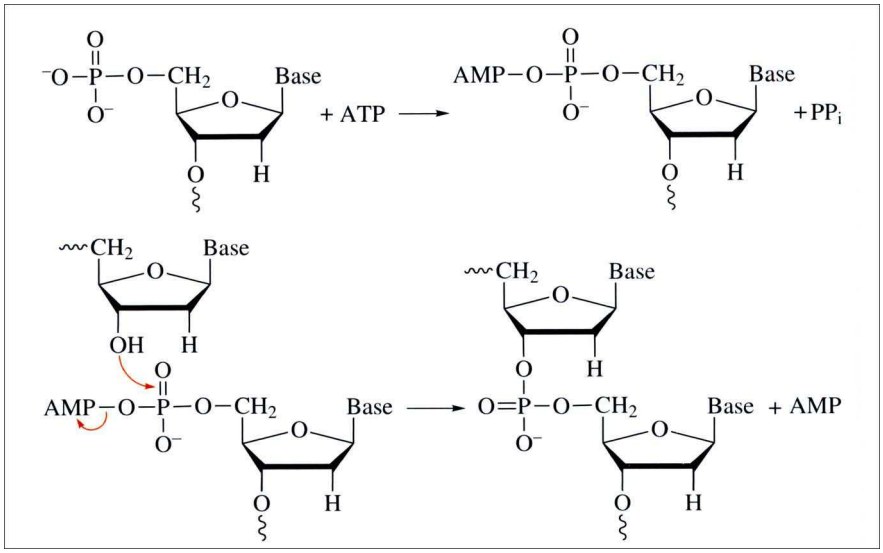
Figure 3.18 Replication of DNA. The 5′→3′ strand is synthesized continuously, whereas the 3′→5′ strand is synthesized as a set of fragments that are subsequently joined together



4

A little thought will show that Scheme 3.2 cannot be the whole story. DNA polymerase catalyses the addition of a nucleotide onto the 3′-end of a growing chain. This is fine for duplication of the upper chain in Scheme 3.1; duplication starts at the 3′-end and goes 5′→3′ for the new chain as shown in Scheme 3.2. However, what about the other chain? It is shown being duplicated 3′→5′, and DNA polymerase cannot catalyse that reaction. The answer to this problem was provided by Reiji Okazaki and his colleagues.[20] What they found was that, during replication, fragments of DNA about a thousand nucleotides long were also formed; these are now known as **Okazaki fragments**. It was proposed that one chain of DNA is synthesized continuously from 5′→3′ (this is called the **leading strand**), whereas the other (the **lagging strand**) is synthesized in fragments, again 5′→3′, and that these fragments are subsequently joined together to form a continuous chain. The process is summarized in Figure 3.18.

Support for this mechanism for synthesis of the lagging strand came from the fact that an enzyme was already known that could join the two ends of a break in a DNA chain. The enzyme is called **DNA ligase**, and it catalyses the reaction shown in Scheme 3.4. First, the residue at the 5′-side of the break is converted to an **adenylate** in a reaction involving ATP; again, pyrophosphate is a product of this reaction and its hydrolysis pulls the reaction to the right. The full structure of the adenylate derivative is given in **4**. The pyrophosphate linkage in this derivative is very reactive, and reacts readily with the 3′-OH group on the other side of the break to yield the desired product. This is another example of the way in which ATP participates in synthetic reactions (compare this with the situation where ATP acts as a phosphorylating agent as described in Box 2.5).



Scheme 3.4

This is by no means the end of the complexities of DNA replication. DNA polymerase is very efficient at adding residues to a growing chain – the enzyme from *E. coli*, for example, adds about 2000 bases per second under optimal conditions – but it cannot start a new chain from scratch. It requires a **primer**; that is, a small polynucleotide to get it going. How is this problem overcome? It turns out that DNA replication is initiated, not by DNA polymerase, but by a special **DNA-dependent RNA polymerase** called **primase**. This enzyme does not require a primer. It synthesizes a short stretch of complementary RNA (about five residues long) at the end of the DNA to be duplicated, after which DNA polymerase takes over and uses the short section of RNA as a primer to duplicate the rest of the DNA molecule. At some stage, the temporary piece of RNA is removed by a 5′→3′ **exonuclease** and the missing residues are replaced by deoxyribonucleotides. This 5′→3′ exonuclease activity is a second activity of the DNA polymerase.

The "exo" part of the name signifies that it hydrolyses residues from the end of the polynucleotide chain – in this case, from the 5′ end. An **endonuclease**, on the other hand, hydrolyses the inter-nucleotide linkage in the middle of a polynucleotide.

Why is this complication necessary? The reason is that DNA polymerase has a **proofreading** function. The specificity of DNA synthesis is determined in the first instance by the need for the incoming base to form the correct Watson–Crick base pair with the template strand. It can be estimated that, relying on this specificity alone, an error rate of about 1 incorrect base in 10,000 would be expected. This is wholly inadequate for the accurate transmission of the genetic information. The error rate is decreased by the fact that the DNA polymerase has an active site that "senses" the shape of the forming base pair and rejects it if it is incorrect. This increases fidelity of copying by perhaps another one 1000-fold. Finally, if an incorrect residue has been incorporated, the DNA polymerase has a method of removing it again before synthesis proceeds. The way this works is that once a nucleotide has been added to the growing chain, the enzyme moves one base along the template chain and the newly formed base pair moves into a site which will only accommodate correct base pairs. If an incorrect base has been incorporated, the enzyme stalls, giving time for the incorrectly paired residue to dissociate from the template strand. When it does so, it encounters yet another catalytic site in the DNA polymerase that has **3′-exonuclease** activity, and the residue is removed by hydrolysis. The correct base can now be added. This proofreading increases accuracy of replication by a further1000-fold and reduces the final error rate of replication to about 1 in 10^{10} residues. Hence a crucial part of the function of DNA polymerase involves testing the preceding base pair for correctness before adding a new nucleotide. At the beginning of synthesis there is no preceding base pair and so the enzyme cannot function. RNA polymerase does not have this proofreading property, and can initiate synthesis without a primer. It does not matter if incorrect residues are incorporated by the RNA polymerase, because they are going to be removed anyway.

DNA polymerase is, of course, enormously important, because it is the key to the accurate transmission of genetic information. It is also very complicated and has three distinct catalytic activities: it is a polymerase, a 5′→3′ exonuclease, and a 3′→5′ exonuclease. The structure of the enzyme has been determined by X-ray diffraction methods and much is now known about its mode of action. To go into this would take us too far from the main theme, but for interested readers the review by Hubscher *et al.*[21] is recommended.

There is one aspect of DNA replication that we have yet to consider and which we will deal with only very briefly. As we know, DNA has a double helical structure. For replication, the helix has to be unwound. Again, there is an enzyme that does this job; the enzyme is called **helicase**. The helicase binds to a region of the DNA referred to as the **origin of replication** and unwinds the helix in that region. Origins of replication are rich in A/T base pairs, and so are regions where the stability of the double helix is lower than average. The helicase then moves along the DNA molecule, in front of the DNA polymerase, unwinding it as it goes to create a **replication fork** (see Figure 3.18). To ensure that the DNA stays unwound long enough for the polymerase to do its job, **single strand binding proteins** attach themselves to the DNA strands and prevent them re-combining. The single strand binding proteins are displaced as the polymerase reaches them. Unwinding of the helix introduces another problem. This is that, as it is unwound, the remaining double helical part of the DNA becomes super-coiled (this is a bit like what happens to the cord on a telephone when the receiver is repeatedly picked up, turned around, and replaced). Again, there are enzymes, called **topoisomerases**, to deal with this difficulty. Topoisomerases catalyse a cut in one strand of the DNA, pass a section of the DNA through this break to relieve the super-coiling, and then join the two ends up again. A review of topoisomerases and how they work has been written by Wang.[22]

Box 3.8 Genetic Recombination During Meiosis

As we have seen, the replication machinery is designed to minimize mistakes in copying DNA and so the question might occur to you as to how **genetic diversity** arises. Why are two daughters of the same parents not genetically identical (unless they are monozygotic twins)? A large part of the answer lies in a process called **crossing over** which results in **genetic recombination**. Genetic recombination during meiosis is illustrated in Figure 3.19.

Cells which contain the full complement of genetic information (one set of chromosomes from the mother, and one from the father) are said to be **diploid** and the process of cell division which produces such cells is called **mitosis**. The gametes, or germ cells, are **haploid**; that is, they contain only one homologue from each pair of chromosomes. Gametes are formed by a process called **meiosis**. In meiosis, two successive cell divisions occur. Replication of the genetic information occurs in the first division but not in the second, so that the amount of genetic material is halved.

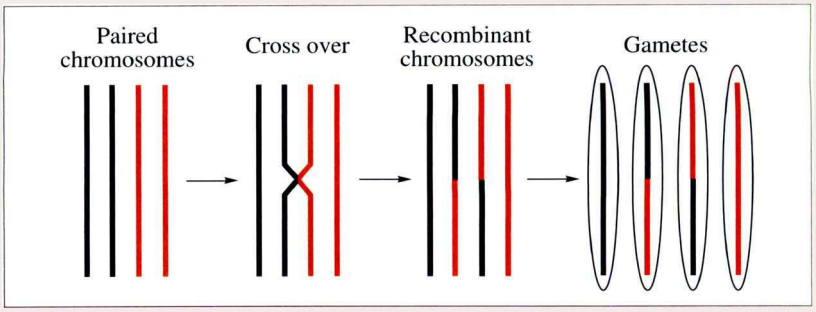


Figure 3.19 Recombination (crossing over) of genetic material during meiosis

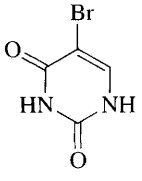
The first part of the figure shows paired homologous chromosomes resulting from replication of the genetic material (see Box 3.6) in the first stage of meiosis; to distinguish between them, one pair is shown red and the other black. A feature of meiosis is that the homologous chromosomes pair up closely along their entire lengths, and at various points they form strong attachments called **chiasmata** (from the Greek *chiasma*, meaning cross). One such chiasma is shown in the second part of Figure 3.19. It is at the chiasmata that crossing over occurs. What happens is that nicks are made in one of each homologous pair, and the chains are then rejoined but in such a way that sections of the chromatids are exchanged. This leads to recombinant molecules in which one part arises from one chromosome and the remainder form the other (shown in Figure 3.19 as chromosomes in which one part is red and the other black). In the second part of meiosis, there is no duplication of the genetic material; rather, the existing chromosomes are partitioned between the new cells so that each of them (the gametes) contain only a single copy of the genetic material. These gametes will be of four different types, one containing the black chromosome, one the red chromosome, and the other two different forms of the recombinant chromosome. Offspring produced from these gametes would then inherit different combinations of paternal and maternal genetic information.

Crossing over can occur at a large number of places in the chromosome, and is a very frequent event. Not surprisingly, therefore, the genetic make up of the gametes in any individual varies enormously, with consequent genetic and phenotypic variation in the offspring.

3.6 DNA Damage and Repair

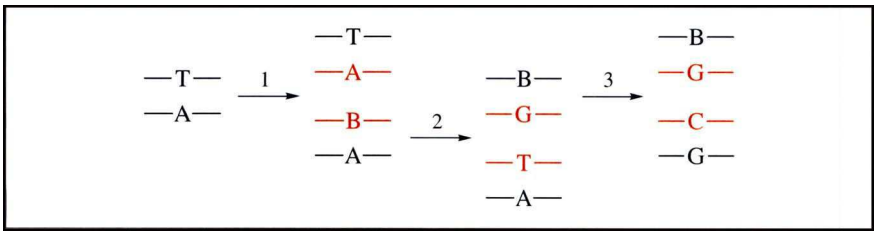
3.6.1 Damage

DNA is subject to a variety of external dangers, both chemical and physical, which can have severe consequences for its structure and hence for the integrity of the genetic information. Changes brought about by these agents are called **mutations**, and chemical agents that lead to mutations are referred to as **mutagens**. Some mutagens lead to substitution of one base for another. An example is 5-bromouracil (**5**). This is an analogue of thymine and can become incorporated into 5-bromodexoyuridine, and hence into DNA in place of thymine. If this

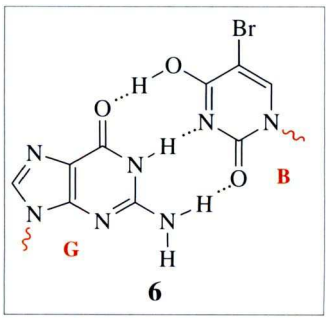


5

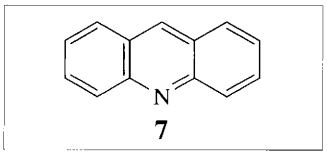
happens, then the consequence may be incorporation of G in the complementary strand in place of A. The way that this can happen is shown diagrammatically in Scheme 3.5. On the left of the scheme, a section of DNA is represented simply as a pair of lines with an A/T base pair shown. In step 1, the DNA is replicated and a 5-bromouracil (abbreviated as B) is incorporated into one strand in place of thymine; old strands are shown in black and new ones in red. In step 2, the 5-bromouracil directs incorporation of a guanine in the new strand. Finally, in step 3, correct base-pairing occurs and the G directs incorporation of a C in the new chain. The net result is a change from an A/T base pair to a G/C pair.



Scheme 3.5



6



7

The reason why B base-pairs efficiently with G seems to be that the electronegativity of the bromine substituent favours formation of the *enol* tautomer of 5-bromouracil. This can form hydrogen bonds with guanine as shown in **6**.

Another effective mutagen is nitrous acid (HNO_2). The reason is that nitrous acid oxidatively deaminates NH_2-containing bases, and converts them to the corresponding carbonyl derivatives. Adenine is converted to hypoxanthine, guanine to xanthine, and cytosine to uracil. Xanthine, like guanine, base pairs with cytosine so there is no effect as a result of this change. On the other hand, hypoxanthine pairs with cytosine, and this leads to a mutation from A/T to G/C (see Problem 3.6).

A different type of mutation is frequently caused by polycyclic aromatic substances such as derivatives of acridine (**7**). These planar molecules can intercalate between the base pairs in the DNA double helix. An example is shown in Figure 3.20 (PDB entry 1G3X[23]). This shows a dodecamer of DNA with a derivative of acridine intercalated into it (the acridine has a tail of four arginine residues attached to it, but that is unimportant for the present purposes). The acridine ring, viewed edge-on, sits between a pair of adenines in one chain, and a pair of thymines in the other. Inserting the acridine ring obviously results in a considerable distortion of the backbones in its immediate vicinity.

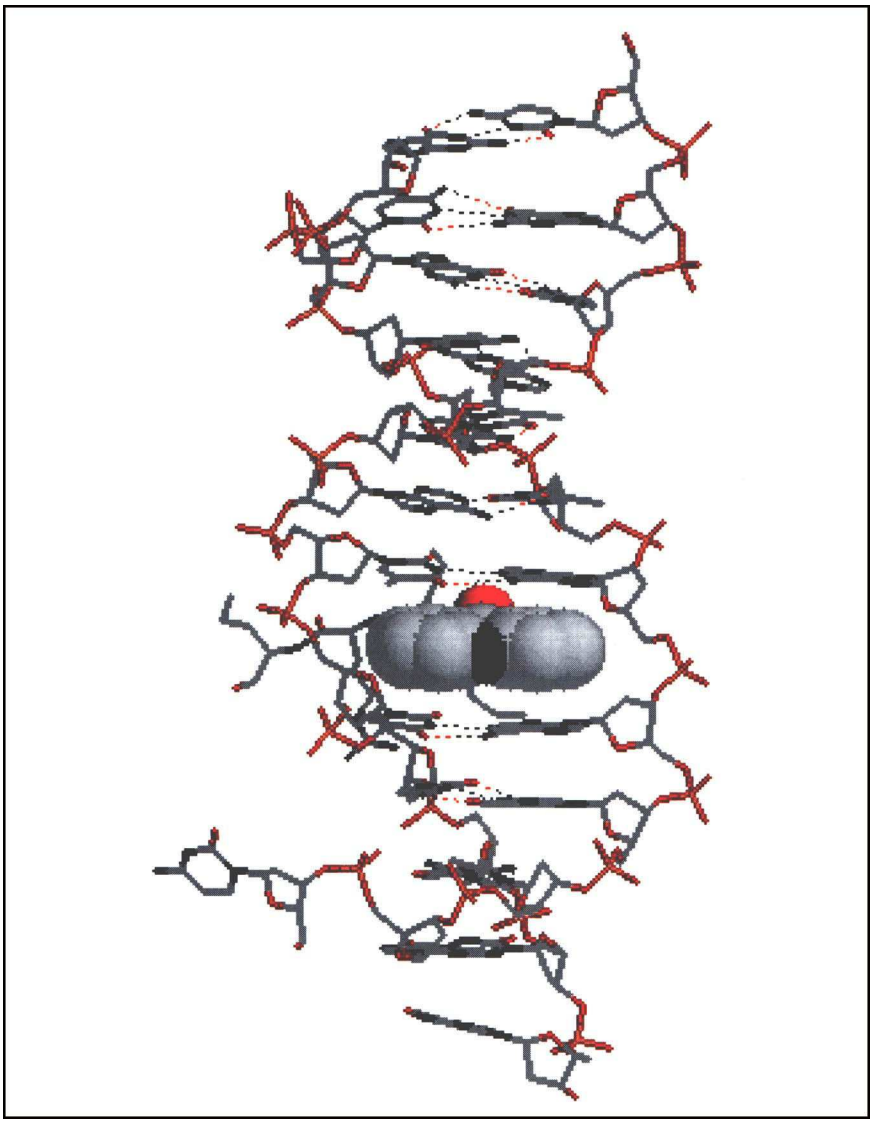


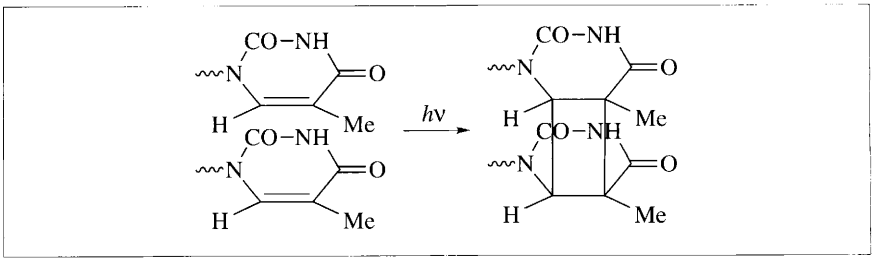
Figure 3.20 An acridine derivative intercalated into DNA

The effect of intercalation of a molecule like acridine on DNA replication will often be the insertion of one, or even more, extra residues in the chain. In protein-coding genes this gives rise to **frame shift mutations** (see Section 1.6). If one or two extra bases are incorporated, then all of the codons after the point of insertion will be changed, with the likelihood that a stop codon will be met at some point, causing premature termination. If three bases are inserted, then this will result in insertion of an extra amino acid in the product protein, but the rest of the protein will be normal. This may or may not affect the activity of the protein product, depending on exactly where the extra amino acid is inserted. Other agents

Testicular cancer caused by exposure to soot was once a common occupational disease experienced by chimney sweeps.

that intercalate in this way include carcinogenic polycyclic compounds found in soot such as anthracenes and benzpyrenes, and a variety of naturally occurring toxins such as the liver carcinogen aflatoxin. Aflatoxin is produced by a mould that grows on peanuts.

A different sort of damage to DNA is caused by ultraviolet radiation. The major effect of this is to cause dimerization of adjacent thymine residues to yield a cyclobutane product. This is shown in Scheme 3.6. The "squiggly bonds" shows the points of attachment of the thymines to the sugar–phosphate backbone. The result of this is to block the action of DNA polymerase, and so prevent replication of the DNA.



Scheme 3.6

3.6.2 Repair

It is not surprising, given the importance of the integrity of the genetic information, that Nature has developed ways of dealing with these various forms of damage to the DNA. To take the last-mentioned case first, just as UV light causes the problem, it also provides a solution. This is in the form of a light-dependent enzyme called **DNA photolyase**. This enzyme catalyses the reversal of the dimerization reaction, and reforms the two thymine residues. There is evidence that DNA photolyase is missing in some people suffering from the rare skin disease *Xeroderma pigmentosa,* which is characterized by extreme sensitivity to sunlight. The disease is inherited recessively (that is, defective genes must be inherited from both parents). Suffers usually develop skin cancers at sites exposed to the sun, and rarely survive beyond the age of 30.

The most general method for correcting errors in DNA is **excision repair**. As the name implies, the incorrect or defective region of the DNA is cut out, and the damage is then repaired. One example of this process involves a second way of dealing with thymine dimers. A protein specialized for the task detects the distortion of the double helix caused by dimer formation. An endonuclease called **excinuclease** then cuts the DNA strand containing the dimer at a point eight nucleotides away on the 5′-side and four nucleotides away on the 3′-side. The liberated 12-mer dissociates from the rest of the molecule and the gap is filled in with the correct nucleotides. The enzyme responsible for doing this is DNA

polymerase I. It uses the 3′-OH group at the gap as a primer. Finally, the newly inserted piece is joined to the 5′-end of the existing chain by DNA ligase. It is likely that deficiency in excinuclease is another cause of *Xeroderma pigmentosa*.

Another type of excision repair involves removal and replacement of only a single residue, and usually operates where a base has been modified, for example by alkyation. What happens here is that the enzyme involved binds to the DNA at the damaged base and causes it to flip out of the helix. The enzyme then acts as a **glycosidase** and hydrolyses the bond between the base and C-1 of deoxyribose. The product is referred to as apurinic or apyrimidinic DNA (or **AP DNA** for short). There are many different glycosidases involved, each specific for a different type of modified base. The AP site is recognized by a specific endonuclease which cleaves the DNA chain adjacent to the AP site and the remaining deoxyribose phosphate is cut out. The gap is then filled by DNA polymerase I and the chain re-sealed by DNA ligase.

Alkyating agents are often powerful mutagens and carcinogens. Typical examples are sulfur mustard [bis(2-chlororethyl) sulfide, **8**, and nitrogen mustard [*N*,*N*-bis(2-chloroethyl)-*N*-methylamine, **9**]. They exert their effects by mainly by alkylating guanine at N-7.

$ClCH_2CH_2-S-CH_2CH_2Cl$

8

$ClCH_2CH_2-N(Me)-CH_2CH_2Cl$

9

Summary of Key Points

1. DNA is a double helical molecule with two polynucleotide chains running in opposite directions, wrapped around a common axis. The sugar–phosphate backbones are on the outside of the structure, and the bases are on the inside. Each base from one chain is hydrogen bonded to a base from the opposite chain. The planes of the base pairs are perpendicular to the helix axis, and the base pairs are stacked on top of one another. In the B-form of DNA there are 10 base pairs for each complete turn of the helix.

2. Wherever an adenine occurs in one chain of the double helix, thymine occurs in the other. Wherever guanine occurs in one chain, cytosine occurs in the other. This specific base pairing forms the basis of the process by which DNA is replicated.

3. DNA forms a double stranded structure as a way of removing the hydrophobic bases from contact with the surrounding water. The double strand adopts a helical twist so as to minimize the distance between the base pairs.

4. When solutions of DNA are heated, the two chains of the molecule separate and adopt random coil configurations. The higher the content of guanine and cytosine, the higher is the temperature at which chain separation occurs.

5. In the cell, DNA is packaged with proteins to form chromatin. The basic unit of this packaging is the nucleosome. In the nucleosome, about 200 bp of DNA are associated with a protein complex called the histone octamer, and with a molecule of the protein histone H1. The string of nucleosomes then wraps itself up into a helical structure, with six nucleosomes per turn, called the 30 nm fibre. In turn, this fibre forms loop structures with the bottoms of the loops attached to a protein scaffold. This is the form in which DNA (probably) exists in non-dividing cells.

6. DNA replication is semi-conservative. That is, the double helix unwinds and two new chains are synthesized, each using one of the original chains as a template.

7. Synthesis of new DNA chains involves adding nucleotides, derived from deoxynucleoside triphosphates, to the 3′-end of the growing chain. The reaction is catalysed by DNA polymerase III. The leading strand is synthesized continuously from the 5′-end to the 3′-end. The lagging strand is synthesized as a set of Okazaki fragments, again in a 5′→3′ direction, and the fragments are joined into a continuous strand by DNA ligase.

8. DNA polymerase has a proofreading role. This enzyme detects the incorporation of an incorrect base into the growing chain and removes it. The fidelity of DNA replication is of the order of one incorrect base for every 10^{10} bases incorporated.

9. DNA is subject to damage by external chemical and physical agents. UV light causes the formation of thymine dimers. Some chemical agents cause substitution of one base for another; others lead to base modification. Intercalating agents tend to cause insertion of one or more extra bases during replication; this results in frame-shift errors in translation.

10. Thymine dimers can be removed by a light-dependent enzyme that catalyses reversal of the dimerization reaction. The most common method for correction of mutations is excision repair. In the case of single incorrect or modified bases, the base is removed by a glycosidase, followed by excision of the ribose-phosphate unit, and repair of the gap by the concerted action of DNA polymerase I and DNA ligase.

Problems

3.1. Refer to Figure 3.6. What is the base sequence of the polynucleotide shown, reading from the bottom of the figure up? What is the chain direction from bottom up?

3.2. The %(G+C) of the DNA from the herpes simplex virus is 72%. What would its melting temperature be under the same conditions as those used in Worked Problem 3.3?

3.3. In Worked Problem 3.4 it was shown that the length of the DNA molecule from the largest chromosome of the mouse, if fully extended in the B-form, would be 6.53×10^7 nm. Given that the average diameter of a cell is 10 μm, calculate the percentage of the volume of the cell, taken to be a sphere, which this DNA molecule occupies. (The volume of a cylinder is given by $\pi r^2 l$ and that of a sphere is $4\pi r^3/3$).

3.4. The histone octamer has a diameter of about 6 nm and the DNA double helix is 2 nm wide. Use this information to calculate the factor by which DNA in the core histone particle is compacted.

3.5. Refer to Figure 3.17. Explain the results obtained in the third generation (the relative intensities of the bands in the centrifuge tube are about 1:3).

3.6. Explain how conversion of adenine to hypoxanthine in DNA by nitrous acid would lead to A/T to G/C transitions.

3.7. The genome of *E. coli* contains 4.8 Mbp of DNA. How long will it take to replicate? What is the likelihood that an error will occur during replication?

References

1. J. D. Watson and F. H. C. Crick, *Nature*, 1953, **171**, 738.
2. S. Doonan, *Peptides and Proteins*, The Royal Society of Chemistry, Cambridge, 2002, p. 117.
3. W. Cochran, F. H. C. Crick and V. Vand, *Acta Crystallogr.*, 1951, **7**, 526.
4. S. Furberg, *Nature*, 1949, **164**, 22.
5. E. Chargaff, R. Lipschitz, C. Green and M. E. Hodes, *J. Biol. Chem.*, 1951, **192**, 223.

6. J. D. Watson, *The Double Helix: A Personal Account of the Discovery of the Structure of DNA*, Athenaeum Press, New York, 1968 (reprinted with an Introduction by Steve Jones, Penguin, London, 1999).
7. R. Olby, *The Path to the Double Helix*, Macmillan, London, 1973.
8. A. Sayre, *Rosalind Franklin and DNA*, Norton, New York, 1975.
9. B. Maddox, *Rosalind Franklin: The Dark Lady of DNA*, Harper-Collins, London, 2002.
10. G. Rhodes, *Crystallography Made Crystal Clear*, Academic Press, San Diego, 1993.
11. H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov and P. E. Bourne, *Nucleic Acids Res.*, 2000, **28**, 235.
12. T. A. Larsen, M. L. Kopka and R. E. Dickerson, *Biochemistry*, 1991, **30**, 4443.
13. J. D. Hepworth, D. R. Waring and M. J. Waring, *Aromatic Chemistry*, The Royal Society of Chemistry, Cambridge, 2002.
14. C. Tanford, *The Hydrophobic Effect: Formation of Micelles and Biological Membranes*, Wiley-Interscience, New York, 1973.
15. C. R. Calladine and H. R. Drew, *Understanding DNA: The Molecule and How it Works*, 2nd edn., Academic Press, San Diego, 1997.
16. D. Hewish and L. Burgoyne, *Biochem. Biophys. Res. Commun.*, 1973, **52**, 504.
17. K. Luger, A. W. Mader, R. K. Richmond, D. F. Sargent and T. J. Richmond, *Nature*, 1997, **389**, 251.
18. J. D. Watson and F. H. C. Crick, *Nature*, 1953, **171**, 962.
19. M. Meselson and F. W. Stahl, *Proc. Natl. Acad. Sci. USA*, 1958, **44**, 671.
20. R. Okazaki, T. Okazaki, K. Sakabe, K. Sugimoto and A. Sugino, *Proc. Natl. Acad. Sci. USA*, 1968, **59**, 598.
21. U. Hubscher, H. P. Nasheuer and J. E. Syoaoja, *Trends Biochem. Sci.*, 2000, **25**, 143.
22. J. C. Wang, *Annu. Rev. Biochem.*, 1996, **65**, 635.
23. L. Malinina, M. Soler-Lopez, J. Aymami and J. A. Subirana, *Biochemistry*, 2002, **41**, 934.

Further Reading

R. L. P. Adams, J. T. Knowler and D. P. Leader, *The Biochemistry of the Nucleic Acids*, 11th edn., Chapman and Hall, London, 1992, Chapters 6 and 7.

J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, 5th edn., Freeman, New York, 2002, Chapter 27.

F. L. Holmes, *Meselson, Stahl, and the Replication of DNA: A History of ldquo;The Most Beautiful Experiment in Biology*, Yale University Press, New Haven, 2001.

4

Transcription and Translation of the Genetic Message

Aims

By the end of this chapter you should understand:

- How the genetic message encoded in DNA is transcribed into a message encoded in mRNA
- How the RNA transcribed from the split genes of eukaryotes is spliced to remove sequences that do not code for protein
- The structures of tRNA molecules and their role in protein synthesis
- The events that occur on the ribosome during protein synthesis

4.1 The Three-dimensional Structure of RNA

To understand some of the processes involved in RNA synthesis, it is necessary to know something of its three-dimensional structure. RNA is a single-stranded molecule, but nevertheless most RNAs are capable of forming double-helical structures over parts of their sequences. RNA does this by folding back on itself so that regions that have complementary, or nearly complementary, sequences can form duplex structures. The hydrogen bonding in these duplexes is mainly of the conventional type, that is A paired with U (remember that U replaces T in RNA) and G paired with C. Some unusual base pairing does, however, occur; for example, it is not uncommon to find G/U base pairs.

The duplex regions, referred to as **stems**, are terminated by **hairpin loops**, usually with between three and six unpaired nucleotides. There may also be **interior loops** within a stem in which the bases do not hydrogen bond, and it is also common to find within a stem a single base that is not paired and forms a **bulge**. The whole structure is referred to as a **stem-loop structure** and a typical example is shown in Figure 4.1.

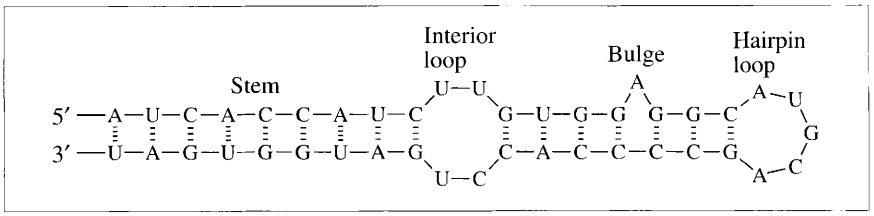


Figure 4.1 Stem–loop structure of a section of an RNA molecule

The hydrogen bonds are shown simply as dashed lines, but recall that there will be two such bonds between A and U, and three between C and G. The complementary regions of such a stem–loop structure are double helical with helix parameters approximating to those of the A-form of DNA; that is, with 11 base pairs per turn of the helix (RNA cannot form a B-type double helix because there is not enough space for the hydroxyl group on C-2′). In general, RNA helices have irregular structures because of the presence of interior loops and bulges.

Double-helical structures can also form between complementary DNA and RNA chains. The usual base pairing rules apply, but again the structure can contain irregular features such as bulges. An example is shown in the stereo model in Figure 4.2 (1EFO[1]). Note that there is an adenine bulge in the RNA chain in this duplex (see Problem 4.1). DNA/RNA hybrids play an important role in transcription, as we shall see in the next section.

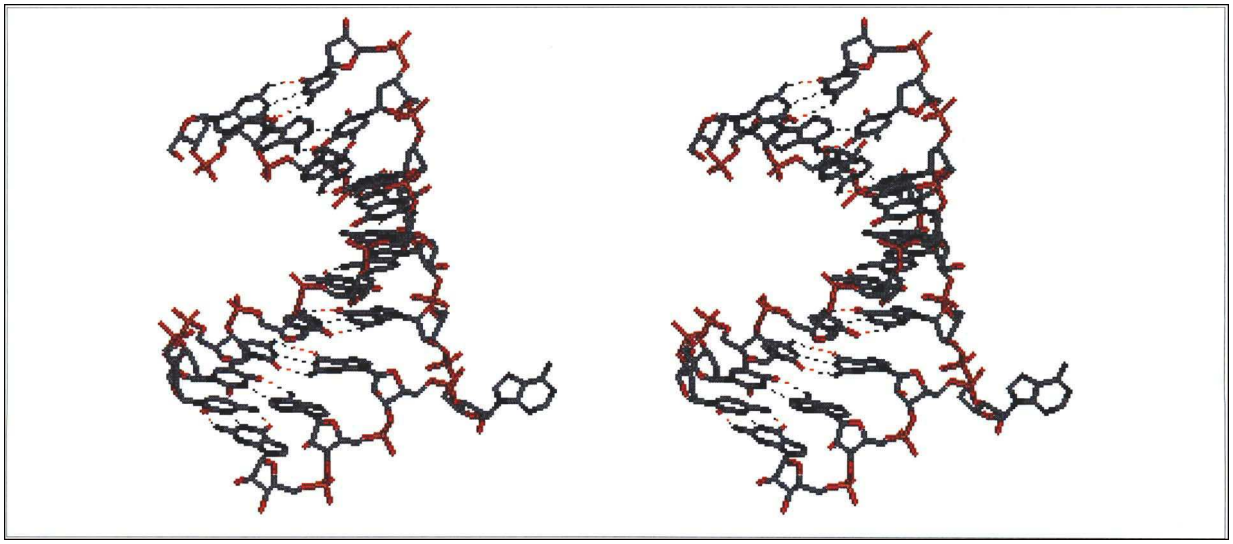


Figure 4.2 Three-dimensional structure of a DNA/RNA hybrid containing a bulge in the RNA chain

4.2 Synthesis of Messenger RNA

A brief account of **transcription**, that is, the process by which the genetic message in DNA is converted into a message contained in the sequence of RNA, was given in Section 1.5. We must now look at the process in a bit more detail. The first point to note is that DNA is a double-stranded molecule, but the information constituting a particular gene is contained in only one strand, called the **coding strand**, read in the 5′→3′ direction. This is not the strand that is transcribed. RNA is synthesized taking its instructions from the complementary strand, or **template strand** as it is called. This is summarized in Scheme 4.1.

This should not be taken to mean that all the genetic messages are contained in the same strand of DNA. Some messages are read from one strand, some from the other. In the very small genomes of viruses it is not unusual for some regions of both strands to contribute to different genetic messages.

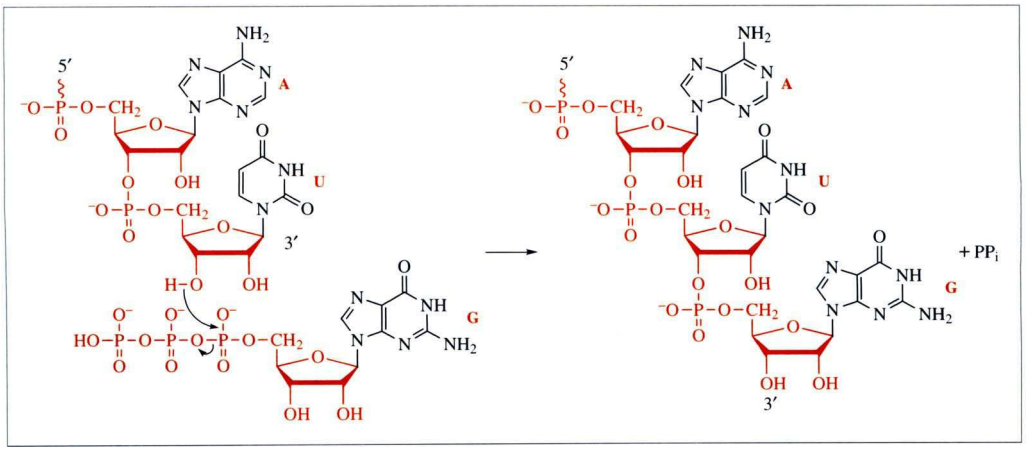
5′	-TCGGATCGGTAATGCCCTGATAGCCTAGCT-	3′	Coding strand of DNA
3′	-AGCCTAGCCATTACGGGACTATCGGATCGA-	5′	Template strand of DNA
5′	-UCGGAUCGGUAAUGCCCUGAUAGCCUAGCU-	3′	mRNA

Scheme 4.1

The genetic code is sometimes written in terms of DNA codons rather than as RNA codons, as was done in Table 1.3. So, for example, the initiator codon is AUG in RNA terms, but ATG in DNA terms.

The DNA coding strand is shown in red, and the template strand in black. In RNA synthesis the template strand is transcribed, the order of bases in the RNA being dictated by the normal Watson–Crick base pairing, except that adenine in DNA pairs with uracil (U) in RNA. The result of this is that the RNA transcript contains just the same message as does the coding strand on the DNA, except that U occurs in place of T.

The chemistry of the chain elongation step in synthesis of RNA is essentially the same as that for the synthesis of the leading strand in DNA replication (Scheme 3.2). Synthesis is in the 5′→3′ direction and the incoming nucleotide is donated by the appropriate nucleoside 5′-triphosphate. This is summarized in Scheme 4.2. The last residue in the growing chain is a U, and the next one to be added is a G, specified by a C in the template strand of DNA. Again, pyrophosphate is liberated, and hydrolysis of this by pyrophosphatase drives the reaction to completion. Note that the 5′-residue of the transcript will retain its triphosphate group. The significance of this will be returned to below.



Scheme 4.2

RNA polymerases were discovered by Severo Ochoa. He shared the Nobel Prize in Medicine in 1959 with Arthur Kornberg "for their discovery of the mechanisms in the biological synthesis of ribonucleic acid and deoxyribonucleic acid".

Proteins that consist of two or more polypeptide chains that are not covalently linked together are said to show **quaternary structure**. The individual polypeptides are specified by separate genes which may or may not be on the same chromosome. A well-known example is the blood protein haemoglobin, which consists of two chains called α, and two chains called β. That is, it has the quaternary structure $\alpha_2\beta_2$.

The three-dimensional structure of the core enzyme from the bacterium *Thermus aquaticus* has been determined by Xhang *et al.*,[2] and that from a eukaryote (yeast) has been solved by Cramer *et al.*[3] These are immensely complex structures containing many thousands of atoms and they are far too complicated to go into here. Suffice it to say that the structures have provided an insight into how these very sophisticated molecular machines work.

The reaction in Scheme 4.2 is catalysed by **RNA polymerase**. RNA polymerases are very complex enzymes. The enzyme from prokaryotes consists of five separate polypeptide chains, or **subunits**; there are two copies of a subunit called α, and one each of subunits called β, β′ and σ. The corresponding enzyme from eukaryotes (called RNA polymerase II, or Pol II) is even more complicated, and contains 12 protein subunits. However, both types of enzyme function in essentially the same way.

Although RNA polymerase catalyses the same basic reaction as does DNA polymerase III, there are some very significant differences. Firstly, RNA polymerase does not require a primer and can start new RNA chains from scratch; consequently, it does not have 5′→3′ exonuclease activity (see Section 3.5). Secondly, it does not proofread the added nucleotide, and so it is not a 3′-exonuclease. It does, however, have other activities that are required to deal with very significant problems involved in RNA synthesis.

Only a very small part of the DNA in a chromosome, the stretch corresponding to a single gene or a small group of genes (as found in **operons**; see Box 4.6), is transcribed into mRNA in any one transcription event. That is, RNA polymerase usually transcribes one gene at a time, and moreover it transcribes only those genes whose products are required by the cell at a particular moment. So the first problem is to search the DNA for a particular site, called the **promoter site**, which signals the beginning of the piece of DNA to be transcribed (see Box 4.1). In bacterial RNA polymerases this is the role of the σ-subunit. The $\alpha_2\beta\beta'$ enzyme, called the **core enzyme**, can catalyse elongation of an RNA chain, but it cannot find the right point to start. Instead, the $\alpha_2\beta\beta'\sigma$ enzyme binds to double-stranded DNA and slides along it until the σ-factor recognizes a promoter site. Transcription then starts, and after the first 10 or so residues have been incorporated, the σ-factor dissociates and the core enzyme carries on transcribing the DNA chain; the σ-factor is now available to bind to another core enzyme and assist it to find a different promoter site.

Transcription cannot start until a segment of the DNA is unwound, because the enzyme uses base pairing with the template strand to specify the identity of the nucleotide to be inserted into the RNA molecule. At the start of transcription, a section of DNA that is 17 base pairs long is unwound. This unwinding activity is an integral part of the function of RNA polymerase. The structure containing the polymerase and the unwound section of DNA is called the **transcription bubble**. This structure now moves along the DNA double helix, unwinding it as it goes and winding back up that section that it has passed, at a rate of about 50 base pairs per second. As it goes the

template strand of the DNA is transcribed. The RNA that is synthesized at first makes a **DNA-RNA hybrid helix** with the template strand, but after eight nucleotides have been incorporated, a region of the polymerase acts to pull the strands apart, thus liberating the growing RNA chain and freeing the transcribed region of the template strand to reform the DNA double helix.

Finally, the RNA polymerase has to know where to stop. The most common **termination signal** consists of a G/C-rich stretch in the template strand, followed by a string of As. What does this do? Firstly, the RNA coded by the G/C rich region forms a stem–loop structure, as shown in Figure 4.3, by base pairing between the Cs and Gs. These are strong interactions, and disrupt the formation of the usual RNA/DNA hybrid. This causes the RNA polymerase to stall after it has incorporated a small number of extra nucleotides. Those extra nucleotides are a string of Us. The hydrogen bonding between A in DNA and U in RNA is weak, and this allows the RNA chain to dissociate from the DNA template, thus terminating transcription.

Some genes lack these termination signals. In such cases, termination is brought about by specific protein factors that recognize particular sequences in the newly synthesized RNA, and then cause the RNA to break away from the RNA/DNA hybrid. We will not pursue this matter here, but information on it can be found in the texts listed under Further Reading.

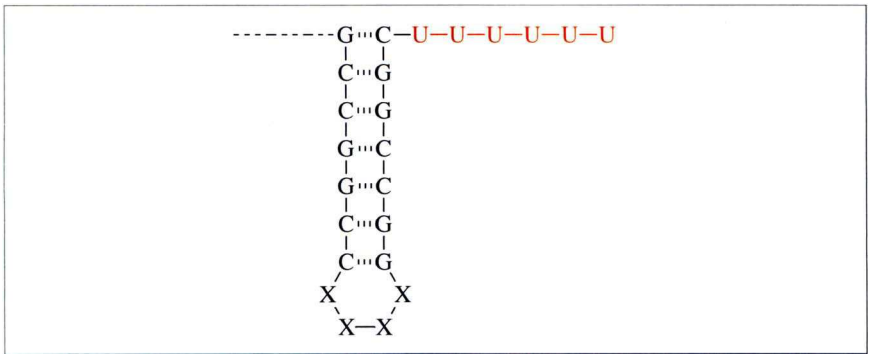


Figure 4.3 Termination of transcription. The runs of G/C residues form a stem–loop structure. This is followed by a string of Us (shown in *red*) which hydrogen bond only weakly to the DNA template; this leads to dissociation of the RNA from the template

Worked Problem 4.1

Q Consider a bacterial gene that is 600 nucleotides long. How long will it take to transcribe this gene if RNA polymerase moves along the DNA at a rate of 50 residues per second? How many amino acids will there be in the translation product of the gene?

A The time taken to transcribe the gene will be $600/50 = 12$ s. The protein product will be 200 amino acids long, since each triplet of nucleotides codes for one amino acid.

Box 4.1 Promoters

Promoters are sequences in DNA that are recognized by RNA polymerase and direct it to start transcription at the proper place. In prokaryotes, there are two regions of the DNA involved, both located **upstream** (that is, on the 5′-side) of the start point. If the first base in the transcribed region is denoted as +1, then the two regions of the promoter are centred at −10 and −35. Analysis of many different promoters provides a **consensus sequence** for these sites; that is, an average of all the sequences found. The consensus for the −10 site is TATAAT (often referred to as the **Pribnow box**), and for the −35 site it is TTGACA. The closer is the actual sequence to the consensus, the stronger is the promoter; that is, the more frequently the gene is transcribed. The role of the σ-factor is, then, to locate and bind to the promoter sequences, and by so doing position the core enzyme at the right place to start transcription. Note particularly that the −10 site is a region where the double helix of DNA would be expected to be weak because it consists of A/T base pairs. Hence it provides a focus for initial melting of the double helix to start transcription.

Promoters are also involved in transcription in eukaryotes, but the situation is much more complex. The most common promoter has the consensus sequence TATAAA, and is referred to as the **TATA box**. In distinction to the situation with prokaryotes, the position of the TATA box is not fixed, but varies from position −30 to −100, depending on the gene in question. There are also other promoters called the **GC box** (consensus sequence GGGCGG) and the **CAAT box** (consensus sequence GGNCAATCT, where N stands for any nucleotide) that are found in positions between −40 and −150. Proteins called **transcription factors** recognize the TATA box. Binding of the transcription factor unwinds the DNA at this point, and also acts as a focus for the recruitment of RNA polymerase II, which then transcribes the gene.

In addition to promoters, transcription of eukaryotic genes is also influenced by **enhancers**. Enhancer sequences do not have promoter activity, but rather increase the effectiveness of promoters. They may be several thousand base pairs away from the promoter site that they enhance, and it may be that binding of specific proteins to enhancers causes a shape change in the DNA which results in better binding of transcription factors to the promoter in question. Enhancers tend to be tissue specific; that is, they are effective in only one type of cell. For example, a given enhancer may work in liver but not in the brain. This is one reason why different tissues express different proteins.

Worked Problem 4.2

Q The consensus sequence of the Pribnow box is TATAAT; that is, it is a hexanucleotide. Calculate the number of times that this sequence would be expected to occur by chance in the genome of *E. coli*, taking the size of the genome to be 5 Mbp.

A There are four possible bases at each position of a hexanucleotide, so the number of possible sequences is 4^6, or 4096. Hence the number of occurrences of the Pribnow box by chance would be about $(5 \times 10^6)/4000 = 1250$. There are about 2000 promoter sites in the *E. coli* genome, but not all of them have the consensus sequence.

In spite of the similarities in the process of mRNA synthesis in prokaryotes and in eukaryotes, there are also some profound differences. In prokaryotes, the mRNA is synthesized essentially in the form required for translation, and indeed translation usually starts before synthesis of mRNA is complete. In eukaryotes, very extensive processing is required before the mRNA is exported from the nucleus and protein synthesis can begin.

The most immediate requirement for processing stems from the fact that most eukaryotic genes are not continuous stretches of DNA. Rather, the coding sequences, or **exons**, are interspersed with stretches of DNA that are not expressed in the final protein product; these are called **introns**. The general arrangement of the DNA in a eukaryotic gene is shown in Figure 4.4. The black regions of the gene are the sequences that will be expressed in the final protein product; the red regions are the intervening sequences, or introns, that will not be expressed. The initial RNA product of transcription contains both the exon and the intron sequences, and is referred to as **pre-mRNA**.

The fact that the coding sequences of genes may be split into sections was discovered independently by groups working with Richard Roberts[4] and Phillip Sharp.[5] Roberts and Sharp were awarded the Nobel Prize in Medicine in 1993 "for their discoveries of split genes".

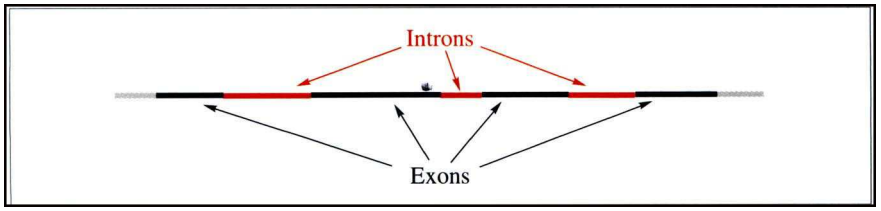


Figure 4.4 Arrangement of introns and exons in a eukaryotic gene

The number of introns in different genes varies over a very wide range. For example, the genes for the α- and β-globin proteins that constitute the oxygen-transporting protein haemoglobin have only two introns. On the other hand, the gene for the protein dystrophin has 78

Dystrophin is an important component of skeletal muscle, and inherited defects in its structure give rise to the disease Duchenne muscular dystrophy.

This is a fatal degenerative disease that affects about one in every 3500 males. It affects only males because the gene is carried on the X-chromosome. Males have only one X-chromosome, so if the gene is defective then dystrophin cannot be made. In females, a defect in the gene on one X-chromosome will be compensated by an active gene on the other. The disease is said to be **sex-linked**.

introns. Dystrophin is a very large protein and the coding regions of its gene contain about 11 kbp of DNA, but this is far outweighed by the amount of DNA in the introns, which amounts to 2.4 Mbp; this is the largest gene known. Generally, introns are smaller than those occurring in the dystrophin gene, and range from about 100 bp to 1000 bp long.

One aspect of the processing of pre-mRNA, then, is **splicing**, which is the cutting out of the introns and joining together of the exons to make a continuous message. Note that this has to be done with great accuracy – if the splice is incorrect by one base, this will cause a frame-shift mutation (see Section 1.6), and the product protein after that splice site will have completely the wrong structure (or will be prematurely terminated). There must, then, be signals in the RNA which specify the beginning and the end of an intron. Most introns conform to the so-called **GU–AG rule**. This means that at the 5′-end of the intron there is an invariant GU sequence, and at the 3′-end there is an invariant AG sequence. These dinucleotide sequences will occur frequently in any RNA molecule, and so there must be other elements in the base sequence that specify the intron/exon junctions. Indeed, there are known consensus sequences for the residues following the GU and preceding the AG, but it is still a very difficult problem to predict in a pre-mRNA molecule where the spicing will occur (see Section 5.9).

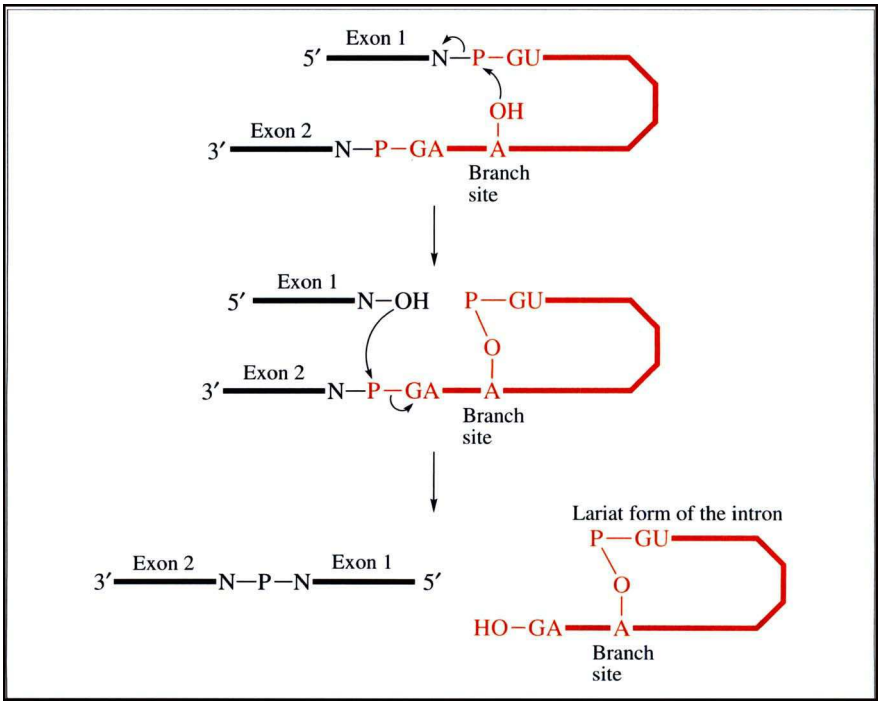
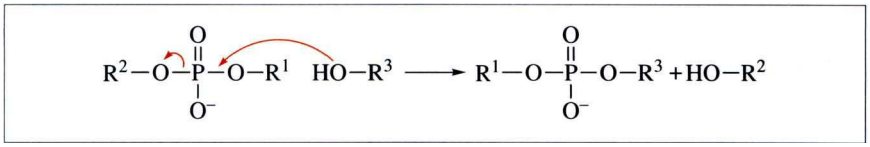


Figure 4.5 Splicing out of an intron. The exons are shown in *black* and the intron in *red*. The 3′-end of exon 1 and the 5′-end of exon 2 are shown as N to denote any nucleotide. Phosphates linking the intron to the exons are shown as P. The OH at the branch site is the 2′-OH of an adenine

The mechanism of splicing is shown in Figure 4.5. An internal region of the intron, referred to as the **branch site**, contains an adenine residue, the 2′-hydroxyl group of which carries out a nucleophilic attack on the phosphate linking the terminal residue of exon 1 to the 5′-G of the intron. The exon is liberated, and has a free OH at position 3′ of the terminal residue. The 5′-terminal G of the intron is linked to the 2′-OH of the branch site adenine to from a cyclic structure (see Problem 4.3). The newly formed 3′-OH of exon 1 now attacks the link between the 3′-G of the intron and the 5′-phosphate of exon 2. This results in linkage of the two exons and liberation of the intron in a cyclic from known as a **lariat** (note that the spliced exons are drawn 3′→5′ in Figure 4.5). The key reactions in Figure 4.5 are **transesterifications**, and for clarity the process is shown in more detail in Scheme 4.3.



Scheme 4.3

Splicing is carried out by a structure called a **spliceosome** that consists of both proteins and several small RNA molecules. These complexes are called **snRNPs** (standing for small nuclear ribonucleoprotein particles, and colloquially called **snurps**). Part of the role of the RNA components of the snurps is to recognize and bind to the conserved sequences at the splice points and at the branch site. Interactions between the components then bring the splice sites together for reaction to occur. The most remarkable thing about the splicing reactions, however, is that they are catalysed by the RNA components of the spliceosome rather than the proteins.

Box 4.2 Catalytic RNA

One of the most exciting, and least expected, developments in nucleic acid chemistry in recent years was the finding that RNA has a catalytic role in some circumstances. It was previously held that biological catalysts, the enzymes, were always proteins. That this is not so was originally shown by Thomas Cech and his co-workers.[6] It was known that the gene for the 26S rRNA molecule in the protozoan *Tetrahymena thermophila* contains a 413-residue intron, and Cech was working on the excision of this sequence from the pre-RNA produced by transcription of the gene (see Section 4.3 for a discussion of the synthesis of rRNA). Remarkably it was found that

the pre-rRNA caused its own splicing *in vitro*. No protein was involved in the process; all that was required was the addition of Mg^{2+} ions and guanosine (GMP or GTP can substitute for G in the reaction). The mechanism of the splicing reaction is essentially the same as that shown in Figure 4.5, except that the initial transesterification reaction involves attack by the OH group of the guanosine residue, rather than by the OH group of adenine at a branch site. The point of cleavage is entirely specific, and the splicing activity depends on the three-dimensional structure of the pre-RNA. In these respects, the pre-RNA had properties characteristic of an enzyme, and the term **ribozyme** was coined to describe catalytic RNA.

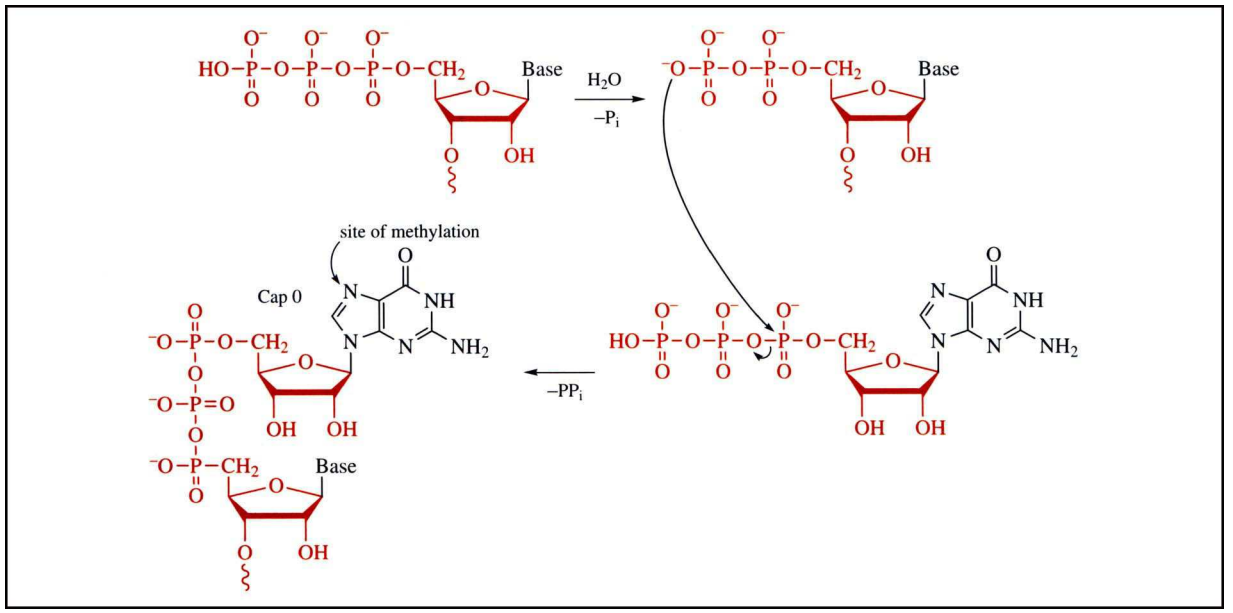
Altman and Cech shared the Nobel Prize in Chemistry in 1989 "for their discovery of the catalytic properties of RNA".

Another example of a ribozyme was not long in coming. This was the enzyme RNase P, an enzyme involved if the processing of pre-tRNA molecules (see Section 4.3), which was studied by Sidney Altman and his team.[7] In this case, the enzyme consists of both a protein part and an RNA part, but again the catalytic activity resides in the RNA component.

A fascinating series of articles on the chemistry and biology of RNA appears as an Insight feature in the edition of *Nature* published on 11 July 2002 (**418**, 214).

Many other ribozymes are now known. For example, catalytic RNA is central to the functions of the ribosome (see Section 4.4.2). These findings have a significance beyond that of the purely chemical interest in how RNA catalyses reactions – they possibly have a bearing on the question of the origin of life. One of the problems associated with theories of the chemical origins of life has been that living systems appear to require both protein enzymes to catalyse their reactions, and an informational macromolecule to store the instructions for the structures of those enzymes. Moreover, enzymes are required for the replication and expression of the informational molecules. This all seemed to be a bit of a "chicken and egg" problem. Which, if either, came first? How could informational macromolecules be made without the enzymes to catalyse the process, and how could the structures of the enzymes be specified without the information in the macromolecules? Catalytic RNA provides a solution to this problem. That is, it is possible that life evolved based on RNA, not on proteins. RNA can certainly play an informational role (and indeed still does in the retroviruses). We now know that it can also act as a catalyst. So both of the functions required for life processes exist in the same molecule. It may be, then, that the first primitive organisms evolved using chemistry based on RNA, and that our ancestors were living in what is now sometimes called an RNA World!

Spicing out of introns is not the only modification that occurs to mRNA in eukaryotes. In fact, both ends of the mRNA are found to be modified. Modification at the 5′-end, a process called **capping**, occurs whilst the pre-mRNA is being synthesized. The first step is hydrolysis of the terminal phosphate of the nucleoside triphosphate at the 5′-end of the chain, followed by reaction with GTP to yield a very unusual 5′→5′ phosphotriester linkage (Scheme 4.4).



Scheme 4.4

The terminal guanosine is then methylated at N-7 to produce a structure known as **cap 0**. The first two ribose units in the original RNA chain may then be methylated on O-2′ to yield **cap 1** (if only the 5′-residue is methylated) or cap 2 (if both residues are methylated); these latter modifications do not, however, always occur. The modification at the 3′-end of the mRNA is more extensive. What happens is that a specific **RNase** recognizes a sequence AAUAAA at the 3′-end of the pre-mRNA transcript and hydrolyses the molecule at that point. Thereafter, an enzyme called poly-A polymerase adds a string of A residues to the 3′-end of the transcript to produce a **poly-A tail**. These tails can be anywhere from 20 to 200 residues long. The purpose of these modifications is not entirely clear. It is thought that the primary function of capping is to protect the newly synthesized pre-mRNA from hydrolysis by exonucleases. The poly-A tail probably also functions in this way, but there is also evidence that its presence is required for efficient translation of the genetic message. There is more on the subject of poly-A tails in Box 5.3.

RNase is the usual abbreviation for **ribonuclease**. Ribonucleases are enzymes that catalyse the hydrolysis of RNA. There are exoribonucleases which hydrolyse residues from the ends of RNA chains, and endoribonucleases which hydrolyse RNA at internal positions.

Box 4.3 Alternative Splicing of Introns

Nearly all the genes of higher eukaryotes are split into exons and introns. Lower eukaryotes such as yeast have fewer split genes, and in prokaryotes split genes are very rare. This could be taken to indicate that the phenomenon has developed late in evolution, but the evidence suggests otherwise. The fact that introns existed very early in evolution is indicated by the fact that splicing mechanisms are the same in fungi, plants and vertebrates. It seems that introns were present in the genes of the ancestors of all forms of life but have been lost in prokaryotes. This presumably reflects their need for very rapid protein synthesis to support their short generation times. Translation of mRNA in prokaryotes starts before mRNA synthesis is complete and this is clearly inconsistent with the existence of introns.

So the question arises as to why split genes have been retained in eukaryotes. At least a part of the answer is that they provide the opportunity for increased diversity of protein products by **alternative splicing**. Consider, for example, a gene with three exons and two introns. Assuming that exon 1 must be included because it contains signals for translation of the product, there are still three different mRNAs (and hence proteins) that could be produced by the alternative splicing patterns 1→2→3, 1→2 and 1→3. At least 20% of genes in higher eukaryotes are thought to be involved in alternative splicing, thus greatly increasing the repertoire of proteins that they can produce.

An interesting example of alternative splicing involves sex determination in the fruit fly, *Drosophila*. A gene involved in this process is called *tra* (for **transformer gene**). The gene has three exons, as shown diagrammatically in Figure 4.6. In male embryos, splicing of all three exons occurs, but the second exon contains an in-frame stop codon, so an active protein product cannot be made. In the female embryos, splicing of exon 1 to exon 2 is prevented by binding to the 3′-end of the first intron of a protein called **sex-lethal protein** (Sxl protein). The result is that exon 1 links to exon 3 instead. The product of this mRNA is called Tra protein. The Tra protein then functions to repress expression of genes required for male sexual development, and the fly develops as a female.

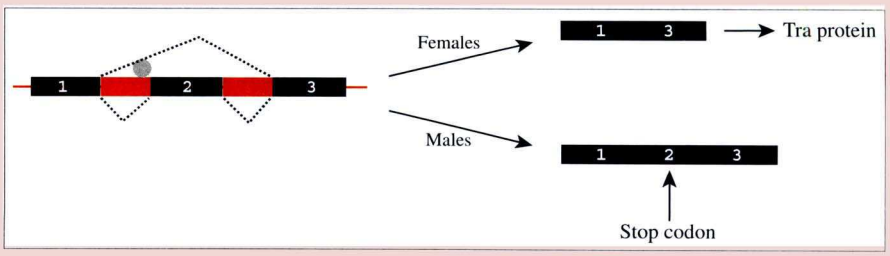


Figure 4.6 Alternative splicing of the *tra* gene transcript in female and male *Drosophila*. In females, the Sxl protein (shown as a *grey sphere*) binds to the end of intron 1 and prevents splicing of intron 1 to intron 2

4.3 Synthesis of Ribosomal and Transfer RNA

In prokaryotes, the genes for rRNAs are organized in operons (see Box 4.6 for further explanation of this term), each of which contains not only the genes for the 5S, 16S and 23S rRNA species, but also genes for one or more tRNA molecules. In *E. coli*, for example, there are seven such operons, the large number presumably a reflection of the fact that the organism has a need for a large number of ribosomes to provide for its rapid rate of protein synthesis. An example of the gene arrangement for one such operon from *E. coli* is shown in Figure 4.7. Transcription of these operons is done by the same enzyme that is responsible for synthesis of mRNA. Once the transcript is produced, it is processed by RNases to liberate the mature rRNA molecules. This is a complex process and we will not go into the details.

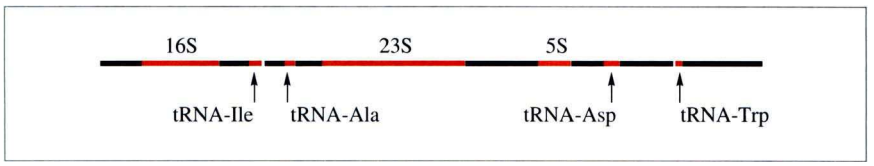


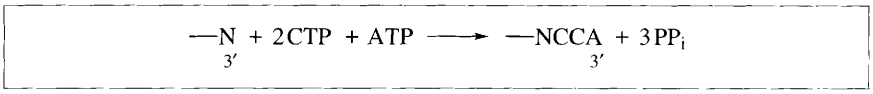
Figure 4.7 Arrangement of rRNA and tRNA genes in an operon from *E. coli*. The regions coding for the final molecules are in *red* and regions which are removed are in *black*

The situation is similar, but rather more complex, in eukaryotes. Eukaryotic ribosomes contain four rRNA species; that is, 5S, 5.8S, 18S and 28S. In most organisms, the genes for 5.8S, 18S, and 28S rRNAs are linked together in **transcription units**, and many of these units are found in clusters throughout the genome. For example, mammals have 100–300 of them. This, and the very high rate at which they are transcribed, is a reflection of the fact that 80% of the RNA in eukaryotic cells is rRNA. These genes are transcribed by RNA polymerase I (Pol I), and again, the mature rRNAs are liberated from the transcripts by RNases. In most eukaryotes the 5S RNA molecules are coded by DNA not linked to that for the other rRNA species. These genes are also spread throughout the genome, but they are transcribed by a different enzyme (RNA polymerase III, or Pol III).

The transcripts of tRNA genes require processing at both the 5′- and the 3′-ends to generate the mature tRNA molecules. A key enzyme in this processing, both in eukaryotes and prokaryotes, is RNase P. RNase P is a ribozyme (see Box 4.2). The role of this enzyme is to generate the correct 5′-terminus of the tRNA molecules. Processing at the 3′-terminus is more complicated. The reason is that all mature tRNA molecules must terminate in the sequence CCA (see Section 4.4) and, particularly in eukaryotes, this sequence is not usually encoded by the gene. Hence after initial trimming away of extra nucleotides, the terminal CCA sequence

has to be added by an enzyme called **tRNA nucleotidyl transferase**. The reaction catalysed by this enzyme is shown in Scheme 4.5, where N represents the 3′-residue of the pre-tRNA molecule.

Scheme 4.5



Finally, many of the bases in tRNA molecules are chemically modified by enzyme systems that are not fully understood. We have already encountered two of the modified bases that occur (pseudouridine and dihydrouridine) in Worked Problem 2.3 and Problem 2.3. Particularly common is methylation to yield derivatives such as 7-methylguanosine, 1-methyladenine and 5-methylcytidine. A different sort of modification also occurs when a base, frequently adenine, is replaced by inosine. In addition, methylation may occur at the 2′-OH groups of some of the ribose residues (see Figure 4.11 for details of the modified bases and nucleosides that occur in one particular tRNA molecule).

4.4 Translation of Messenger RNA

4.4.1 The Role of Transfer RNA

The bare outlines of how mRNA is translated into protein have already been covered in Section 1.7. To recap, the mRNA is read from the 5′-end three bases at a time, each triplet of bases coding for insertion of an amino acid into the growing polypeptide chain. The protein chain is synthesized from the N-terminal end. At some point, a termination codon is reached, and translation stops. We now need to look at how these processes happen.

Worked Problem 4.3

Q The synthetic polynucleotide 5′-A(A)$_n$AC-3′ can be used to direct protein synthesis *in vitro*. The product protein has the amino acid sequence Lys-(Lys)$_m$-Asn. Explain how this result shows that mRNA is translated 5′→3′.

A Reading from the 5′-end, there are repeating AAA codons which specify Lys (see Table 1.3). The 3′-codon is AAC, which codes for Asn. This amino acid is at the C-terminus of the protein and hence must have been added last. The message is, therefore, read 5′→3′.

The first problem that came into focus after the nature of the genetic code was established was how a code written in terms of a sequence of nucleotides could be used to specify the order of a set of very different molecules, the amino acids, in a protein chain. It did not seem likely that the amino acids could interact directly with the codons in the mRNA. The solution to the problem was provided by Francis Crick, who came up with the brilliant idea that **adaptors** were required to do the job. The suggestion was that an amino acid would be linked with a specific adaptor, and that the adaptor would then recognize the codon for its amino acid in the mRNA. It seemed reasonable to suppose that the adaptors would have, as part of their structures, a triplet of nucleotides that would recognize the codon using conventional Watson–Crick base pairing. The adaptor hypothesis turned out to be essentially correct, and the adaptors were identified as group of small RNA molecules called **transfer RNA** or **tRNA**.

Within a relatively short time of the formulation of the adaptor hypothesis, the nucleotide sequence of a tRNA molecule, the alanyl-tRNA from yeast, was determined by Robert Holley and his co-workers[8] (alanyl-tRNA is the particular tRNA that is specific for alanine). This was a remarkable achievement. Not only was it the first nucleic acid to be completely sequenced, it was also found to contain a large number of modified bases, the structures of which had to be determined. The molecule contains a total of 76 nucleotides. Not only did Holley and his team determine the nucleotide sequence, they also proposed that the molecule was folded into what is known as a **clover leaf structure** by base pairing between complementary regions of the polynucleotide chain. This proposal has now been shown to be correct.

Holley shared the Nobel Prize in Medicine in 1968 with Khorana and Nirenberg "for their interpretation of the genetic code and its function in protein synthesis".

The structures of many tRNA molecules are now known, and they all have some features in common, as shown in the cartoon in Figure 4.8. The 3′-end of the molecule always has the sequence CCA, and the terminal adenine is not phosphorylated. The 5′-residue, most often a G, is phosphorylated on the 5′-OH. There are four regions where the chain runs antiparallel to itself and forms hydrogen bonds between complementary sequences (hydrogen bonds are shown as single lines between the cells in Figure 4.8); the molecule in these regions is double helical. Three of these hydrogen-bonded stems end in loops. The **DHU loop** is so called because it usually contains one or two residues of dihydrouracil. The **TΨC loop** always contains that triplet of bases at the 5′-end of the loop (recall that thymine does not normally occur in RNA). Finally, the **anticodon loop** contains the three bases that constitute the anticodon. In addition to these, there is an **extra loop** which is of variable size,

and accounts largely for the differences in length from one tRNA molecule to another.

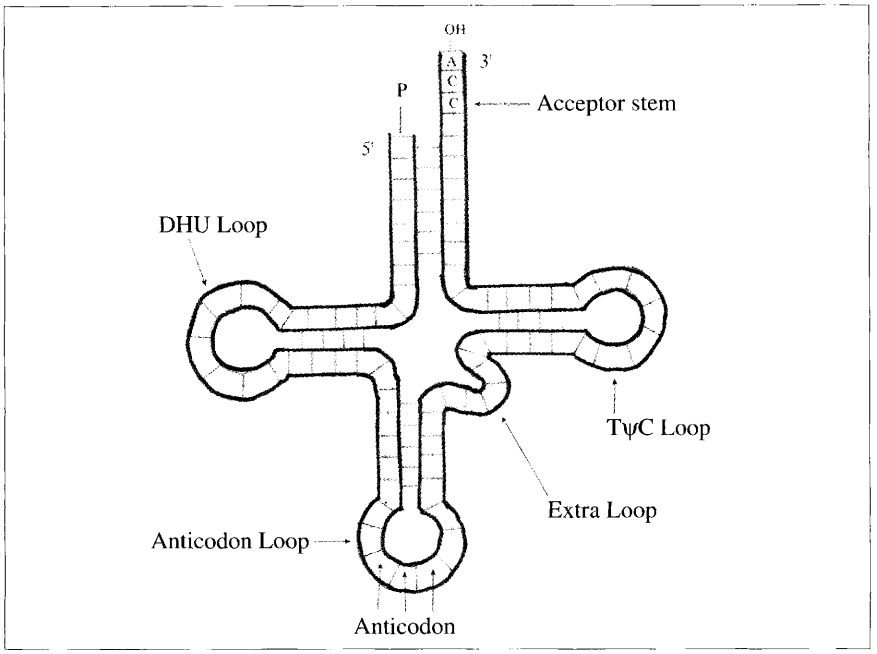
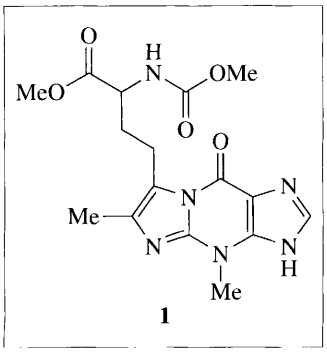


Figure 4.8 Cartoon of the structure of a tRNA molecule. Each of the cells represents a nucleotide residue, and the lines between residues are hydrogen bonds



1

Confirmation of the correctness of the clover leaf structure came when the structure of a t-RNA molecule was determined by single-crystal X-ray diffraction. The particular t-RNA concerned was the phenylalanyl-tRNA from yeast. Its backbone structure, taken from 6TNA,[9] is shown in Figure 4.9. The molecule is L-shaped, with the point of attachment of the amino acid at the top right, and the anticodon triplet of bases at the bottom. The positions of the loops and the stems are marked so that comparison can be made with the cartoon in Figure 4.8. The same t-RNA in the same orientation is shown in Figure 4.10, this time as a wireframe model, with the anticodon triplet and the 3′-terminal CCA coloured red. Figure 4.11 gives the nucleotide sequence of this t-RNA molecule. Note the large number of modified bases and nucleosides that it contains. **Wybutosine** is a very highly modified guanine with the structure shown in **1**. It is frequently found near the anticodon in t-RNA (see Problem 4.6).

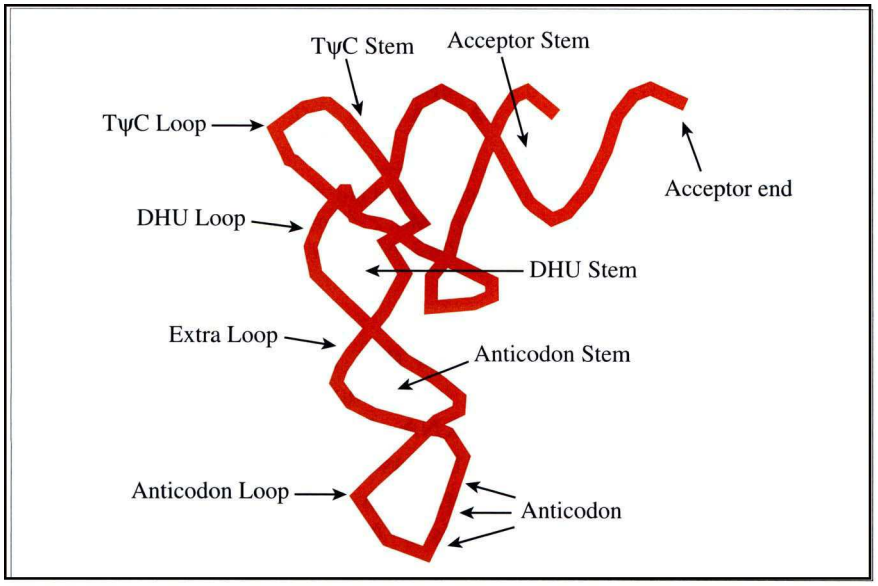


Figure 4.9 Backbone model of phenylalanyl-tRNA from yeast

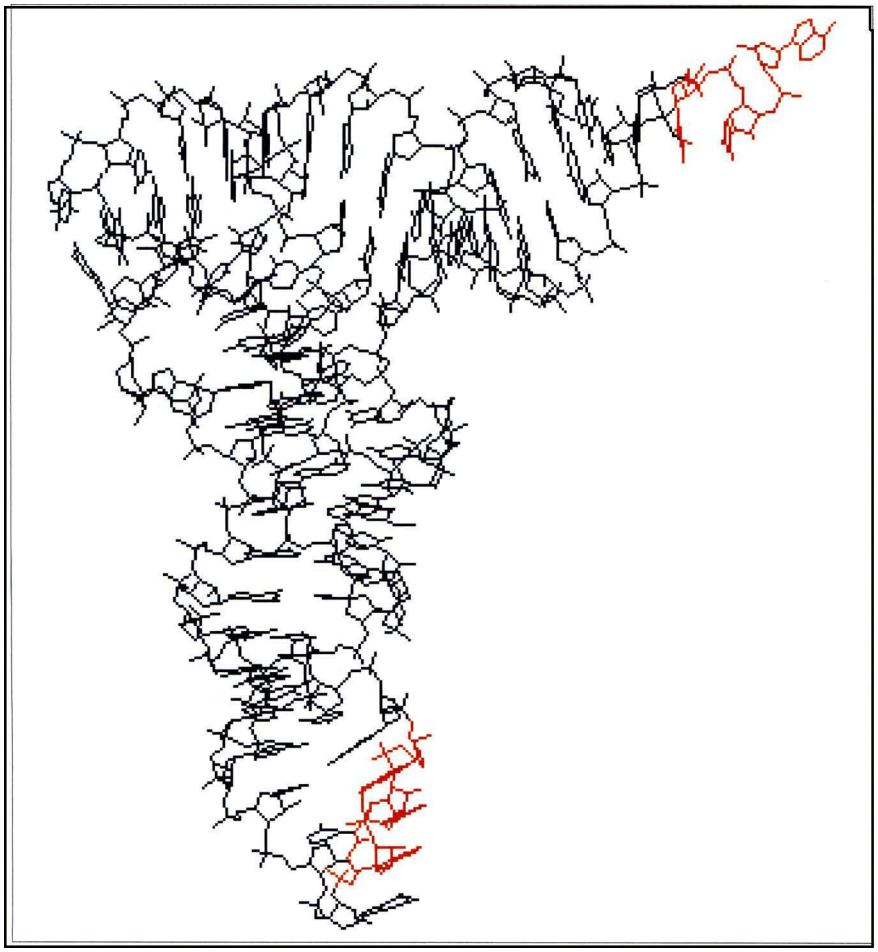


Figure 4.10 Wireframe model of phenylalanyl-tRNA from yeast. The 3′-terminal CCA and the anticodon are shown in *red*

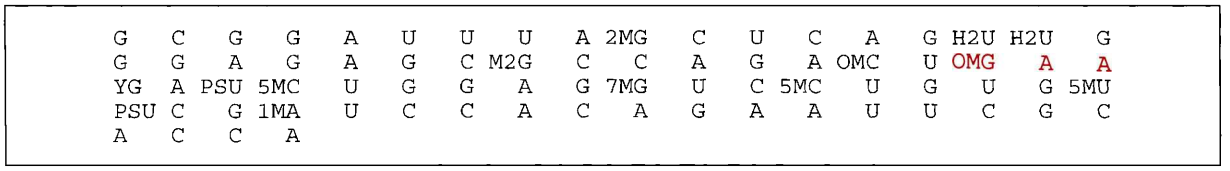


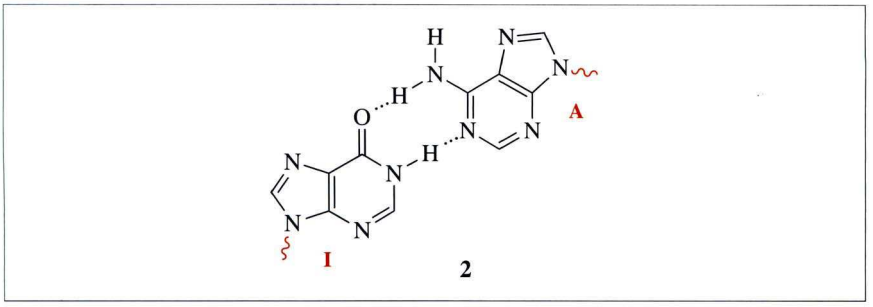
Figure 4.11 Nucleotide sequence of phenylalanyl-tRNA from yeast. Abbreviations used are: 2MG, 6-hydroxy-2-methyaminopurine; H2U, dihydrouracil; M2G, 2-(dimethylamino)-6-hydroxypurine; OMC, 2′-*O*-methylcytidine; OMG, 2′-*O*-methylguanosine; YG, wybutosine; PSU, pseudouridine; 5MC, 5-methylcytosine; 7MG, N^7-methylguanine; 5MU, 5-methyluracil (thymine); 1MA, N^1-methyladenine. The anticodon triplet is shown in *red*

Table 4.1 Allowed pairings between the 5′-base of an anticodon and the 3′-base of a codon

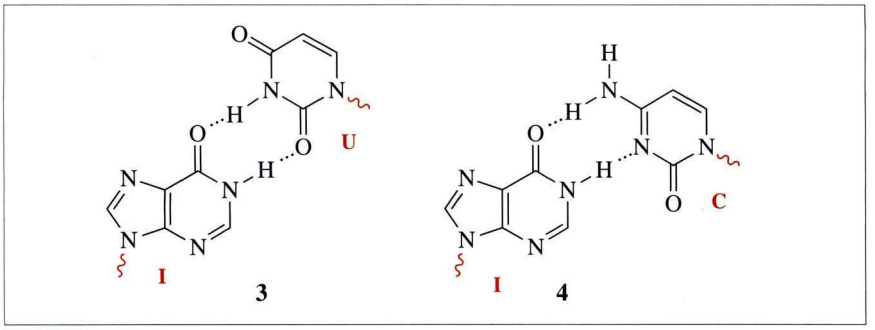
5′-Base of the anticodon	*3′-Base of the codon*
C	G
A	U
U	A and G
G	U and C
I	U, C and A

There are two essential properties that the t-RNA molecule must have: it must accept the correct amino acid, and it must recognize the appropriate codon in the mRNA. To take the second of these first, recognition of the codon is by base paring, with the codon read in the 5′→3′ direction and the anticodon in the 3′→5′ direction. The anticodon of phenylalanyl-tRNA is 3′-AAG-5′ (in fact, the G is 2′-*O*-methylguanosine, but this base pairs in the same way as G), so what codon will it recognize? This will be 5′-UUC-3′. Table 1.3 shows that this is indeed one of the two codons for phenylalanine. Table 1.3 also shows that there is a second codon for phenylalanine, namely UUU. Does this mean that a second t-RNA must exist with the anticodon sequence AAA? It turns out that this is not the case. The same t-RNA can read both of these codons. This is possible because there is some latitude in the base pairing between the 3′-base of the codon and the 5′-base of the anticodon. This possibility was put forward by Crick in what is known as the **wobble hypothesis**.[10] A list of the allowed pairings is shown in Table 4.1. Note that inosine can base pair in the wobble position with three other bases (U, C and A) and consequently it is very frequently found in anticodons so as to maximize the number of codons that can be read. For example, the alanyl-tRNA from yeast has the anticodon sequence 3′-CGI-5′, and it can recognize the codons GCA, GCC and GCU.

The base pairs that I makes with A, U and C are shown in **2–4**, respectively. It can be seen that the geometry is nearly the same in all cases, and the small deviations from the accurate geometry of Watson–Crick base pairs can be accommodated by small movements of the polynucleotide backbone of the t-RNA. The reason why wobble is restricted to the 5′-base of the anticodon seems to be that the 3′-base is always followed by a G, usually heavily modified (wybutosine in the phenylalanyl-tRNA), which forms strong stacking interactions with the preceding two bases and anchors them in a configuration that accepts only standard Watson–Crick base pairs (see Problem 4.6).

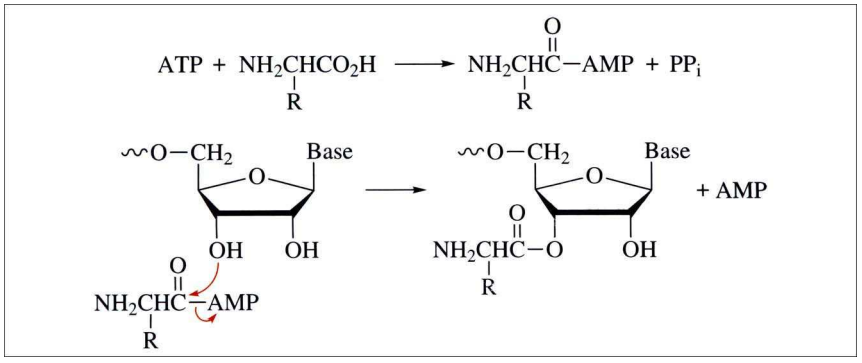


2



3 4

Linking of an amino acid to the 3′-OH group of t-RNA occurs in a reaction involving ATP and catalysed by specific **aminoacyl-tRNA synthetases**. The reaction proceeds *via* an intermediate aminoacyl adenylate compound, analogous to the process described in Scheme 3.4. The reactions involved are shown in Scheme 4.6. The aminoacyl adenylate formed in the first stage of the reaction is a reactive anhydride (see Problem 4.7); the equilibrium for its formation is driven to the right by hydrolysis of the product pyrophosphate. This intermediate reacts with the 3′-OH group of the terminal adenine residue in the tRNA to form an ester linkage. Some aminoacyl-tRNA synthetases initially link the amino acid to the 2′-OH of the ribose, from which it migrates to the 3′-position.



Scheme 4.6

Raney nickel is finely divided nickel on to which hydrogen is adsorbed. It reduces cysteine (side chain $-CH_2SH$) to alanine (side chain $-CH_3$).

It is essential that the correct amino acid is linked to the tRNA, because once the tRNA is charged with an amino acid it will deliver it to the growing protein chain as specified by the codon in the mRNA. If the tRNA is incorrectly charged, then it will incorporate the wrong amino acid. This was shown in a classic experiment where cysteinyl-tRNA was charged with the correct amino acid, which was then converted to alanine by treatment with Raney nickel. This wrongly charged tRNA was found to incorporate alanine into proteins in response to the codon for cysteine.

From the determination of the three-dimensional structures of tRNA–enzyme complexes, a great deal of information is available about how tRNA interacts with its specific aminoacyl-tRNA synthetase. A fairly detailed knowledge of the structural features of proteins is required to fully appreciate the significance of these structures, and so no examples will be included here. If you have already studied protein chemistry and want to look at representative examples, the structure of the enzyme specific for glutaminyl-tRNA is entry 1GTS[11] in the PDB, and that of the enzyme specific for aspartyl-tRNA is entry 1ASY.[12] These are examples of what are known as **Class I enzymes** and **Class II enzymes**, respectively. The way in which the tRNA interacts with the enzyme is different in the two classes. All tRNA synthetases fall into one or other of these two classes.

Hence the correct translation of the genetic code depends as much on the fidelity of charging the tRNA molecules as it does on the recognition of the codon by the anticodon. The error frequency of incorporation of the amino acid seems to be between 1 in 10^4 and 1 in 10^5. This is achieved partly by the specificity of the interactions between the enzyme and its cognate t-RNA, but most aminoacyl-tRNA synthetases also have a proofreading function. Once the amino acid has been linked to the 3′-A of the t-RNA, it visits a second catalytic site on the enzyme, where, if the attached amino acid is incorrect, the bond to the t-RNA is hydrolysed. For example, if threonyl-tRNA synthetase is incubated with its cognate t-RNA that has been artificially charged with serine, the serine is rapidly removed. This is an error that is likely to occur in practice because serine and threonine are chemically similar amino acids (see Table 1.2). The editing site seems to work on the basis that the smaller amino acid (serine) can enter the site where hydrolysis occurs, whereas the correct amino acid, threonine, cannot enter the editing site and so is not hydrolysed.

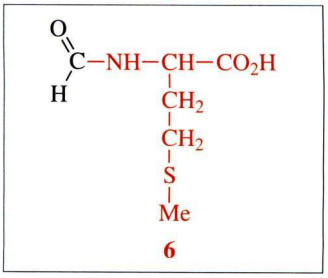
4.4.2 Protein Synthesis

We turn now to the events of protein synthesis, concentrating on mainly as they occur in prokaryotes. The key role here is played by the ribosome, the outline structure of which has already been described in Section 1.7. At the start of protein synthesis, the ribosome exits as separate subunits (30S and 50S particles in the case of prokaryotes). The process begins by binding of the mRNA to the 16S rRNA molecule in a small ribosomal subunit. This binding involves base pairing between the 16S rRNA and a region centred about 10 nucleotides upstream (on the 5′-side) of the initiator AUG codon. This upstream region is rich in purines and is referred to as the **Shine–Delgarno box**. A typical sequence is shown in **5**, where the first group of red nucleotides is the Shine–Delgarno sequence, and the initiator codon (AUG) is also in red.

5′-CACGAGGGGAAAUCUGAUGGAACGCUAC-3′

5

In prokaryotes, protein synthesis always starts with *N*-formylmethionine (**6**, the formyl group is in black). The formyl group is added to the methionine after the latter has been attached to $tRNA_f$. Internal methio- nines are inserted in the protein chain by a different tRNA denoted as $tRNA_m$; methionine attached to this tRNA cannot be formylated. In most cases, the *N*-formylmethio- nine residue is removed from the growing protein chain when it is about 10 amino acid residues long.



6

At this point, a special tRNA, called initiator $tRNA_f$, binds to the AUG codon at the start of the message. This tRNA is charged with methionine which has been modified by formylation of the α-amino group. The situation is now as illustrated schematically in the left-hand part of Figure 4.12, where the black oval represents the small ribosomal subunit, the black line is the mRNA and the red rectangle is $tRNA_f$ charged with *N*-formylmethionine. This complex then recruits a large subunit as shown in the central part of Figure 4.12 (open oval) to produce the complete translation complex. The initiator tRNA is bound at a site in the complex called the **P-site** (P stands for peptide).

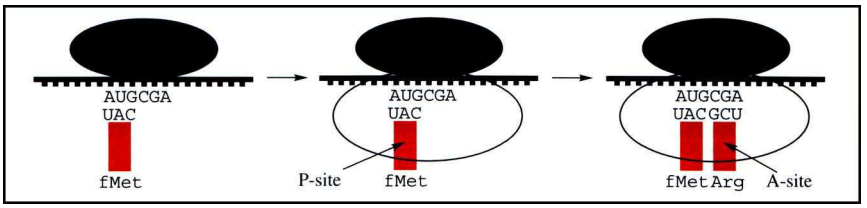
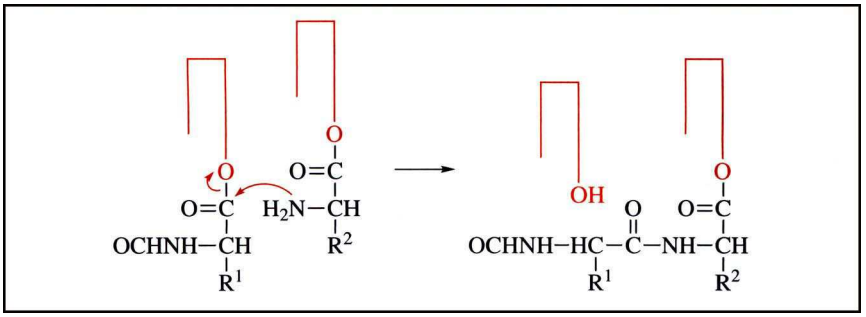


Figure 4.12 The initial stages of protein synthesis. The *black oval* represents the small ribosomal subunit and the *black line* is the mRNA. *Red rectangles* represent tRNA molecules, and the *open black oval* is the large ribosomal subunit

Protein synthesis now proceeds. The charged tRNA specified by the next codon in the mRNA is recruited to the ribosome, and binds at what is called the **A-site** (A stands for aminoacyl). This is shown in the right-hand part of Figure 4.12; the codon in this example is CGA, which specifies insertion of an arginine. The amino group of the amino acid linked to this tRNA is positioned adjacent to the ester linkage between *N*-formylmethionine and its tRNA. A **peptidyl transferase** in the ribosome then promotes transfer of the *N*-formylmethionine residue to this amino group, with formation of a peptide bond. The reaction is given in Scheme 4.7, where the tRNA molecules are shown schematically in red. It is interesting to note that the peptidyl transferase activity is a property of the 23S rRNA and does not involve any of the protein components of the ribosome; that is, the peptidyl transferase is a ribozyme.



Scheme 4.7

At this stage, the growing protein chain is attached to the tRNA molecule in the A-site and the initiator $tRNA_f$ in the P-site is uncharged. Both are still hydrogen bonded to the mRNA. The next event is movement of the mRNA through the ribosome by one codon in the direction right to left. This is promoted by an enzyme called **elongation factor G**. The result is shown in the left-hand part of Figure 4.13. The initiator $tRNA_f$ is detached from the mRNA and occupies a third site on the ribosome called the **E-site** (E stands for exit). The tRNA carrying the growing protein chain occupies the P-site, and the A-site is empty ready to receive the next charged tRNA. Next the initiator $tRNA_f$ exits from the ribosome, and the tRNA coded by the third codon (shown as alanyl-tRNA in Figure 4.13) enters the A-site. Peptide bond formation ensues (right-hand part of Figure 4.13), and the whole process repeats itself to extend the protein chain further.

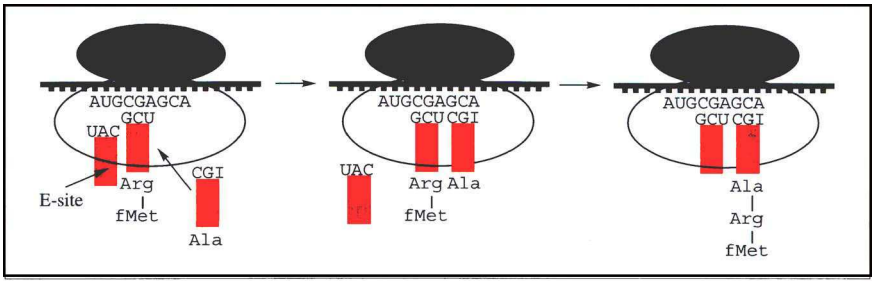


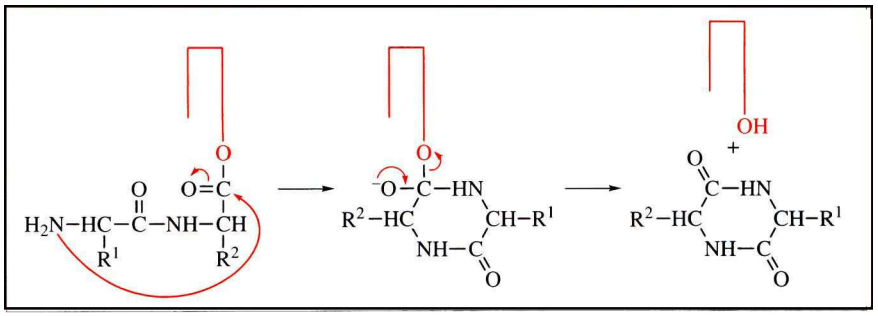
Figure 4.13 Elongation phase of protein synthesis

Worked Problem 4.4

Q Suggest a possible explanation for the fact that prokaryotic protein synthesis is initiated with *N*-formylmethionine.

A The answer may lie in the nucleophilic nature of a free amino group. Consider the situation where *N*-formylmethionine has been added to the second amino acid in the chain (left-hand part of Figure 4.13). If the methionine had a free amino group, then a nucleophilic attack could occur on the ester linkage joining the second amino acid to its tRNA. This would result in the formation of six-membered ring structure and liberation of this derivative (a diketopiperazine) from the tRNA, as shown in Scheme 4.8.

Eukaryotes, in which initiation occurs with underivatized methionine, must have some other way of overcoming this problem.



Scheme 4.8

Box 4.4 Evidence for Three tRNA Binding Sites in the Ribosome

Experimental evidence for the existence of three tRNA binding sites in the ribosome has come from solution of the three-dimensional structure of a complex between the ribosome from the bacterium *Thermus thermophilus* and three copies of phenylalanyl-tRNA. The small subunit with the bound tRNA molecules is shown in Figure 4.14 (1GIX[13]). The RNA molecules are shown as backbone models with the rRNA coloured pink and the tRNA molecules coloured red. The protein components are shown in spacefill and coloured grey. The tRNA molecules occupy the E-, P- and A-sites as indicted. Note that the rRNA component represents the majority of the mass of the subunit (about two thirds).

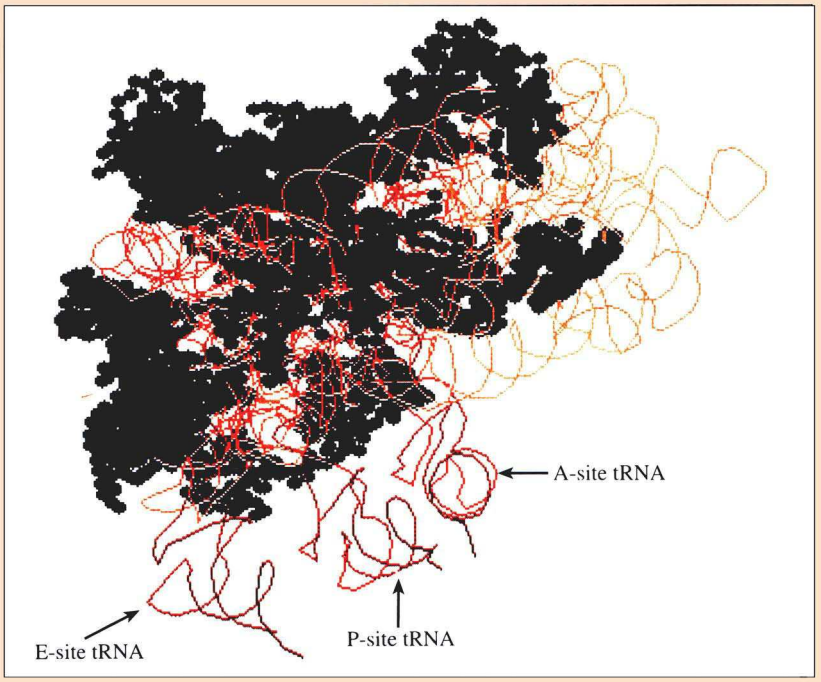


Figure 4.14 Small ribosomal subunit with three bound tRNA molecules

For clarity, the large subunit has not been shown. Its structure can be found in file 1GIY.

Termination of translation occurs when one of the three possible stop codons is reached. There are no tRNA molecules with anticodons complementary to these codons. Rather they are recognized by proteins called **release factors** (RFs). There are two of these. One, RF1, recognizes the codons UAA and UAG. The other, RF2, recognizes UAA and UGA. They appear to work by coordinating a water molecule which is transported to the peptidyl transferase centre of the ribosome, where it leads to hydrolysis of the ester linkage between the C-terminal amino acid and the tRNA to which it is attached. The protein then leaves the ribosome. Finally, a protein factor called **ribosome release factor** causes the ribosomal subunits to separate and the mRNA to be released.

Box 4.5 Protein Factors Involved in Protein Synthesis

Many protein factors that are not part of the ribosome are involved in protein synthesis. The first of these are the **initiation factors** IF1, IF2 and IF3. IF1 and IF3 form a complex with the small ribosomal subunit and prevent it from combining with the large subunit in the absence of mRNA; this would constitute an inactive complex. IF2 binds to *N*-formylmethionyl-$tRNA_f$, and the complex so formed finds and binds to the initiation codon. The three initiation factors are released when the 50S ribosomal subunit binds to the small subunit.

Next in the process comes a pair of **elongation factors** called EF-Tu and EF-Ts. EF-Tu collects aminoacyl-tRNA molecules from the synthetases and delivers them to the ribosome. One part of its role is to prevent the ester linkage joining the amino acid to the tRNA from being hydrolysed *en route* to the ribosome. The second part of its function is to help to ensure that the correct tRNA is delivered to the codon in the A-site. If the base pairing between the codon and anticodon is not correct, then EF-Tu will not release the aminoacyl-tRNA. The mechanism of action of EF-Tu depends on the fact that the free factor binds a molecule of GTP. When the correct codon is located, a conformational change in the protein results in hydrolysis of the GTP to GDP and in release of the aminoacyl-tRNA. The other elongation factor, EF-Ts, is required to promote the exchange of bound GDP in EF-Tu with GTP so as to return the EF-Tu to its

active form. It is important to note that EF-Tu does not bind to *N*-formylmethionyl-$tRNA_f$ and so *N*-formylmethionine is never delivered to the A-site. It does, however, bind to methionyl-$tRNA_m$ so that methionine can be incorporated into proteins in response to internal AUG codons.

As already mentioned in the text, another elongation factor, EF-G, is involved in the translocation of the mRNA through the ribosome once peptide bond formation has occurred. The structure of EF-G has been determined and is deposited in the PDB as entry 1DAR.[14] The molecule consists of two distinct parts. There is a globular part at the bottom which has a structure similar to that of EF-Tu; this part of the molecule has a binding site for GTP. The part at the top of the molecule resembles the shape of a tRNA molecule. After the peptide bond stage of elongation is complete, the globular part of the EF-G binds to the 50S ribosomal particle. The tRNA-like part interacts with the 30S particle adjacent to the A-site. When binding occurs, the GTP is hydrolysed to GDP, and this results in a conformational change in the protein in which the tRNA-like part moves to occupy the A-site of the 30S particle. This forces the tRNA carrying the growing protein into the P-site and pushes the mRNA one codon along. EF-G then leaves the ribosome, which is now ready to receive the next aminoacyl-tRNA molecule.

The final proteins involved in protein synthesis are the release factors. These have been described in the main text and will not be dealt with further here.

It should be noted that soon after protein synthesis has started, the region of the mRNA containing the Shine–Delgarno sequence emerges from the ribosome. Once that has happened, another small ribosomal subunit can bind to it and create a new translation complex. Hence, particularly under conditions of rapid protein synthesis, several ribosomes can be found translating the same mRNA. A mRNA with several ribosomes attached and actively synthesising protein is called a **polysome**.

Translation in eukaryotes is somewhat more complex than in prokaryotes, but the broad outlines are similar. One difference is that eukaryotes do not have an equivalent to the Shine–Delgarno sequence. Rather, the small ribosomal subunit binds to the 5′-cap region of the mRNA, and then works its way along until it reaches the first AUG triplet. This is then taken as the start signal for translation. This difference compared with prokaryotes reflects the fact that many prokaryotic mRNAs contain more than one message (that is, they are polycistronic)

and so must be able to bind ribosomes at several sites (see Box 4.6). Another difference is that eukaryotic translation starts with methionine rather than *N*-formylmethionine, although, as with prokaryotes, a special tRNA is used to recognize the initiation codon, and a different methionyl-tRNA recognizes internal methionine codons. The chemical events in chain elongation are essentially the same as in prokaryotes. There are differences in the protein factors involved in the various phases of translation, but these need not concern us.

Worked Problem 4.5

Q Protein synthesis is a complex process, and consequently it is quite slow. An average value for the rate of incorporation of amino acids into proteins is about 10 residues s^{-1}. Use this value to calculate the time required to synthesize a molecule of dystrophin.

A In Section 4.2 it was stated that the coding regions of the dystrophin gene occupy about 11 kbp. This means that the protein is about 3670 residues long. The time required to make it is, therefore, about 367 s, or 6.1 min.

Box 4.6 Operons

Many biochemical processes involve a sequence of reactions, each catalysed by a specific enzyme (see, for example, Scheme 1.2). In prokaryotes, it is commonly found that the enzymes for a particular pathway are coded by genes that are adjacent to one another in the genome, together with a region of DNA that encodes, as a minimum, a promoter and an **operator**. This group of regulatory and structural genes is called an **operon**. The promoter has the usual function of directing the start of RNA synthesis, but in an operon the product mRNA codes for all of the structural genes contained in the operon; it is referred to as a **polycistronic** mRNA. Note that the gene for each of the structural proteins has its own translation start and termination signals, so that the proteins are made individually, not as a single large protein product. The operator is a region of DNA between the promoter site and the first structural gene. This stretch of DNA binds a specific protein called a **repressor protein**, the function of which is to prevent transcription of the structural genes.

What is the point of all this? Firstly, having all of the genes for the enzymes involved in a particular pathway transcribed together ensures that they are synthesized in a coordinated manner and in equal quantities. Secondly, having the transcription of the structural genes under the control of an operator ensures that the enzymes are produced only when they are required. For example, the *lac* operon in *E. coli* encodes three proteins that are required for the metabolism of lactose. When the organism is growing in the absence of lactose, the genes are switched off. When lactose is present in the growth medium, the genes are switched on. To take another example, the biosynthesis of the amino acid tryptophan requires five enzymes, all of which are encoded by the *trp* operon. When there is sufficient tryptophan in the cell, the transcription of the operon is switched off. When the level of tryptophan drops, the transcription is switched on so that the enzymes are synthesized and more tryptophan can be produced. Note the fundamental difference between the *lac* and *trp* operons. The *lac* operon switches *on* in response to presence of a substance (lactose) in the cell; the *trp* operon switches *off* in response to the presence of a substance (tryptophan) in the cell. Nevertheless, the ways in which they work are basically similar.

So how do they work? We will look at the *lac* operon in *E. coli* as an example. This is understood largely as a result of work of Francois Jacob and Jacques Monod.[15] A diagram of the operon is given in Figure 4.15. The protein products of the operon are β-galactosidase (coded by the *lacZ* gene), lactose permease (coded by *lacY*) and thiogalactoside transacetylase (coded by *lacA*). The permease is a protein which spans the *E. coli* cell membrane and allows influx of lactose from the growth medium into the cell. The β-galactosidase is required to hydrolyse lactose into galactose and glucose, which are then metabolized further by other enzyme systems in the cell. The hydrolysis of lactose is shown in Scheme 4.9. The role of the transacetylase is not understood.

Monod and Jacob shared the Nobel Prize in Medicine in 1965 with André Lwoff "for their discoveries concerning genetic control of enzyme and virus synthesis".

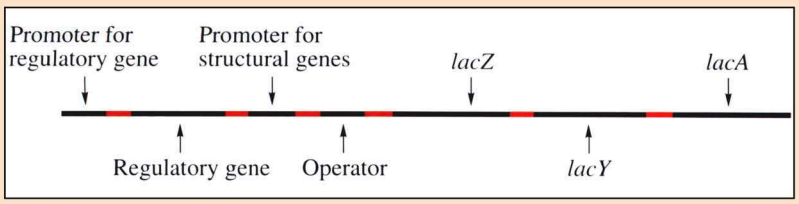
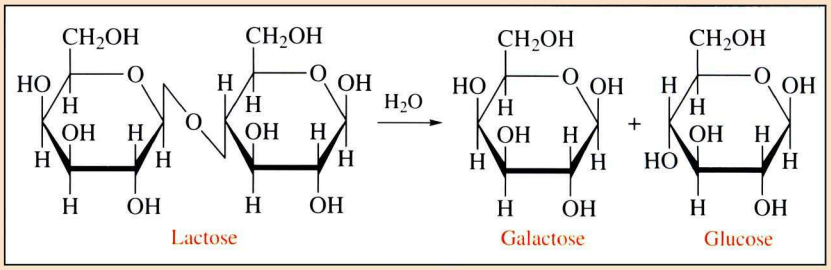
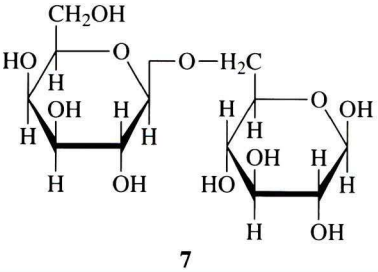


Figure 4.15 The *lac* operon in *E. coli*. The *red bars* are sections of DNA separating the functional regions of the operon. These sections and the functional elements are not drawn to scale

Scheme 4.9



The operon is a little more complicated than suggested above. In addition to the components already described, it contains at the 5′-end a regulatory gene with its own promoter. It is this regulatory gene that codes for the repressor protein. In the absence of lactose, the repressor protein binds tightly to the operator and blocks transcription of the downstream genes. When lactose is present in the cell, a small amount of it is converted to allolactose (**7**) which acts as an **inducer**. The inducer binds to the repressor protein. This results in a change in shape of the protein, the consequence of which is a dramatic decrease in the affinity of the repressor for the operator. The repressor detaches from the operator and allows transcription of the downstream genes to proceed. When lactose is no longer present, synthesis of allolactose ceases, the repressor regains its original conformation and re-binds to the operator. Transcription is then shut down.



7

Summary of Key Points

1. RNA can form double-helical hydrogen-bonded stem structures by folding back on itself in regions where sections of the chain are complementary to one another. These stems terminate in loops in which the bases are not hydrogen bonded. RNA can also form duplex helical structures with complementary strands of DNA.

2. Transcription is the process in which the genetic message in the coding strand of DNA is converted into a message encoded in RNA. The DNA strand transcribed is the complement of the coding strand.

3. RNA synthesis is catalysed by RNA polymerases. Promoter sites in the DNA are recognized by the RNA polymerase as signals for the start-point of transcription. RNA synthesis is in the 5′→3′ direction and continues until a termination signal is reached.

4. In eukaryotes, most genes are split into coding regions (exons) and non-coding regions (introns). The entire gene is transcribed into pre-mRNA and then the introns are spliced out to yield the mature mRNA molecule. Splicing is preceded by capping and by poly-A tailing.

5. RNA polymerases are also responsible for the synthesis of rRNA and tRNA molecules. Both classes of molecule are synthesized as parts of larger transcripts from which they are cut out by the action of ribonucleases.

6. tRNA molecules act as adapters between the amino acids and the codons in mRNA. Amino acids are linked to their specific tRNA molecules by *via* ester bonds between the carboxyl group of the amino acid and the hydroxyl group of the 3′-base of the tRNA. This reaction is catalysed by specific aminoacyl-tRNA synthetases.

7. Each tRNA molecule has a triplet of bases called the anticodon that recognizes the codon in mRNA by base pairing. Some tRNA molecules can recognize more than one codon by a process of "wobble", which arises from some latitude in the pairing of the 3′-base of the codon and the 5′-base of the anticodon.

8. Protein synthesis occurs on the ribosome. mRNA attaches to a small ribosomal subunit by base pairing between the Shine–Delgano sequence in mRNA and a complementary base sequence in 16S rRNA. Initiator $tRNA_f$ then binds to the initiator codon in the mRNA, and the large ribosomal subunit is recruited to form the complete translation complex with the initiator tRNA in the P-site. This initiator $tRNA_f$ is charged with methionine in the case of eukaryotes, and *N*-formylmethionine in the case of prokaryotes.

9. The charged tRNA specific for the next codon in the mRNA binds at a site on the ribosome (the A-site) adjacent to the initiator

tRNA. The amino group of the amino acid attached to this tRNA attacks the ester linkage between the initiator tRNA and the methionine (or *N*-formylmethionine) that it carries. The result is formation of a peptide bond between the two amino acids with the product dipeptide joined to the tRNA in the A-site.

10. Translocation of the mRNA by one codon through the ribosome occurs. This puts the initiator tRNA in the exit site, the tRNA bearing the dipeptide in the P-site, and leaves the A-site vacant. Another charged tRNA enters the A-site and peptide bond formation occurs. The process repeats until a termination signal is reached.

Problems

4.1. Access the file 1EFO in the RCSB database (see Worked Problem 3.1 for how to do this). From this file, retrieve the sequences of the RNA and DNA chains shown in Figure 4.2. Show how the sequences form a hydrogen-bonded duplex structure.

4.2. Assuming that RNA polymerase II incorporates nucleotides into a transcript at a rate of 50 s^{-1}, calculate how long it would take to transcribe the dystrophin gene.

4.3. Refer to Figure 4.5. Draw a detailed scheme showing the reactions leading to the cleavage of exon 1 from the pre-RNA molecule.

4.4. Refer to the legend of Figure 4.11. Draw the structures of the bases 2MG, 5MC and 1MA.

4.5. The amino acid cysteine is coded by two triplets of bases (Table 1.3) but requires only one tRNA. What is its anticodon sequence? The amino acid arginine is coded by six triplets of bases. How many tRNAs are required for this amino acid, and what are their anticodon sequences?

4.6. Produce a stereo model of the anticodon and the following wybutosine residue of the tRNA in Figure 4.10 (see Worked Problem 3.2 for how to do this).

4.7. Draw the structure of the aminoacyl adenylate intermediate in Scheme 4.6. Which base is present in the final product?

References

1. C. Sudarsanakumar, Y. Xiong and M. Sundaralingam, *J. Mol. Biol.*, 2000, **299**, 103.
2. G. Zhang, E. A. Campbell, L. Minakhin, C. Richter, K. Severinov and S. A. Darst, *Cell*, 1999, **98**, 811.
3. P. Cramer, D. A. Bushnell and R. D. Kornberg, *Science*, 2001, **292**, 1863.
4. L. T. Chow, R. E. Gelinas, T. R. Broker and R. J. Roberts, *Cell*, 1977, **12**, 1.
5. S. M. Berget, C. Moore and P. A. Sharp, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 3171.
6. T. R. Cech, A. J. Zaug and P. Grabowski, *Cell*, 1981, **27**, 487.
7. C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace and S. Altman, *Cell*, 1983, **35**, 849.
8. R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick and A. Zamir, *Science*, 1965, **147**, 1462.
9. S. R. Holbrook, J. L. Sussman, R. W. Warrant and S.-H. Kim, *J. Mol. Biol.*, 1978, **123**, 631.
10. F. H. C. Crick, *J. Mol. Biol.*, 1966, **19**, 548.
11. J. J. Perona, M. A. Rould and T. A. Steitz, *Biochemistry*, 1993, **32**, 8758.
12. M. Ruff, S. Krishnaswamy, M. Boeglin, A. Poterszman, A. Mitschler, A. Podjarny, B. Rees, J. C. Thierry and D. Moras, *Science*, 1991, **252**, 1682.
13. M. M. Yusupov, G. Z. Yusupova, A. Baucom, K. Lieberman, T. N. Earnest, J. H. D. Cate and H. F. Noller, *Science*, 2001, **292**, 883.
14. A. A. Evarsson, E. Brazhnikov, M. Garber, J. Zheltonosova, Y. Chirgadze, S. Al-Karadaghi, L. A. Svensson and A. Liljas, *EMBO J.*, 1994, **13**, 3669.
15. F. Jacob and J. Monod, *J. Mol. Biol.*, 1961, **3**, 318.

Further Reading

R. L. P. Adams, J. T. Knowler and D. P. Leader, *The Biochemistry of the Nucleic Acids*, 11th edn., Chapman and Hall, London, 1992, Chapters 9–12.

J. M. Berg, J. L. Tymoczko and L. Stryer, *Biochemistry*, 5th edn., Freeman, New York, 2002, Chapters 28 and 29.

R. F. Gesteland, T. R. Cech and J. F. Atkins (eds.), *The RNA World*, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY, 1999.

5 Modern Tools of DNA Analysis

Aims

By the end of this chapter you should understand:

- How DNA molecules are purified using electrophoresis
- The use of restriction enzymes to fragment DNA molecules at specific base sequences
- How DNA molecules are identified using blotting and hybridization
- How recombinant DNA molecules are made and how cloning is carried out
- How DNA can be amplified using the polymerase chain reaction
- How the base sequences of DNA molecules are determined
- How computers are used in the study of DNA
- How oligonucleotides are synthesized

5.1 Introduction: Recent Advances in DNA Technology

The story of DNA for the first 100 years was one of slow increases in understanding of its biological role and of its chemical structure. This culminated in the discovery of the double helical structure of DNA, and consequent elucidation of the way in which the genetic information is replicated and expressed. The rate of development in the subsequent years has been dramatic, not only in the amount of information that we have about DNA sequences, but also in the ways in which that knowledge has been applied to the benefit of humankind.

Ensuring a constant supply of a therapeutic protein is not the only reason for producing it using cloned genes. If the protein is

For example, it is now possible to **genetically engineer** bacteria so that they are able to produce human proteins (see Section 5.6). This has enormous importance for the production of therapeutic agents. To give

one example, **human growth hormone** is a protein of 217 amino acids which is produced by the pituitary gland. Underproduction of the hormone in children leads to **dwarfism** arising from failure of the long bones to develop normally. The condition can be treated by injection of growth hormone, but it is very difficult to obtain human pituitary glands from which to isolate the protein, so the number of children that could be treated in the past was very small. Now, however, the gene coding for human growth hormone has been inserted into bacterial cells (or **cloned** into those cells, to use the modern terminology). These cells can be grown in culture, and provide a virtually limitless source of growth hormone. Many therapeutic agents are now produced by this route.

isolated from human tissue there is always a possibility that the product will be contaminated with viruses, and in particular with HIV. The problems arising from injecting a contaminated agent into a patient are obvious. Use of cloned proteins does not carry this risk.

Genetic engineering is not restricted to micro-organisms. More recently, there have been great advances in the ability to genetically modify plants. The objectives of this include increasing productivity of crop plants, introduction of pest resistance thus reducing the need for artificial pesticides, developing crop plants that will grow in hostile environments, and modifying plants to compensate for nutritional deficiencies. For example, about 1 million children in the third world die every year, and 5 million become blind, because of lack of vitamin A in their diet. This arises from the fact that their staple food is rice, which lacks vitamin A. A strain of rice has now been engineered that synthesizes vitamin A (and also has increased iron content), the use of which will help to solve these problems.

In a different area, **DNA fingerprinting** (see Box 5.10) is now a widely used technique in forensic science. The method has been developed to the point where the DNA from the follicle of a single hair from a crime scene can be used to unambiguously identify the perpetrator of a murder or a rape. Equally important, DNA fingerprints obtained from archived scene-of-crime material have now been used to put right several cases of miscarriage of justice where the wrong person had been convicted.

Most recently, the DNA sequences of complete genomes have been determined. These include the genomes of many bacteria, some of them pathogens that are responsible for human disease. Knowledge of their genome structures will lead to advances in ways of combating infection. Similarly, the genome sequence of the malaria parasite, *Plasmodium falciparum*, has recently been completed, as has that of its primary mosquito host, *Anopheles gambiae*. Malaria kills over 1 million people every year, the majority of them children in sub-Saharan Africa, and over 500 million people suffer from the disease. The hope is that knowledge of the genome sequences will provide pointers towards preventative strategies. Already, for example, a genetically modified form of the related mosquito *A. stephensi*, whose parasite *P. berghei* causes malaria in mice, has been constructed. The modified mosquito is unable to transmit the parasite. Perhaps it will prove possible to do the same thing with

A. gambiae and produce a mutant which, if released into the wild, would supplant the wild-type population. However, it is early days yet.

Perhaps most public interest focuses on the structure of the human genome, a "first draft" of which was published in 2001. This has already provided a wealth of information, and much more will come in the future as the structure of the genome is examined in detail. A very large amount of work is now being done on mapping differences between the genomic DNA of different individuals. Most of these differences are **single nucleotide polymorphisms** (or **SNPs**, pronounced *snips*); that is, differences in single residues in particular positions. It is estimated that there are about 10 million SNPs in the human genome, and it is largely from these that individual differences in characteristics such as susceptibility to diseases arise. The prospect is that one day it will be possible to predict from an individual's DNA profile which diseases he or she is likely to get in the future, and to take preventative action to avoid them.

In fact, two drafts were published. One was produced by the International Human Genome Sequencing Consortium (IHGSC) led by John Sulston, and the other by a biotechnology company called Celera. A nearly complete (>99% in coding regions) version of the sequence is now available on the website of the University of California, Santa Cruz (http://genome.ucsc.edu).
Sulston shared the 2002 Nobel Prize in Physiology or Medicine with Sydney Brenner and Robert Horvitz, but this was not for genome sequencing. Rather, it was "for their discoveries concerning genetic regulation of organ development and programmed cell death". They studied the nematode worm *Caenorhabditis elegans* and followed cell division and differentiation during the life cycle of the organism. This allowed them to identify key genes regulating organ development and programmed cell death.

One exciting area that has already developed very considerably from comparison of the patterns of differences in the genomes of human populations is the study of the history of mankind's evolution. It is now clear that all existing human populations are descended from a small group of people who lived in Africa a little more than 100,000 years ago, and it is possible to trace the migration of their descendants out of Africa to their present locations. We cannot go into the details of this fascinating story, but if you are interested in your evolutionary history, then you cannot do better than read the account given by Steve Olson.[1]

Box 5.1 Gene Therapy

One of the great hopes for application of the knowledge obtained from the human genome sequence is that it will lead to development of effective gene therapies. Many illnesses are genetic in origin (see Box 1.4), and the prospect is that treatment may be possible by replacing the defective gene in samples of the patient's own cells, and then returning those cells to the body. Using the patient's' own cells removes all the problems associated with immune rejection of cells from a donor.

The problem is how to insert a corrected version of the defective gene into the chromosomes of the patient. So far, the method that has proved most effective is by using modified retroviruses. The virus is stripped of most of its own genes, so as to avoid deleterious effects arising from viral infection, and the gene for the function that the patient is lacking is then inserted into the residual viral genome. Infection of the patient's cells with this modified virus then results in insertion of the viral RNA transcript into the cellular genome and, hopefully, expression of the desired protein. The cells are then

re-introduced into the patient where they divide in the normal way, thus correcting the genetic defect.

So far, the best results from use of this technique have been obtained with a disease called **severe combined immune deficiency** (**SCID**). One form of this disease is X-linked, and results in failure of the sufferer to develop the specialized cells required to mount an immune response and to fight off infection. A pioneering treatment for this condition was developed by Alain Fischer and Marina Cavazzana-Calvo working in France. They treated 11 children with the condition by introducing the required gene into their bone marrow cells and then replacing the cells. Of the 11 children treated, 9 were cured of the disease, and the treatment was subsequently adopted by other doctors.

Unfortunately, one of the children in the French study has subsequently developed leukaemia-like symptoms, apparently as a result of the treatment. The problem seems to be that there is no way of controlling where in the patient's DNA the retroviral genome inserts itself. If it inserts in the middle of another gene, and in particular one that is involved in regulation of cell division, then cancer could result. That is what is thought to have happened here.

This is clearly a very serious setback for gene therapy, but one that will no doubt be overcome. Possibly this will be done by developing methods for controlling the sites at which retroviral genomes insert into the recipient's DNA. Alternatively, new ways of delivering the required gene to the patient's DNA may be devised. Whichever of these may be the case, there is little doubt that in the future gene therapy will offer the prospect of good health and a normal lifespan to people born with potentially fatal genetic diseases.

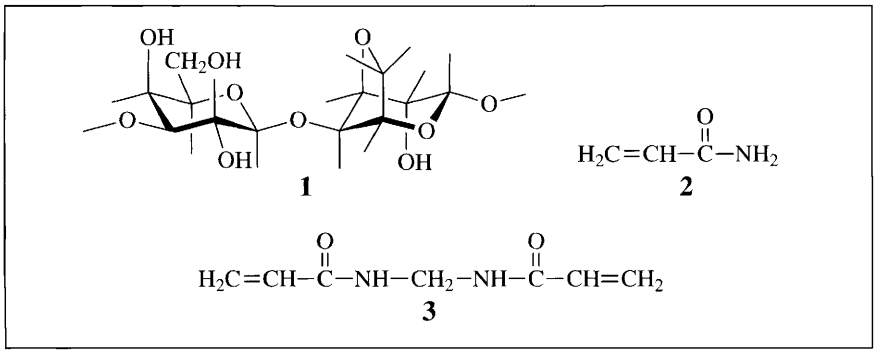
By any standards, this (very partial) list of advances in DNA technology over the last few decades is remarkable, and the question arises as to how so much has been done in a relatively short time. The answer lies essentially in the development of new experimental tools. It is a general phenomenon that rapid advances in scientific knowledge are made as a result of discovery of new experimental techniques. A description of what those tools are in the case of DNA studies, and how they are used, is the focus of this chapter.

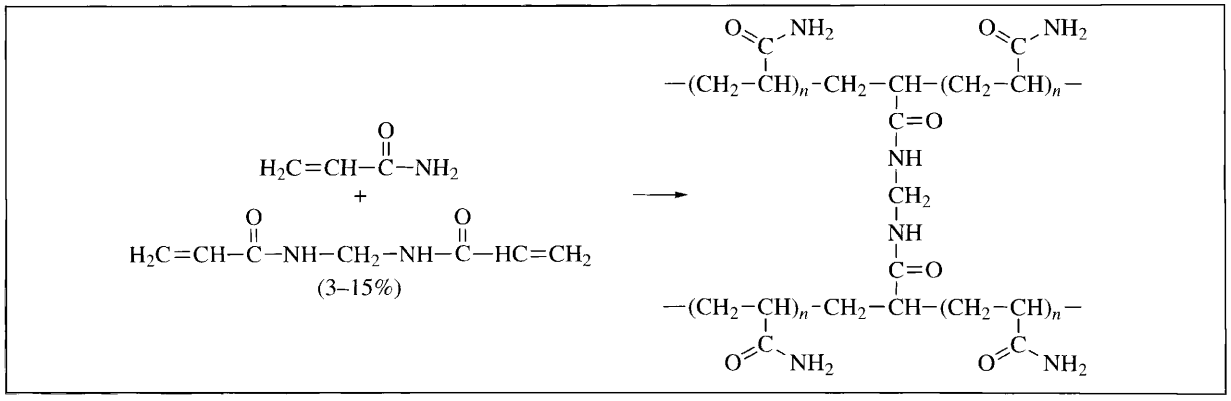
5.2 Gel Electrophoresis

Gel electrophoresis is a key technique for both the purification and the analysis of DNA samples. The term electrophoresis is used to describe

Polyacrylamide gel electrophoresis (PAGE) is used for DNA molecules between 2 and 1000 bp long. Agarose gels can be used to separate molecules from 200 bp to 100 kbp long. Recently, a technique called **pulsed field electrophoresis** has been developed by which DNA molecules greater than 1 Mbp can be separated. The method involves inverting the direction of the applied field at regular intervals, and probably works on the basis that smaller molecules re-orientate themselves more rapidly as the field direction changes, although there is still debate about this. The important point is that the method allows for separation of chromosomal DNA molecules.

the movement of charged molecules in an electric field. In practice, the molecules are made to move through the liquid phase of a **gel** which, depending on the particular application, may be made of **agarose** or of **polyacrylamide**. Agarose is a polysaccharide obtained from seaweed; the structure of the monomeric unit is shown in **1**. If agarose is dissolved in hot water and then allowed to cool, the polysaccharide chains interact to form a gel, the pore size of which depends on the concentration of agarose used. Acrylamide has the structure shown in **2** (the proper name of this compound is **propenamide**, but you would search in vain in the index of any textbook of biochemistry or molecular biology for a reference to that name or to polypropenamide, and so we will not use them here). It undergoes free-radical polymerization to form a linear polymer, but the resulting material lacks mechanical strength. If, on the other hand, some *N,N′*-methylenebisacrylamide (**3**) is included in the reaction, then the polyacrylamide chains are cross linked as shown in Scheme 5.1. The degree of cross linking, and hence the pore size of the gel, depends on the percentage of the bisacrylamide included in the reaction mixture.

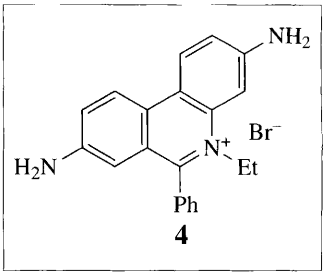




Scheme 5.1

In both cases, the gel is made in a buffer solution at neutral or alkaline pH, under which conditions nucleic acids carry a net negative charge. The gel is cast between rectangular plates and small wells are cut to receive the samples. An electric field is then applied, with the negative pole at the top and the positive pole at the bottom. The individual components of the mixture of nucleic acids will move down the gel at rates dependent on their sizes. The smaller the molecule, the faster it moves; this is essentially because small molecules can move more easily through the pores of the gel than can large ones. It turns out that there is a linear relationship between the distance moved and the logarithm of the length of the DNA molecule.

At the end of the experiment, it is necessary to visualize the nucleic acid bands in the gel. There are several ways in which this can be done, but the easiest is to incubate the gel in a solution of ethidium bromide (**4**) and then to wash out the excess. Ethidium bromide intercalates between the bases of the DNA (see Section 3.6.1). In this non-polar environment, ethidium bromide is fluorescent, and so the bands of DNA can then be seen by exposing the gel to UV light. A diagrammatic representation of an agarose gel is shown in Figure 5.1. A gel with only two wells is shown; in practice, gels with 20 or more wells can be prepared, which allows for comparison of multiple DNA samples (an actual example of an agarose gel is shown in Figure 5.8).



4

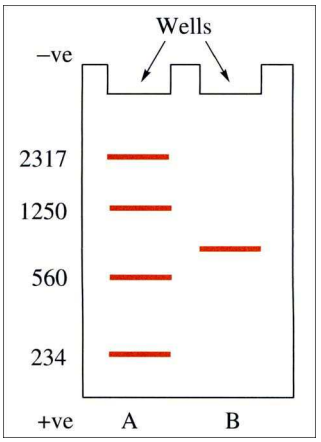


Figure 5.1 Agarose gel electrophoresis. Track A contains DNA molecules of known lengths (given in bp at the side). Track B contains a molecule of unknown length

Worked Problem 5.1

Q From the distances of migration of the bands in track A of Figure 5.1, obtain a value for the size of the DNA molecule in track B.

A Figure 5.2 shows a plot of the relative distances moved by the components in track A (that is, the ratio of the distance moved from the bottom of the well divided by the length of the gel), against the logarithm of the number of base pairs. The relative distance moved by the molecule in track B is 0.50, which corresponds to a log (length) of 2.91, and hence to a length of 813 bp.

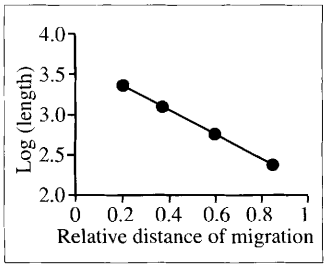


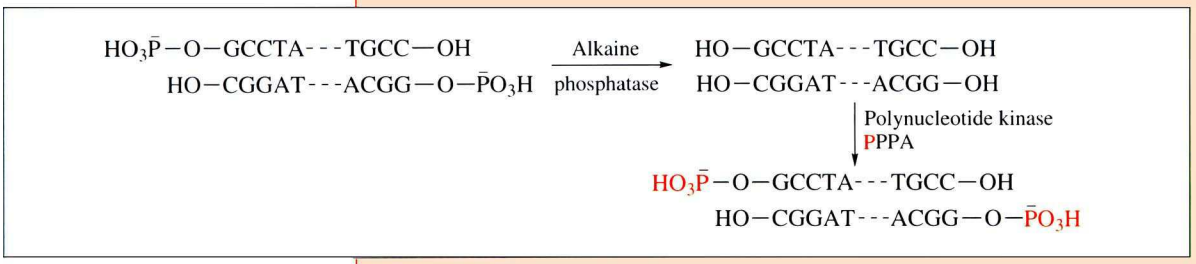
Figure 5.2 Plot of relative distance of migration against logarithm of the length for the bands in track A of Figure 5.1

Box 5.2 Autoradiography

Detection of DNA on gels using ethidium bromide is quite sensitive, and loadings of less than 100 ng are routinely used. Sometimes, however, it is necessary to work on a smaller scale than

this, and in such cases **autoradiography** provides a solution. The DNA sample to be analysed is radiolabelled with ^{32}P. After electrophoresis, the gel is layered onto a sheet of X-ray sensitive photographic paper and left for an appropriate time. The radiation from the bands of DNA exposes the film, and on development of the film the banding pattern is revealed. With very small amounts of DNA the time of exposure can be increased until acceptable levels of exposure are achieved.

Clearly the use of this technique requires methods for labelling of the DNA. One method is end-labelling. The DNA sample is first treated with alkaline phosphatase, which removes phosphate from the 5′-end of the DNA molecule should it be present (this will depend on how the DNA specimen has been obtained). The product of this reaction is then incubated with γ-^{32}P-labelled ATP and **polynucleotide kinase**. The enzyme transfers the terminal phosphate from ATP onto the 5′-OH of the DNA and so produces a radioactive product. The process is summarized in Scheme 5.2. The ATP is abbreviated to PPPA with the radioactive phosphate in red.



Scheme 5.2

The down-side of autoradiographic methods is, of course, the dangers inherent in the handling of radioactive materials. For this reason, the method is avoided if possible. It should be pointed out, however, that use of ethidium bromide is also potentially hazardous because the substance is a powerful mutagen arising from its ability to intercalate into DNA.

A slightly more complicated method of labelling is by **nick translation**. Here, the DNA is treated briefly with a deoxyribonuclease which introduces single-stranded breaks (nicks) in the double stranded molecule. The 5′-phosphate ends produced are substrates for the 5′→3′ exonuclease activity of DNA polymerase I (see Section 3.5). This enzyme removes residues from the DNA strand but also, in the presence of a mixture of deoxynucleoside triphosphates, mends the gap so created by addition of nucleotides to the 3′-OH end of the nick. If the deoxynucleoside triphosphates used are α-^{32}P-labelled, then the repaired DNA will be radiolabelled. This method is preferable to end-labelling because a higher level of radioactivity can be achieved. If a radiolabel of high specific activity is used, then detection of bands after electrophoresis can be done at the sub-nanogram level.

5.3 Restriction Enzymes

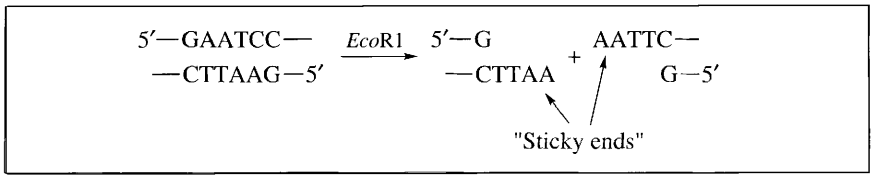
In the previous section, we considered the use of gel electrophoresis to separate DNA fragments without saying where those fragments come from. The answer is by digestion of larger DNA molecules using **restriction enzymes**. The restriction enzymes are endonucleases which cut DNA chains at highly specific sequences, and their discovery marked a major leap forward in DNA chemistry because they provide a tool for producing defined pieces of DNA. They have been colourfully referred to as "molecular scissors" because they allow the DNA to be cut according to the pattern provided by the sequence of the molecule.

The restriction enzymes were discovered in the late 1960s by scientists who were interested in a process called **host-controlled restriction**. What this means is that when phage propagated on one strain of bacteria are transferred to a different strain, the ability of the phage to infect the new strain is often severely restricted. The phenomenon was studied by groups lead by Werner Arber, Daniel Nathans and Hamilton Smith, and it was shown that restriction occurs because the bacteria produce enzymes (restriction enzymes) that degrade the phage DNA and so prevent it from being replicated. The DNA of the host bacterium is not degraded because some of the bases in the site recognized by the restriction enzyme are methylated and this prevents the activity of the enzyme. A few copies of the phage DNA also become methylated and so escape digestion. These can lead to the production of new phage, which is why the infection is only *restricted* rather than *prevented*.

Arber, Nathans and Smith were awarded the Nobel Prize in Physiology or Medicine in 1978 "for the discovery of restriction enzymes and their application to problems of molecular genetics".

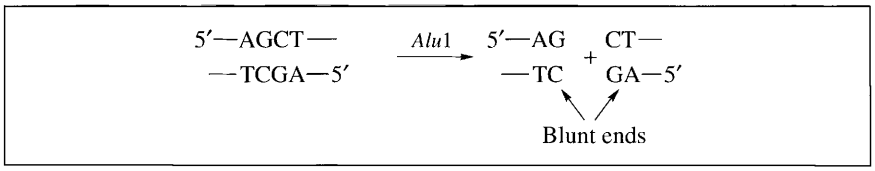
There are in fact three different types of restriction enzyme that work in somewhat different ways, but only one type, the **Type II enzymes**, are used in DNA studies and we will concern ourselves only with those. The first such enzyme was isolated by Smith and Wilcox,[2] but now many different ones are known. Typical, and very widely used, is the enzyme ***Eco*RI**, so called because it was isolated from *E. coli* and was the first such enzyme isolated from that source. The recognition sequence of the enzyme is 5′-GAATTC-3′ and the way in which it acts is shown in Scheme 5.3. The essential feature of its action is that it cuts both chains, but not at points opposite one another. Rather, it cuts between residues 1 and 2 in the 5′→3′ chain, and residues 5 and 6 in the 3′→5′ chain. This produces two chains, each with four unpaired residues. These sections of single-stranded DNA are known as **sticky ends** for reasons that will become apparent in Section 5.5. The recognition sequence of *Eco*R1 has the special feature that it is **palindromic**; that is the sequences of the two chains read 5′→3′ are the same. Put another way, the recognition site has two-fold symmetry. This is a feature of nearly all Type II enzymes, and is related to the way that the site is recognized by the restriction enzyme.

The structure of a restriction enzyme, *Bgl* (see Table 5.1), in complex with a 16mer substrate (5′-TATTATAGATCTATAA-3′) has recently been determined.[3] The structure can be found as entry 1D2I in the PDB. The protein has two identical monomers related by a two-fold axis of rotation vertically through the centre of the dimer. Hence the symmetry of the cleavage site in the DNA molecule is mirrored by the symmetry of the protein. One monomer catalyses hydrolysis of one DNA strand, and the other catalyses hydrolysis of the other strand.



Scheme 5.3

Not all restriction enzymes produce sticky ends, and not all of them have a six-base recognition site. For example, *Alu*I recognizes the sequence 5′-AGCT-3′ and cuts in the middle of the sequence as shown in Scheme 5.4. In this case the product fragments have **blunt ends**. The enzyme *Not*I, on the other hand, has an eight-base recognition site and produces sticky ends. Some of the most commonly used restriction enzymes are listed in Table 5.1.



Scheme 5.4

Table 5.1 Commonly used restriction enzymes. Only the sequences of the 5′→3′ strands are shown. The cleavage point is indicated with a vertical line (|)

Enzyme	*Recognition sequence*	*Organism*	*Ends*
*Alu*I	AG|CT	*Arthrobacter luteus*	Blunt
*Taq*I	T|CGA	*Thermus aquaticus*	Sticky
*Hae*III	GG|CC	*Haemophilus aegyptius*	Blunt
*Hin*fI	G|ANTC[a]	*Haemophilus influenzae* Rf	Sticky
*Eco*RI	G|AATTC	*Escherichia coli*	Sticky
*Bam*HI	G|GATCC	*Bacillus amyloliquefaciens*	Sticky
*Bgl*II	A|GATCT	*Bacillus globigii*	Sticky
*Hin*dIII	A|AGCTT	*Haemophilus influenzae* Rd	Sticky
*Pvu*I	C|GATCG	*Proteus vulgaris*	Sticky
*Pvu*II	CAG|CTG	*Proteus vulgaris*	Blunt
*Sma*I	CCC|GGG	*Serratia marcescens*	Blunt
*Not*I	GC|GGCCGC	*Nocardia otitidis-caviarum*	Sticky

[a]The N in this sequence represents any base, so *Hin*fI has four recognition sequences

The enormous importance of the restriction enzymes is that they are completely specific and so can be used to cut large DNA molecules in a very precise way into pieces that are useful for other applications such as cloning (see Section 5.6) and sequencing (see Section 5.8).

Plasmids are small, circular DNA molecules found in many species of bacteria. They have their own origins of replication, and so are replicated independently of the chromosomal DNA (referred to as

An example of their use might be helpful. Consider the results reported in Figure 5.1. How were the pieces of DNA of defined lengths in track A

obtained? The answer is that they were produced by the combined action of two restriction enzymes, *Bam*HI and *Bgl*I, on a **plasmid** called pBR322.

The plasmid is a circular DNA molecule 4361 bp in length, and a diagram showing some of its features is given in Figure 5.3. The numbering of the bases is done by reference to the unique *Eco*RI site; base number 1 is taken as the first T in this recognition site. A description of the plasmid, and a list of all the restriction sites that it contains, are available at http://www.fermentas.com/techinfo/NucleicAcids/mappbr322.htm.

The positions of the recognition sites for *Bam*HI and *Bgl*I are as shown. The latter enzyme has the recognition site 5′-GCCNNNN|NGGC-3′ where, as usual, N signifies any nucleotide. The numbers given for the restriction sites in Figure 5.3 are those for the first base in the site. Hence cleavage at the *Bam*HI site occurs after base 375 and at the first *Bgl*I site at base 935 (if you find this confusing, reference to Figure 5.22 which gives the complete sequence of pBR322, with the restriction sites in red, should help). The restriction fragment produced is therefore 560 bp long, as shown in Figure 5.1. You can check that the lengths given for the other fragments are correct.

autonomous replication). Many plasmids carry genes that confer antibiotic resistance on the host organism. They can be transferred between one bacterium and another, and so are responsible for some cases of appearance of new drug resistance in previously susceptible micro-organisms. Plasmids carrying genes that confer resistance to several drugs have evolved as a result of the increased use of antibiotics. Organisms carrying such plasmids are a major problem, particularly in hospitals.

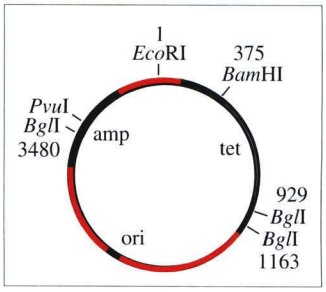


Figure 5.3 Map of pBR322 showing the positions of the origin of replication (ori), the genes for resistance to tetracycline (tet) and ampicillin (amp), and selected restriction sites

5.4 Blotting and Hybridization

Suppose that we want to isolate a particular DNA fragment from a restriction enzyme digest, and further suppose that we know at least a little of its base sequence. How do we proceed? We can separate the fragments by gel electrophoresis, but the problem is how to identify the piece of DNA that we want. The answer to this is provided by **hybridization**. A small piece of single-stranded DNA (ssDNA), or **oligonucleotide**, can be synthesized (how this is done is described in Section 5.10) that is complementary to a section in one of the chains of the DNA molecule we wish to find in the gel. Because the synthetic oligonucleotide has a sequence complementary to that of a section of the target DNA, it will bind to it and form a hybrid molecule.

Note that these molecules should properly be called **oligodeoxyribonucleotides**, but because the context of the discussion is DNA chemistry the abbreviated form oligonucleotide can be used as a shorthand. They are frequently referred to by the even shorter name of **oligos**.

The problem is that it is not possible to do this directly, because the gels used for electrophoresis are not sufficiently stable mechanically to allow the required manipulations to be carried out. The solution to this problem was developed by Edward Southern in 1975.[4] The technique essentially involves blotting of the DNA fragments from the gel onto a mechanically and chemically stable membrane made of nylon or nitrocellulose. In recognition of its inventor the method is known as **Southern transfer**, or sometimes as **Southern blotting**. It is carried out as shown in Figure 5.4. The gel is placed on a paper wick dipping into a reservoir of alkaline buffer, and overlaid with the membrane. On top of the membrane is

placed a stack of paper towels pressed down by a weight. The buffer is sucked up through the paper towels by capillary action, and carries the DNA molecules onto the membrane. After transfer, the membrane is heated to fix the DNA molecules in place and to promote their dissociation into single strands. The pattern of DNA molecules on the membrane will be a replica of the pattern in the gel, and the molecule of interest can be located by probing with the synthetic oligonucleotide. Once it is located, the appropriate band can be eluted from a duplicate gel and will lead to a purified sample of the DNA molecule required.

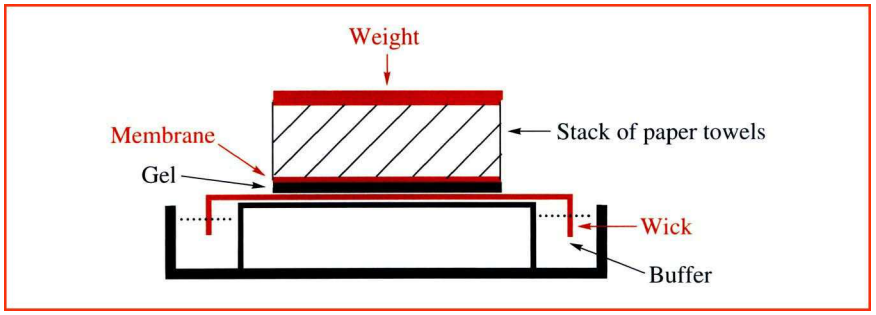


Figure 5.4 Experimental set-up for Southern blotting

Variants of the technique have been developed for transfer from gels of both RNA (**northern blotting**) and proteins (**western blotting**). There is no technique of eastern blotting!

How long should the oligonucleotide be so as to ensure that it is specific for the DNA of interest? If it is too small, then it is quite likely that the sequence will occur by chance in some other DNA molecule. In addition, to be sure that the probe binds only to the piece of DNA to which it is completely complementary, we need to use stringent hybridization conditions; in practice, this means carrying out the hybridization at temperature just below the T_m for the duplex formed between the probe and the target DNA. Larger probes allow more stringent conditions to be used so that duplexes with one or two mismatched bases are not allowed to form. There is no firm rule about the optimum size, but oligonucleotides in the range of 15 to 20 residues long are routinely used. The temperature required for the hybridization step can be worked out using equation (5.1), which relates T_m to the number of residues of each base in the oligonucleotide:

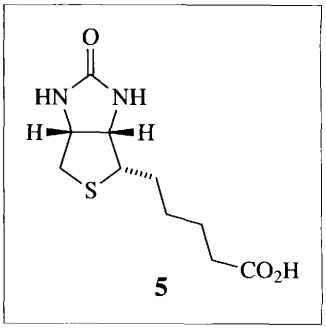
$$T_m = (4[\mathrm{G} + \mathrm{C}] + 2[\mathrm{A} + \mathrm{T}])\ ^\circ\mathrm{C} \qquad (5.1)$$

Worked Problem 5.2

Q Suppose that a 15mer is used as a probe for detection of a particular piece of DNA on a Southern blot. What is the chance that the same sequence will occur in another unrelated DNA molecule?

A Since there are four possible bases at each position, the probability that the sequence will occur by chance is 1 in 4^{15}, or 1 in about 1.07×10^9. Hence its occurrence by chance would be an unlikely event except in very large chromosomes.

It is necessary, of course, to be able to detect the position of the hybridized oligonucleotide on the membrane. One way of doing this is to incorporate a radiolabelled residue into the oligonucleotide and then to carry out autoradiography (see Box 5.2). Because of the dangers involved in using radioisotopes, this method is decreasing in popularity. A very good alternative is to couple **biotin** to the probe. Biotin has the structure shown in **5**. The value of doing this is that biotin binds extremely strongly and specifically to a protein molecule called **avidin**. After the blotting procedure has been carried out, the membrane is treated with the biotin-linked oligonucleotide and then washed to remove any unhybridized material. The membrane is then incubated with avidin to which a fluorescent tag has been covalently linked. The position of the desired band can then be seen by viewing the membrane under UV light.



5

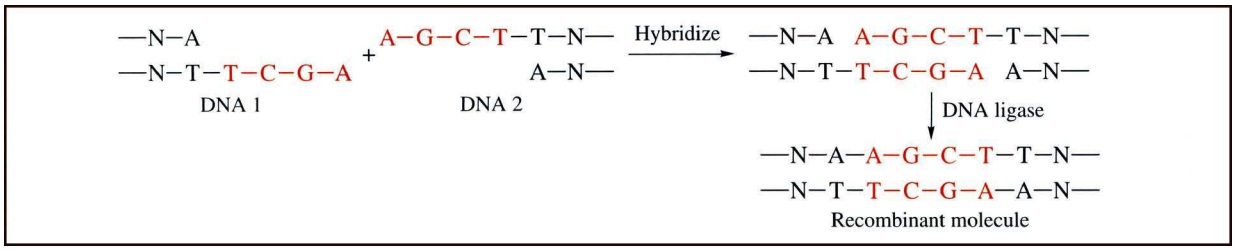
A somewhat different situation frequently arises where we know the amino acid sequence of a protein and wish to isolate the DNA that codes for it. It might seem at first sight that this is straightforward, because we know the genetic code and so we can work backwards from a section of the amino acid sequence to the base sequence that codes for it, and synthesize an appropriate oligonucleotide to use as a probe. However, this ignores the degeneracy of the code. Recall that most amino acids are encoded by at least two codons, and sometimes by as many as six. So knowing the amino acid sequence is not the same as knowing the actual base sequence that codes for it. The only approach is to use a mixture of oligonucleotides that contains all the possible coding sequences for the piece of protein selected; one of them will have the correct base sequence to hybridize with the DNA. Clearly, the number of possible coding sequences, and hence the number of individual oligonucleotides required in the mixture, may be very large. Suppose that we select a sequence of five amino acids, each of which is specified by four codons. In this case we will need a mixture of 1024 (4^5) different 15mers to be sure that the correct sequence is represented. The problem can be simplified by looking for a region of the amino acid sequence which contains amino acids coded by only one or two codons. Methionine and tryptophan are particularly desirable because they are each specified by a single codon (see Table 1.3). Reference to Problem 5.1 shows the value of working with a section of the protein containing those amino acids.

5.5 Making Recombinant DNA Molecules

The initial development of recombinant DNA technology was due to Paul Berg and his co-workers.[5] Berg was awarded half the Nobel Prize in Chemistry 1980 for "his fundamental studies of the biochemistry of nucleic acids, with particular regard to recombinant DNA".

For many purposes, some of which will be described in the next section, it is necessary to join DNA molecules together. This process is referred to as **ligation**, and the product is a **recombinant DNA** molecule. Key to this process is the enzyme DNA ligase, an enzyme which has already been described in Section 3.5 in the context of DNA replication. The essential property of the enzyme is that it can form the phosphodiester bond between two adjacent nucleotides in a DNA chain. The question is how to bring the ends of the DNA molecules that we wish to join together adjacent to one another. The answer is provided by the sticky ends that are produced by some restriction enzymes.

Suppose, for example, that the two pieces of DNA that we wish to join have been produced by fragmentation using *Hin*dIII (see Table 5.1). In this case the situation will be as shown in Scheme 5.5, where the line joining the nucleotides represents the phosphodiester linkage.



Scheme 5.5

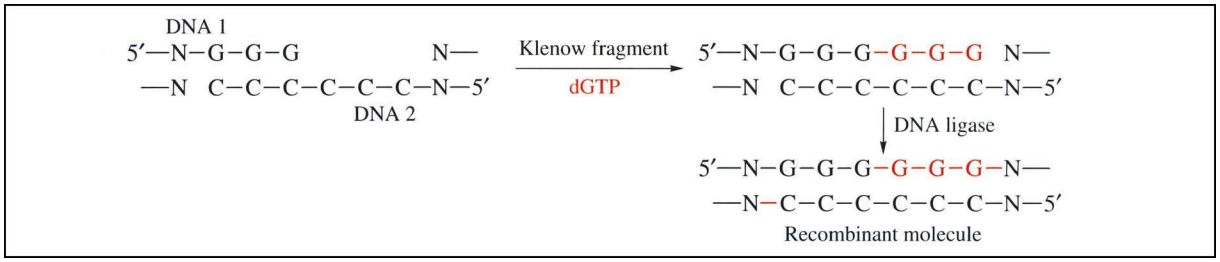
One DNA molecule (DNA1 in Scheme 5.5) will have a sticky overhang with the sequence TCGA at its 3′-end (shown in red). The second DNA molecule will have an overhang with the sequence AGCT at its 5′-end. These, of course, are complementary and will hybridize (hence the name "sticky end"). At this stage, the phosphodiester linkage between the A residues is missing and the fragments are held together only by base pairing. If DNA ligase is added, however, the linkages are formed and the two original molecules are now covalently linked into a single recombinant molecule.

Making recombinant molecules from fragments with blunt ends is more difficult. Although DNA ligase can catalyse the necessary reactions to join blunt ended molecules, the process is very inefficient. The reason is that the concentration of DNA fragments will be very low, and so the chance coming together of the fragments for ligation has a very low probability. The approach to this problem is to put sticky ends on the blunt ended molecules.

One way of doing this is to use the enzyme **terminal transferase**, which adds nucleotides on to the 3′-ends of DNA chains. So, for example, if one DNA fragment to be ligated is incubated with the enzyme in the presence of dGTP, a string of Gs will be added. If the other fragment is incubated with the enzyme and dCTP, a string of Cs will be added. These homopolymers will then act as sticky ends. There is an extra problem here in that the tails are unlikely to be the same length so that, in addition to ligation, it will be necessary to fill in any gaps. The enzyme used for this purpose is the **Klenow fragment** of DNA polymerase I.

Recall that DNA polymerase I has both a polymerase and a nuclease activity (see Section 3.5). Here, we want only the polymerase activity. It turns out that the nuclease activity of the enzyme is associated with the N-terminal region of the molecule. Removal of this region abolishes the nuclease activity, but leaves the polymerase activity intact. This truncated enzyme is what is referred to as the **Klenow fragment**.

The process is summarized in Scheme 5.6. It is assumed that the G-tail is shorter than the C-tail (note that very short tail lengths are shown for convenience; in practice the tails might be 10–20 residues long). The Klenow fragment is used to fill in the missing Gs, and then DNA ligase links the chains together.



Scheme 5.6

5.6 Cloning

Cloning is the name given to the process by which foreign DNA molecules are introduced into living cells, although, as we shall see, the term properly applies to only a part of the process. Cloning was originally carried out using bacterial cells as the host, and we will concentrate mainly on that process. A cloning experiment proceeds it the following stages:

- The foreign DNA is incorporated into a **vector**, frequently a plasmid, using recombinant DNA technology
- The recombinant vector is introduced into the host bacterial cells, which are then spread thinly on a nutrient medium on which they can grow and divide. As this occurs, the vector with its inserted DNA is replicated and passed on to the daughter cells
- Each of the bacterial cells multiplies and produces a **colony** or **clone**. The clone originated from a single bacterial cell, and so all the cells in it express only a single type of recombinant DNA molecule

The first functional plasmids for use in cloning were constructed by Herbert Boyer, Stanley Cohen and their co-workers.[6]

The importance of this technology lies in the following features. Firstly, it allows for the production of large amounts of a particular DNA molecule. We have already seen that DNA can be purified using gel electrophoresis, but the amounts that can be obtained by this route are small – typically a few nanograms. Once a clone containing the DNA of interest is obtained, the bacterial cells can be grown in culture, and it is then easy to obtain the cloned DNA in milligram amounts. Secondly, it provides a method of purifying the DNA. It is quite likely that the DNA sample obtained by elution from a gel may not be pure. After all, the method separates molecules based only on size, and it may be that two or more DNA molecules migrate at a very similar rate and so are extracted together. This will mean that more than one sort of recombinant plasmid is produced at the start of the cloning experiment. This is not, however, a problem. Each of the clones obtained will have originated from a bacterial cell containing only one sort of plasmid, and so it is simply a matter of identifying a clone that contains the plasmid with the DNA we want.

A final benefit of cloning, but one of a rather different sort, arises if a special type of vector called an **expression vector** is used. Expression vectors contain not only an origin of replication that allows them to be copied, but also a set of expression signals that allow the inserted DNA to be transcribed, and the product RNA translated. This means that if the inserted DNA codes for a protein, the bacterial cell will now synthesize the foreign protein. This is the way in which proteins are produced by genetic engineering, the importance of which was described in Section 5.1.

To see how cloning is actually carried out, we will take as an example the use of the plasmid pBR322, which was one of the earliest plasmids developed and still one of the most popular.

Like all useful plasmid vectors, pBR322 is not a naturally occurring molecule. It was constructed starting from a natural plasmid found in *E. coli* and then, using recombinant DNA techniques, engineered to produce its desirable features. It was made by Bolivar and Rodriguez[7] (hence its name: "p" for plasmid, "BR" for Bolivar and Rodriguez, and 322 for the particular version).

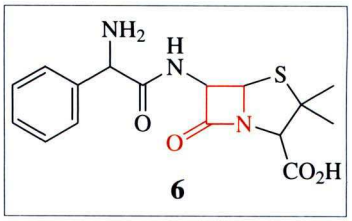
An outline of some of the features of this plasmid has already been given in Figure 5.3, and it will now become clear how these features can be exploited in cloning. Note first the restriction sites for *Eco*RI, *Bam*H1 and *Pvu*I. These are sites at which restriction fragments can be inserted in the plasmid. Suppose, for example, that we have carried out a restriction digest of DNA with *Bam*H1, and have separated a fragment by gel electrophoresis. If the plasmid is cut with the same enzyme, the break in the plasmid will have the same sticky ends as the restriction fragment, and so can be ligated into the gap to yield a **recombinant plasmid**. The recombinant plasmid can then be introduced into *E. coli* cells. The trick to do this is to mix the cells and the plasmid in cold $CaCl_2$ solution, and then to raise the temperature briefly to 42 °C. Why this works is not clear, but the result is uptake of the plasmid into the cells, which are then said to be **transformed**.

Transformation is not a very efficient process, and the next thing to do is to select those cells that have taken up plasmid from those that have not. This is where the genes for antibiotic resistance come in. pBR322 has a gene for ampicillin resistance, so if the bacteria are grown on a medium containing this antibiotic, the organisms that have taken up plasmid will survive but non-transformed bacteria will not. The ampicillin gene is referred to as a **selectable marker** because it allows selection of transformed cells.

There is another problem. Some of the transformed cells will contain plasmid without an insert; that is, molecules that have simply re-closed during the ligation reaction. The reason for this is that an excess of plasmid over restriction fragment will have been used, and plasmid molecules that do not contain an insert will simply re-close. However, notice where the *Bam*HI site is. It is inside the gene for tetracycline resistance. Inserting a restriction fragment in this site will disrupt the resistance gene and prevent it from exerting its effect. So transformants containing plasmid with an insert will be sensitive to tetracycline, but those containing non-recombinant plasmid will be resistant. How can we make use of this fact? What is done is firstly to grow the bacterial cells on agar plates containing ampicillin. Only those cells which are transformants will grow and produce colonies. The colonies are now **replica plated** onto agar containing tetracycline. A filter paper is touched lightly onto the first plate so that a few cells from each colony are picked up, and is then touched onto the second plate so that the cells are deposited in the same positions as they occupied on the first. Now only the non-recombinant cells will grow. Colonies from cells with recombinant plasmids are those represented on the first plate but not on the second. This is summarized for clarity in Figure 5. 5.

Once the recombinant colonies have been identified, it is relatively straightforward to identify those that contain the DNA of interest using hybridization methods (see Section 5.4). A membrane is overlaid onto the agar plate and some of the bacteria from each colony become attached. The bacterial cells are disrupted by treating the membrane with an alkaline solution followed by heating. This renders the DNA single stranded and fixes it to the membrane. The membrane can then be probed with a synthetic oligonucleotide as described for Southern blots.

Ampicillin (**6**) is a derivative of penicillin. The resistance gene codes for an enzyme called **β-lactamase** which hydrolyses the lactam ring (shown in red).



6

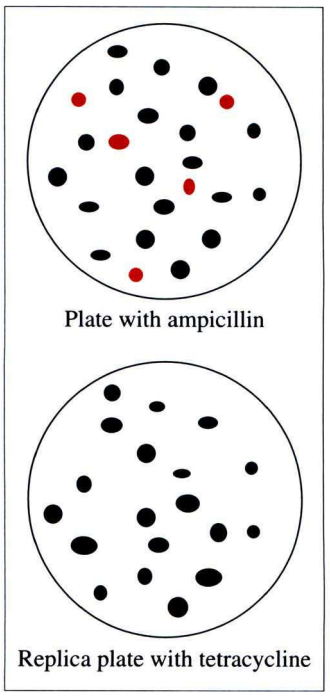


Figure 5.5 Screening for colonies containing recombinant plasmids. The colonies on the ampicillin plate are all transformants. Those in *red* do not appear on the replica plate containing tetracycline and so must be recombinants

Box 5.3 Other Prokaryotic Cloning Vehicles and Genetic Libraries

There are many other cloning vehicles in use and we cannot deal with them all here. Interested readers are referred to the books under

Further Reading. That being said, there are a few aspects of the subject that are worth a brief discussion.

One of the problems with plasmid vectors such as pBR322 is that, for reasons we will not go in to, they allow for cloning of pieces of DNA only up to about 6 kbp long. It is often necessary to clone much larger molecules than this, in which case an entirely different approach is taken. This involves the use of vectors based on a phage called λ. The phage consists of a DNA molecule surrounded by a protein coat. When it infects an *E. coli* cell, the phage attaches to the cell and injects its DNA into the bacterium. The viral DNA is incorporated into the host's DNA and is replicated (this is called the **lysogenic phase**). At some point, the cell enters the **lytic phase** in which many new phage particles are produced and the bacterial cell ruptures (or **lyses**).

The way in which λ DNA is used in cloning is as follows. The DNA is cleaved with a restriction enzyme and the foreign DNA inserted. This can be up to 23 kbp long in the case of a vector called **λGEM**. The recombinant DNA is then packaged into coat protein and the virus particles are used to infect *E. coli* cells growing as a continuous layer (known as a **lawn**) on a nutrient medium. When the lytic phase is reached, the bacteria lyse and appear as clear areas (called **plaques**) on the lawn. The plaques can be probed for the presence of the desired DNA in the usual way.

One very important use of λ vectors is the production of **genomic libraries**. A genomic library is a set of clones containing DNA fragments covering the whole of the genome of a particular organism. The intention is that the library will contain at least one copy of every gene in the genome, so that an individual gene can be "withdrawn" from the library for further study. The approach to constructing the library is to fragment the genome into large pieces (say 20 kbp long), package the pieces into a λ vector, and then produce a set of clones, each of which contains one of the fragments. These libraries can be quite large. For example, if a library is made from *E. coli* with fragments 20 kb long, then nearly 700 clones are required to be 95% certain that all the genome is represented. With yeast, the corresponding number is 2700, and for the human genome it is 428,000, which is a very big library indeed!

One of the reasons why the libraries from higher eukaryotes are so large is that their genomes contain a vast amount of "junk" DNA which seems not to have a function (see Box 5.10). The other reason is that, most genes in higher eukaryotes have introns, and so individual genes may be very large.

The existence of introns causes another sort of problem which puts some limits on the usefulness of a genomic library. If we take a gene from such a library and clone it into a bacterium, the gene will not be expressed as a protein product. The reason for this is that bacteria do not carry out splicing. For this reason in particular it would be very nice to have a library constructed from the mRNA molecules expressed by a particular cell type. There would be no "junk", and the genes in the library would be capable of being expressed in bacteria. Unfortunately, RNA cannot be cloned directly, but there is a very elegant way of overcoming this problem, namely to make a library from what is called **complementary DNA** or **cDNA**. cDNA is the molecule formed when mRNA is reverse-transcribed using the enzyme reverse transcriptase (see Section 1.8).

How is it done? The first problem is to isolate the total mRNA from the cell type of interest. There is an easy way of doing this. Recall that one of the properties of eukaryotic mRNA is that it contains a poly-A tail added as a post-transcriptional modification (see Section 4.2). This fact is made use of in a technique called **oligo(dT)-cellulose chromatography**. Chains of about 20 dT residues are covalently attached to cellulose and the material packed into a column equilibrated with buffer. If total cellular RNA is passed through the column, the mRNA molecules will bind to the oligo(dT)-cellulose by the normal base-pairing interactions, whereas all the rest of the components will pass through unretarded. The bound material can then be eluted from the column.

The next step is to reverse-transcribe the mRNA preparation using a retroviral reverse transcriptase. The enzyme needs a primer, and poly-dT is used for this purpose. Figure 5.6 shows the beginning of the process with poly-T hybridized onto the poly-A tail of the mRNA (shown as only six residues long for simplicity). The reverse transcriptase, in the presence of the four deoxyribonucleoside triphosphates, is then used to synthesize single-stranded DNA (ssDNA) using the mRNA as the template. It is now necessary to remove the mRNA strand. This was originally done by mild treatment with base, under which conditions DNA is stable but RNA is hydrolysed. More recently, an alternative method has been introduced where the RNA is fragmented using a nuclease called RNase H. This breaks the mRNA into small pieces, some of which will remain hybridized to the ssDNA. The advantage of this is that if a DNA polymerase such as Pol I is added, the polymerase uses the 3′-OH groups of the RNA fragments to prime second-strand DNA synthesis and the complementary strand of the DNA is produced.

All that is required now is to ligate the mixture of DNA molecules into a vector such as λ and clone to produce the library.

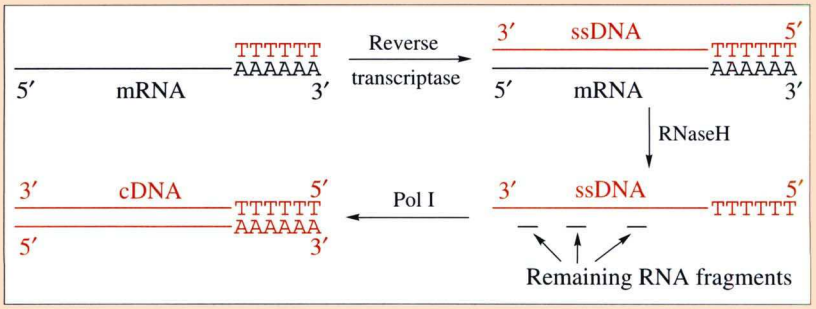


Figure 5.6 Procedure from the production of cDNA from mRNA

Screening a cDNA library can be done using radiolabelled oligonucleotide probes as usual. There is, however, another possibility. If the library is constructed in a particular vector called λgtII, the protein products in the clones are expressed, and screening can be done by searching for the clone that produces the protein of interest. The way this is done is by using an antibody that is specific for the protein.

Box 5.4 Vectors for Eukaryotes

This is a large topic, which we will only touch on briefly. More information is available in the texts under Further Reading.

Much attention has been focused on the use of baker's yeast (*Saccharomyces cerevisiae*) as a host organism for cloned genes. One reason for this is that the native forms of many eukaryotic proteins are glycosylated (that is, they have covalently attached polysaccharide chains). Prokaryotes generally cannot synthesize these polysaccharides but yeast can, and so expression of the proteins in yeast is necessary to produce the native forms. Yeast is unusual among the eukaryotes in having a plasmid, called the **2μm plasmid** (or, sometimes, the **2μ circle**), which behaves like a prokaryotic plasmid and has been extensively used for cloning. An exciting recent development in this field has been the construction of **yeast artificial chromosomes (YACs)**, molecules that are stably replicated during division of the yeast cells. The advantage of these is that they can be used to clone very large pieces of DNA up to 1 Mbp in length. This means that they can be used to clone even the largest genes, or to

make libraries of eukaryotic genomes that require far fewer clones than do prokaryotic libraries. For example, a library of the human genome can be constructed with only about 10,000 clones.

Genetic modification of plants is also a very active area of research. The most common approach to this is by using a plasmid called the **Ti plasmid**, which occurs in the bacterium *Agrobacterium tumefaciens*. This is a bacterium that infects plants and induces a disease called **crown gall** (a type of tumour). The name Ti plasmid is derived from this activity (*t*umour *i*nducing). The genes that cause the tumour are in a section of the plasmid which is stably inserted into the genome of the host plant. What is done is to remove some of those genes, so that insertion does not result in disease, and replace them with the gene or genes that it is intended to introduce into the plant. Once plant cells have been infected with *A. tumefaciens* that contains the modified Ti plasmid, mature plants can be generated from these cells, and the inserted genes will be inherited by every cell in the plant.

One problem with this is that *A. tumefaciens* infects only dicotyledonous plants such as tomato, potato and pulses. Many interesting crop plants such as wheat and rice are monocots. This difficulty has been overcome by the unlikely sounding technique of shooting the plasmid DNA directly into the cells of plant embryos! The DNA is deposited on gold microprojectiles and then fired directly into the plant; the technique is called **biolistics**.

The use of retroviruses for introducing DNA into humans has already been mentioned in Box 5.1. This approach is also used for genetically modifying other animals. The other technique which has met with success is direct injection of DNA into nuclei, which results in integration of the foreign gene into the chromosome. Attention is being focused on the use of **embryonic stem cells** for this purpose. The advantage of this is that these cells can be used to produce a complete animal. There are, however, ethical problems surrounding this sort of research which have yet to be fully explored.

5.7 The Polymerase Chain Reaction

Another method for the amplification of DNA was developed in 1985 by Kary Mullis and his co-workers.[8] This is the **polymerase chain reaction**, or **PCR** as it is widely known. It is an amazingly powerful yet relatively simple method that has become one of the most widely used tools in molecular biology. It is essentially a chain copying method, the principles of which are outlined in Figure 5.7.

Mullis was awarded the Nobel Prize in Chemistry in 1993 for "his invention of the polymerase chain reaction (PCR) method". His Nobel Lecture can be found at http://www.nobel.se/chemistry/laureates/1993/mullis-lecture.html.

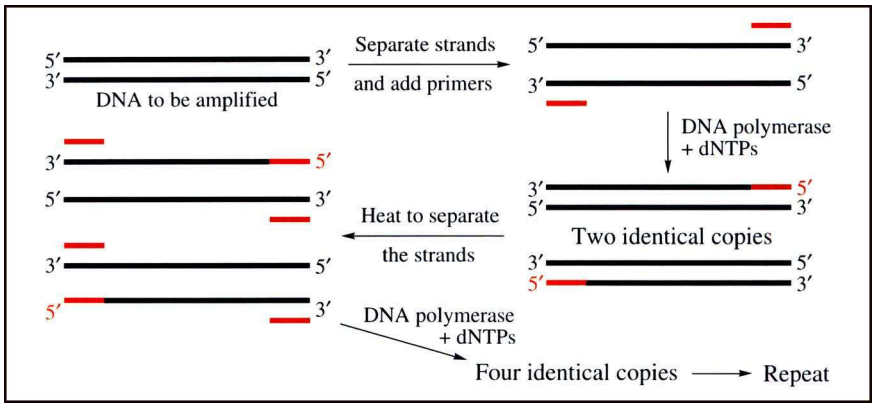


Figure 5.7 The polymerase chain reaction

Essential to the method are a pair of primers complementary to the two ends of the piece of DNA to be amplified. One of them is complementary to the 3′-end of one strand of the target DNA, and the other is complementary to the 3′-end of the other strand. Generally the primers used are about 20 nucleotides long. At the beginning of the experiment, the target DNA is heated at about 95 °C to separate the strands. The primers, DNA polymerase, and a mixture of the four deoxynucleoside triphosphates are added. The temperature is decreased to a value suitable for hybridization of the primers to the separated chains, as shown in Figure 5.7. An annealing temperature just below the melting temperature for the primer/DNA template is used (see equation 5.1). Chain copying will then occur in the 5′→3′ direction to yield two identical copies of the original molecule. The whole process of chain separation, annealing and chain synthesis is then repeated 30 or 40 times, depending on the degree of amplification required. The process is very rapid and is amenable to automation. All that is required is to cycle the reaction mixture between the temperatures for chain separation, annealing and chain elongation. The cycle time is 4–5 minutes.

Thermus aquaticus is a member of a group of micro-organisms called **thermophiles**. These bacteria are found in places such as hot springs, and are adapted to life at high temperatures. In consequence, their proteins tend to be much more thermostable than those from organisms living in more conventional habitats.

A little thought will show that the polymerase used must have rather unusual properties in that it must be able to withstand heating to 95 °C without losing its activity. The enzyme used is called ***Taq polymerase*** because it is obtained from the organism *Thermus aquaticus*. The enzyme is very thermostable and also works best at a relatively high temperature of about 75 °C; this is, therefore, the temperature used for the chain elongation step.

A simple calculation shows that the degree of amplification achievable by PCR is very large indeed. At each step the amount of DNA present is doubled. So after 30 steps, an amplification of about 10^9-fold is obtained.

Worked Problem 5.3

Q Suppose that 40 cycles of PCR are used to amplify 1 molecule of DNA 1000 bp long. What weight of product will be obtained? Take the Avogadro constant as 6×10^{23} mol^{-1}.

A 40 cycles will give an amplification of 2^{40}-fold, which is about 10^{12}-fold. 10^{12} molecules represent $10^{12}/(6 \times 10^{23}) = 1.8 \times 10^{-12}$ mol of DNA. The M_r of a piece of DNA 1000 bp long is about 600,000, so the weight of DNA produced would be about 1 μg. By DNA standards this is a large amount of material (but note that 40 cycles is at the limit of what is practical).

As well as leading to a very high degree of amplification, PCR is also a very robust technique and can be used with partially degraded samples. The result of this is that it has proved possible to obtain sequence information from the tiny amounts of DNA that can be extracted from sources such as the fossilized bones, or preserved tissues, of extinct organisms (to this extent at least, the story line of Jurassic Park is not entirely science fiction!).

Box 5.5 An Example of the Use of PCR

The organism *Helicobacter pylori* is one of the causative agents of duodenal ulcers. Some strains of the bacterium are sensitive to the therapeutic agent metronidazole, but others are resistant to the drug. Figure 5.8 shows the results of PCR analysis of the catalase gene from three different strains. The gene, which is about 1500 bp long, was amplified using the primers 5′-GTGA-AACAAACCACTGCTTTTGG-3′ and 5′- GTGGTGCATGTC-TTTTCCACG-3′. Electrophoresis of the PCR products showed that the gene from strains a and c, which are metronidazole-sensitive, was amplified by these primers (shown by the heavy band in the duplicate tracks labelled a and c in Figure 5.8). The catalase gene from the resistant strain (c) could not be amplified with these primers, showing that it has significant differences from the gene in sensitive strains. These results provide a line of investigation for the mechanism of metronidazole resistance in *H. pylori*.

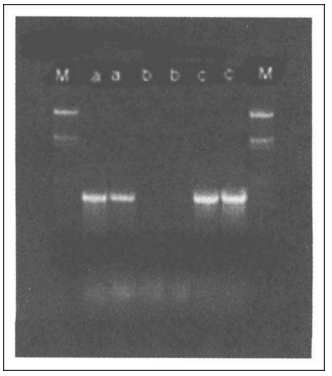


Figure 5.8 PCR analysis of the catalase gene from three strains of the bacterium *Helicobacter pylori*. The tracks labelled M contain markers obtained by digesting λ DNA with *Hin*dIII and *Eco*RI (courtesy of Dr Ravi Nookala)

5.8 DNA Sequencing

Probably the single most important development in molecular biology in the second half of the 20th century was the invention of methods for the determination of base sequences of DNA molecules. In fact, the first nucleic acids to be sequenced were small RNA molecules (see Section 4.4). The reasons for this were two-fold. Firstly, they were relatively easy to purify. Secondly, and more important, RNases were available that allowed cleavage of the polynucleotide chains at specific bases. Nevertheless, RNA sequence analysis was a slow and laborious process and is now very rarely undertaken. Sequence analysis of an RNA molecule is now done by sequencing the DNA that codes for it, and we will not deal with the classical methods here.

Sequencing of DNA had to wait for the discovery of methods for specific fragmentation of these much larger molecules into pieces of manageable size, and for methods for increasing the quantities of the fragments available. That is, the discovery of restriction enzymes and the development of cloning were essential prerequisites for DNA sequencing. Equally important, someone had to have the inspiration to come up with a method for doing it. In fact, two methods were invented at more or less the same time.

These methods have one thing in common: they depend on being able to measure the lengths of DNA fragments using polyacrylamide gel electrophoresis. They differ, however, in how those fragments are derived.

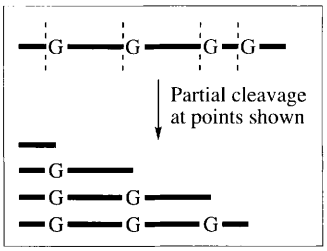


Figure 5.9 Fragments produced by partial cleavage of a DNA molecule at G residues

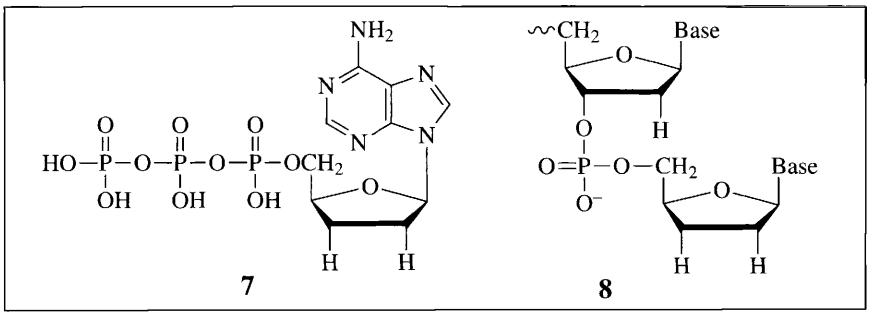
Allan Maxam and Walter Gilbert[9] developed a method which depends on chain cleavage. In particular it involves the use of chemical procedures to break the DNA chain at a particular base. Consider for example, a reaction that cleaves DNA at G residues, and suppose that it is applied to a piece of DNA in such a way that only partial cleavage occurs at each G in the chain. The result would be as shown in Figure 5.9. That is, a set of fragments is generated, each of which terminates where a G was present in the original chain. By measuring the lengths of the fragments produced, the positions of the G residues are established. If four such cleavage methods could be developed, then application of each of them in turn to the DNA molecule would produce the sequence. It turns out that it is not quite as easy as that, because it is not possible to devise procedures that cleave the chain completely specifically at each of the four bases. Nevertheless, a viable method was developed and led to successful sequencing of DNA molecules. We will not, however, look any further at the details because the method is now used infrequently. This is because it is more difficult to use than the other available approach, and is not so amenable to automation.

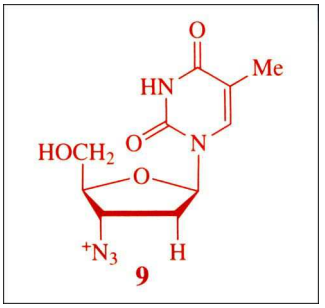
The method invented by Frederick Sanger and Alan Coulson involves chain copying. The initial approach was to carry out copying in reaction mixtures that contained a limited amount of one of the four deoxynucleoside triphosphates. The result of this is to generate a set of partial copies that terminate at the position where the residue in limited supply was to be inserted in the chain. Even though this method worked well, and was successfully used to determine the sequence of the genome of a virus, it was soon superseded by an even better one.[10] This is known as the **dideoxy chain termination method**, and is the basis of nearly all modern DNA sequencing. We will look at it in detail.

Gilbert and Sanger shared the second half of the Nobel Prize in Chemistry in 1980 for "their contributions concerning the determination of base sequences in nucleic acids". It is worth noting that this was the second Nobel Prize awarded to Frederick Sanger. He won the first one in 1958 for his work on the amino acid sequence of the protein insulin. It is remarkable indeed that he was centrally involved in working out methods for the determination of the structures of both major classes of biological macromolecules.

First it is necessary to look at what a **dideoxynucleoside triphosphate** A typical example (dideoxyadenosine triphosphate, ddATP) is shown in **7**. The particular feature to note is that the ribose ring lacks hydroxyl groups at both the 2′- and the 3′-positions. These molecules are recognized by DNA polymerase and are incorporated into polynucleotide chains in the normal way, but it should be obvious that once a dideoxy residue has been incorporated, chain growth stops. There is no 3′-OH to which to add the next residue. The 3′-end of the chain would now have the structure shown in **8**.

Dideoxynucleosides were originally developed as anti-viral agents, the hope being that they could be used to specifically prevent viral DNA synthesis. This approach to chemotherapy has not been successful. On the other hand, **azidothymidine (AZT, 9)** is an effective therapeutic agent against AIDS. It works by inhibiting the enzyme reverse transcriptase.





Suppose that we have a ssDNA molecule for which we know a short stretch of sequence at the 3′-end. We can then synthesize an oligonucleotide that is complementary to this know piece of sequence, and use it as a primer for chain copying by DNA polymerase (a Klenow fragment must be used to avoid exonuclease activity). Further suppose that, as well as the usual four deoxynucleoside triphosphates, we include in the reaction mixture a small amount (say 1%) of a dideoxynucleoside triphosphate. If the added dideoxynucleoside triphosphate is ddTTP, the outcome will be as shown in Figure 5.10.

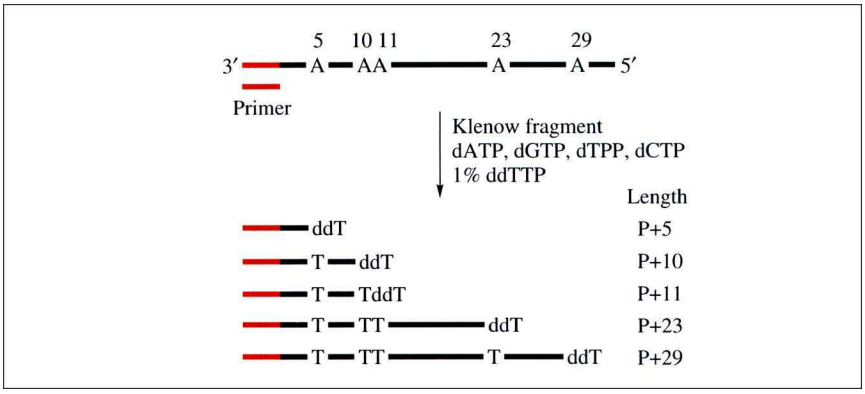


Figure 5.10 Sequence analysis by the dideoxy method

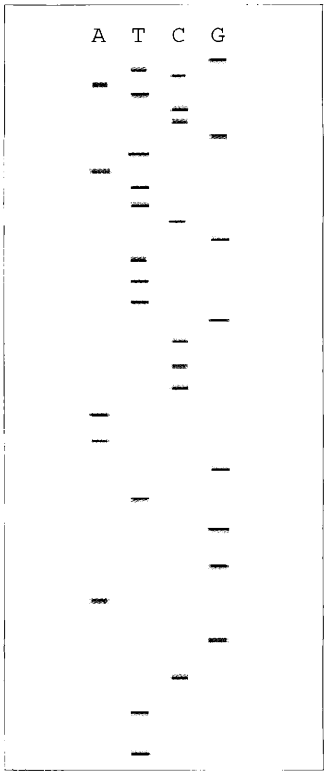


Figure 5.11 Gel electrophoresis pattern expected from sequence analysis of the ssDNA molecule with the sequence shown below. The letters above the tracks are the bases at which copying terminates. For example, track A contains fragments obtained by copying in the presence of ddTTP.
3′-TTCGAGGTGAACCCGTTTG-CTTATGCCTACTG-5′

The 3′-end of the ssDNA for which we know the sequence and for which a complementary primer has been synthesized is shown in red. The unknown part of the molecule is represented simply as a black line with the positions of the A residues specified. When chain copying is started and the first A at position 5 is reached, a dT will be inserted for most of the molecules being copied, but for a small fraction (about 1%) ddT will be inserted and chain elongation will stop. This produces a fragment of length P + 5, where P is the length of the primer. For the rest of the molecules, chain elongation proceeds to the next A at position 10, where again a small fraction of the new chains will have ddT inserted and will terminate giving a fragment of length P + 10. Clearly, what we will have at the end is a mixture of new chains, all of which terminate with a ddT residue (except the longest one, unless the last residue in the chain happens to be an A). Measurement of the lengths of those fragments will yield the positions of the A residues in the original ssDNA molecule.

The experiment is now repeated three more times, in each case with a different ddNTP in the reaction mixture. At the end, the four reaction mixtures are run side by side on a polyacrylamide gel under conditions where the fragments are dissociated from the parent ssDNA molecule, and the positions of the bands observed. This is usually done by using radiolabelled ddNTP molecules in the reactions so that the gel can be viewed after autoradiography.

Suppose that the actual sequence of the ssDNA molecule in Figure 5.10, excluding the part used for priming, was as in the legend of Figure 5.11. The band pattern expected after electrophoresis would be as shown in the figure.

It is unnecessary to estimate the lengths of the fragments. All that is necessary is to read upwards from the bottom of the gel and see in which lane successive bands fall. The smallest fragment runs the furthest

through the gel, and so the first residue in the sequence is a T followed by another T. Next comes a C then G, and so on. It is a simple matter to read the complete sequence from the gel. Remember, however, that the sequence as read is 3′→5′, which is the opposite way around from the way that sequences are usually written.

For simplicity, the sequence in Figure 5.11 is quite short. In practice, sequences up to 400–500 bases can be read from a single gel (longer sequences can be obtained using automated methods as described in Box 5.8). Note that as the pieces become longer, the bands move closer together. With very long sequences, the separation of the longer bands becomes less and less until a point is reached where it is no longer possible to distinguish them. This effect is apparent in Figure 5.12, which shows part of the autoradiograph obtained in an actual DNA sequencing experiment.

One problem should immediately come to mind concerning this technique. It requires single-stranded DNA, whereas DNA is usually double stranded. The solution to this problem is provided by a phage called **M13**. This phage has a genome consisting of a circle of ssDNA. After infection of the *E. coli* host, the ssDNA acts as a template for synthesis of its complementary strand. The resulting double-stranded DNA (the **replicative form** replicates in the usual way. When new phage particles are formed, the dsDNA is converted back into the single-stranded form and packaged into the viral proteins. These properties have been used to advantage in DNA sequencing. The double-stranded replicative form of the viral genome can be used just like a plasmid, and can be used for cloning dsDNA. Many different constructs have been produced, but a particularly useful one is called M13mp8. This variant of the replicative form DNA contains a **polylinker**, that is, a stretch of DNA containing recognition sites for a variety of restriction enzymes. The sequence of this part of the molecule is shown in Figure 5.13. The polylinker is in red, and some of the restriction sites that it contains are indicated. Pieces of dsDNA that have been produced using any of these restriction enzymes can be inserted into the polylinker. If *E. coli* cells are then transformed with the recombinant dsDNA, viral particles will be produced that contain only one of the two DNA strands (the upper one in Figure 5.13) and the inserted DNA can be sequenced directly. There is another advantage to this method. Since the inserted DNA is sequenced *in situ*, the ssDNA on the 3′-side of the polylinker can be used as the primer binding site, and the same primer can be used irrespective of the sequence of the inserted DNA. That is, a **universal primer** can be used. On possible universal primer is shown in Figure 5.13 above the stretch of DNA to which it is complementary (in pink).

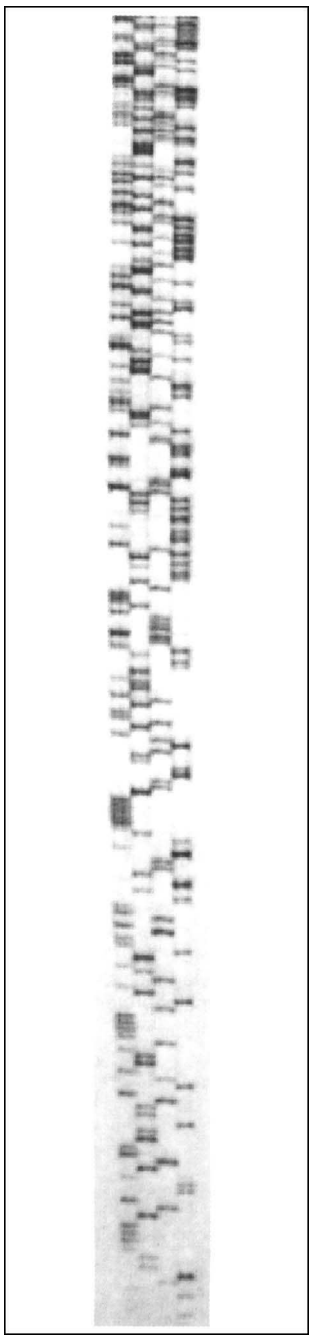


Figure 5.12 The autoradiograph obtained in an actual DNA sequencing experiment (courtesy of Dr Dunstan Rajendram). The top part of the gel has been cut off to fit it on the page. The tracks are G, A, T and C from left to right

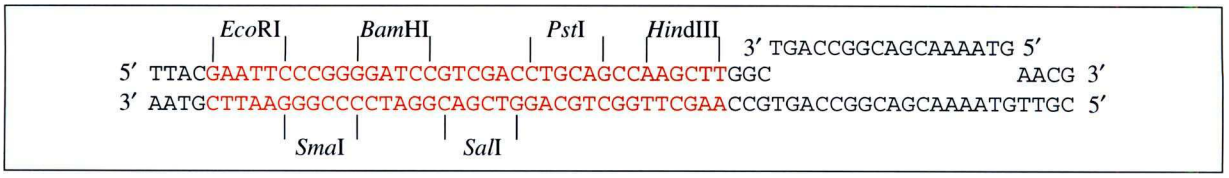


Figure 5.13 The polylinker of the vector M13mp8 (in *red*) and a region that can be used for binding a universal primer for DNA sequencing (in *pink*). The primer sequence (17mer) would be as shown

Box 5.6 Site-directed Mutagenesis

Phage M13 is also a very valuable tool for a technique called **site-directed mutagenesis**. This is a method which allows production of a protein with an amino acid changed from that which normally occurs at a specific position in the polypeptide chain. There are several reasons why this is a useful thing to be able to do. For example, if a particular amino acid is suspected of being involved in the biological activity of a protein, changing it and observing the effects on activity can provide valuable confirmatory evidence. Sometimes the therapeutic properties of a protein can be improved by changing one of the amino acids. There are many other possibilities.

Site-directed mutagenesis was originally developed by Michael Smith. He was awarded half of the Nobel Prize in Chemistry in 1993 "for his fundamental contributions to the establishment of oligonucleotide-based, site-directed mutagenesis and its development for protein studies".

How is it done? The first requirement is that the DNA coding for the protein is available and that its sequence is known. The DNA can be inserted into a double-stranded M13 vector in the usual way, and the corresponding single-stranded molecule isolated from recombinant phage. Suppose that we now synthesize an oligonucleotide which is complementary to the region of the protein that we wish to modify, but in which the codon for one of the amino acids has been changed. For example, if we wished to replace a cysteine specified by the codon 5′-TGC-3′ with a serine, specified by the codon 5′-TCC-3′, then the complementary oligo would contain the sequence 3′-AGG-5′. If the oligo is sufficiently long (say 20 residues) it will still hybridize, even though there is a single base mismatch. This is shown diagrammatically in the left-hand part of Figure 5.14. The second strand of the replicative form can now be completed by using DNA polymerase, and the gaps sealed with DNA ligase. If the replicative form of the molecule is now used to transform *E. coli* cells, two types of colony will be produced, one containing replicative form molecules with the original gene and the other containing the mutated gene. The latter can be cut out, inserted into an expression vector, and used to produce mutant protein.

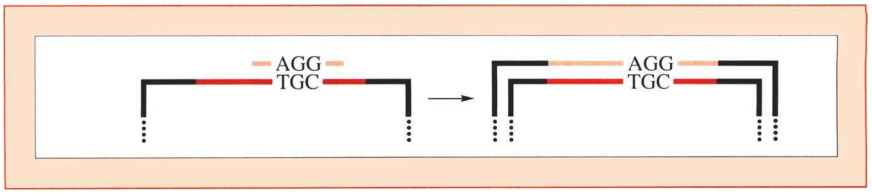


Figure 5.14 Site-directed mutagenesis. The left-hand diagram shows a part of the M13 DNA genome in *black*, with an insert containing the coding sequence for a protein in *red*. A codon for cysteine is specified. Hybridized to it is a synthetic oligonucleotide (*pink*) containing a single base mismatch (note that the oligo is not drawn to scale). Addition of DNA polymerase and DNA ligase results in synthesis of the remainder of the second strand of the replicative form of the M13 DNA

The other problem that might come to mind is what happens if we wish to sequence a piece of DNA that is longer than 400–500 bases? The answer depends on how long the DNA is. If it is, say, 800–1000 bases long, one approach would be to sequence it from both ends. The object of this is to find an overlapping complementary region at the ends of the two sequences that have been determined, as illustrated in Figure 5.15. The final bases assigned in the two sequences are shown explicitly, whereas the earlier ones are represented simply as lines. The ends of the two sequences are complementary, and hence the sequence of the entire molecule has been determined. Note that in practice a longer overlap than that shown would be required. The reason is that there are likely to be errors in the bases assigned at the very ends of the sequences, and so a long overlap is required to be certain that the two sequences are genuinely complementary. Note also that, even with shorter molecules, it is good practise to sequence them from both ends to provide confirmatory evidence for the sequence assigned.

There is a very cunning way of sequencing from both ends. A second vector called M13mp9 is available in which the polylinker has been put in the opposite way around. So if a sample of the DNA is inserted into both vectors, the ssDNA molecules obtained will be in opposite orientations as well.

Figure 5.15 Sequencing a long DNA molecule from both ends to provide a complete sequence. The *black* and *red lines* may represent up to 400–500 sequenced residues

For somewhat longer molecules, a technique involving "walking" along the sequence can be used. The sequence of the first few hundred residues is determined in the usual way, and an oligonucleotide is synthesized which is complementary to the 5′-end of the sequence obtained. This oligonucleotide can then be used as a primer to extend the sequence further. The technique is illustrated in Figure 5.16. Several steps can be carried out, thus allowing the sequence analysis of quite large DNA molecules. The disadvantage of the method is, of course, that it requires a new oligo to be synthesized at each step in the walk along the molecule.

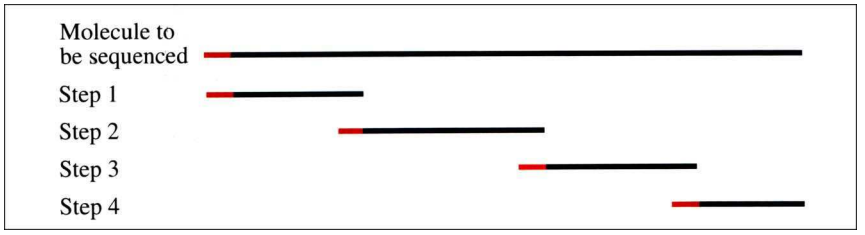


Figure 5.16 Sequencing a long molecule by "walking" along it. As much sequence as possible is determined in step 1. An oligonucleotide is then made with the sequence determined at the end of step 1. This is used to extend the sequence in step 2. The process is repeated until the end of the original molecule is reached. Primers are shown in *red*

With DNA molecules thousands of nucleotides long, a more complicated technique is used. The approach is to do what is known as **shotgun sequencing**. The molecule is fragmented using a restriction enzyme and the fragments cloned into an M13 vector. Clones are randomly selected and sequenced. This will result in a set of partial sequences, but their order in the original molecule will be unknown. Fragmentation is then repeated with a second restriction enzyme to produce a second set of fragments, and these again are cloned and sequenced. This second set is then searched to find fragments which **overlap** ones from the first digest. The sorts of results obtained are illustrated in Figure 5.17.

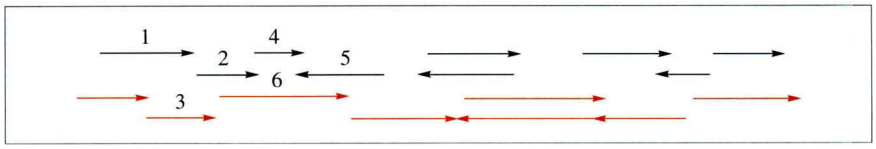


Figure 5.17 Shotgun sequencing of a large DNA molecule. The *black arrows* represent sequences of fragments produced by a restriction enzyme, with the arrow head showing the direction of sequencing. The *red arrows* show sequences of fragments produced by a second restriction enzyme

For example, fragment 3 from the second set contains the 5′-end of fragment 1 and the 3′-end of fragment 2 from the first set, thus establishing the order of fragments 1 and 2 in the original molecule. Fragment 6 contains the 5′-end of fragment 2, all of fragment 4, and the complement of the 5′-end of fragment 5 (which has been sequenced in the reverse direction). Working through the two sets of fragments allows the original sequence to be reconstructed. The method is called a shotgun approach because it is unnecessary to obtain and sequence all of the fragments from both digests. What is required is a randomly selected set of fragments that allows the original sequence to be reconstructed (see Problem 5.6).

Box 5.7 A Few Early Results from DNA Sequencing

It is impossible to overestimate the contribution that DNA sequence analysis has made to science. The following are just a small number of illustrative examples.

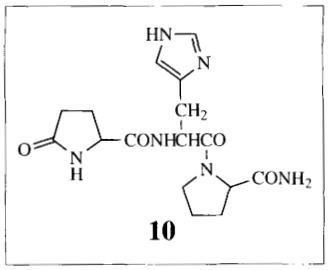
The first genome to be sequenced was that of the virus ΦX174.[11] Because it was the first, this genome provided a wealth of information about the processes of gene expression, but perhaps one of the most interesting features was how this small genome (5375 nucleotides) could pack in so much information. Apart from the fact that there is no wasted space, it turns out that two stretches of the DNA both code for two different proteins but in different reading frames. This has now been found with other viruses as well

and presents an interesting solution to the problem of how to keep a genome as small as possible.

The next major achievement of DNA sequencing, also by Sanger's group, was the determination of the structure of the human mitochondrial genome.[12] The genome is 16,569 residues long and codes for several proteins. When the genome was first sequenced, only five of those proteins were known. Subsequently, the identities of the other coded proteins were discovered and all found to be integral proteins in the mitochondrial inner membrane. The genome also codes for a 12S and a 16S rRNA, but for only 22 tRNA molecules. The interesting thing about that is that the standard mechanism of base wobbling requires a total of 31 tRNA molecules to read all of the coding triplets. One possibility was that some other tRNA molecules are imported into the mitochondria to make up a complete set. Alternatively, it was possible that some of the codons are simply not used in mitochondrial DNA. It turns out that neither is the case. Rather, in those cases where all four codons in a family code for the same amino acid, a single tRNA recognizes them all. So, for example, the codons for threonine (CUU, CUC, CUG, CUA) are all recognized by a tRNA with the anticodon GAU; that is, the tRNA effectively ignores the third base. Most remarkable of all, it turns out that the genetic code is not universal. Some of the codons have a different meaning in mitochondria, and there are even differences between mitochondria from different organisms. In mammalian mitochondria, the codon UGA means Trp instead of STOP, AGA and AGG mean STOP instead of Arg, and AUA and AUU mean Met instead of Ile. How and why these differences arose is a mystery.

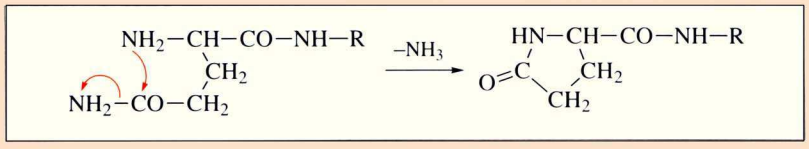
One of the fascinating things about mitochondrial DNA is that it is inherited only from one's mother. Variations in the DNA sequence have, therefore, been used to study maternal lineages in populations. The conclusion that has been arrived at is that all women in the world today are descendants of a single woman who lived in East Africa about 150,000 years ago. This person is often referred to as the Mitochondrial Eve. You can learn more about this from Olson's book.[1]

As a final example, consider the rather strange looking tripeptide shown in **10**. This is called **thyrotropin releasing factor** (TRF). It is one of a group of peptide hormones that are produced in a part of the brain called the hypothalamus, and that provide a link between the activity of the brain and the chemical processes going on in the body. The peptide has the structure pyroglutamylhistidinylproline amide. That is, the N-terminal glutamine is cyclized to pyroglutamate (see Scheme 5.7), and the C-terminal proline is amidated. The question is, how is it made?



10

The pioneers in the study of neuropeptide hormones were Roger Guillemin and Andrew Schally. They were warded the Nobel Prize in Physiology or Medicine in 1977 "for their discoveries concerning the peptide hormone production of the brain". The story of the isolation and structure analysis of TRF is one of the epics in this field. Hypothalamus fragments were obtained from 300,000 sheep. Only 1 mg of the peptide was isolated from this material and used for structure determination. A very readable account is given in Guillemin's Nobel lecture at http://www.nobel.se/medicine/laureates/1977/guillemin-lecture.pdf.



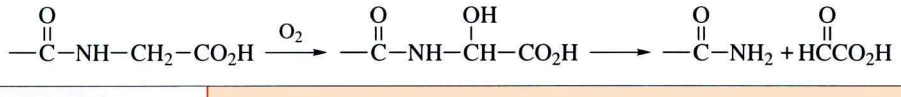
Scheme 5.7

The answer is provided by the amino acid sequence of the **prohormone** shown in Figure 5.18.[13] This sequence was obtained by translation of the DNA sequence of the cloned gene.

```
  1  mpgpwlllal  aliftltgip  escalpeaaq  eegavtpdlp  glenvqvrpe  rrflwkdlqr
 61  vrgdlgaald  switkrqhpg  kreeeekdie  aeergdlgeg  gawrlhkrqh  pgrranqdky
121  swadeedsdw  mprswlpdff  ldswfsdvpq  vkrqhpgrrs  fpwmesdvtk  rqhpgrrfid
181  pelqrsweek  egegvlmpek  rqhpgkralg  hpcgpqgtcg  qtgllqllgd  lsrgqetlvk
241  qspqvepwdk  eplee
```

Figure 5.18 Amino acid sequence of the precursor for thyrotropin releasing factor

Within the prohormone there are five repeating units, shown in red, with the sequence KR**QHP**G(K,R)R. What happen first is that a specialized enzyme with trypsin-like activity recognizes the pair of basic amino acids at the beginning and end of each of these sequences, and liberates the tetrapetide QHPG. Next, the C-terminal glycine is cleaved oxidatively to leave proline amide at the end of the tripeptide (Scheme 5.8). Finally, the N-terminal glutamine is cyclized to give the biologically active TRF. Many other neuropeptides are made in this way, and the story is a nice example of how DNA chemistry and protein chemistry used together can solve interesting biological problems.



Scheme 5.8

Box 5.8 Automated DNA Sequence Analysis

Manual methods of DNA sequence analysis are rapid, but by no means rapid enough for very large sequencing projects. At least five sets of samples can be run at the same time on gels such as that in Figure 5.12, allowing perhaps 2000 residues to be identified in a day's work. However, that still means that sequencing a Mb of DNA would take 500 days and would involve a vast amount of gel preparation and data analysis. This is not how it is done nowadays. Rather, the process has been automated, and many commercial machines are now available that not only speed sequencing up but take out much of the manual labour involved.

Key to automated sequencing was the development of sets of fluorescent markers with different emission wavelengths that could be use to label the chain-terminated fragments. In early applications,

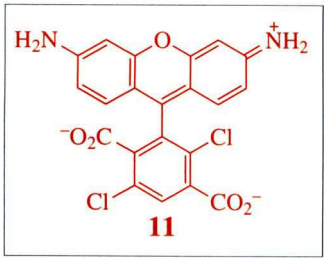
the fluorescent tags were attached to the primers, with a different one for each of the four chain-termination reactions. After the copying experiment, the four sets of reaction mixtures were run through a polyacrylamide gel, and at a point towards the bottom of the gel a detector was used to scan each of the four tracks in the gel in turn at the emission wavelengths appropriate for the reaction mixture in that track. The sequence could then be deduced from the order in which signals were detected. (Imagine the experiment in Figure 5.11, but with the bands run completely through the gel rather than being stopped and visualized by autoradiography. The first band to pass the detector would be in track T and would be detected as such by the specific fluorescent tag on the primer. The second band would also be a T, followed by a C, and so on.)

Most recently, a novel method has been developed in which the dideoxynucleoside triphosphates themselves are labelled with fluorescent tags. Remarkably, DNA polymerases have been engineered that still recognize these tagged ddNTPs and incorporate them into the correct positions in a growing DNA chain. The procedure is thereby much simplified. It is no longer necessary to carry out four separate copying reactions because the fluorescent tag is on the chain terminator, and so all four reactions can be carried out with the same sample. The ddNTP that has led to chain termination is identified by the wavelength of emission of its tag.

Similarly, the fragment mixture is separated in a single electrophoretic procedure. In many commercial machines, polyacrylamide gel electrophoresis has been replaced by **capillary electrophoresis** in free solution; components still migrate in order of decreasing size, but the reason for this is complex and we will not go into it. The benefits of capillary electrophoresis are that it removes the necessity for making gels, is very rapid, and increases the sensitivity of detection; sequence analysis can be carried out on as little as 100 ng of material. The equipment used is shown diagrammatically in Figure 5.20.

As each component of the mixture of fragments moves down the capillary past the detector, the nature of the terminating dideoxynucleoside is identified by the fluorescence detector, and the results are fed to an on-board computer. The computer stores the results, and at the end of the run produces a trace of the fluorescence signals and reads the sequence from it. A very small part of the output from the middle of such a sequencing run is shown in Figure 5.21. The deduced sequence is shown at the top of the trace.

A widely used set of labels are derivatives of dichlororhodamine (**11**). The fluorescence emission spectra are shown in Figure 5.19, with each spectrum labelled with the trade name of the derivative. The emission maxima are well separated and allow for easy discrimination between the labels.



11

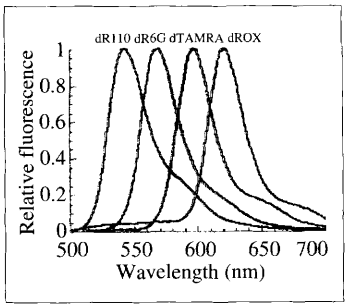


Figure 5.19 Fluorescence spectra of dichlororhodamine derivatives

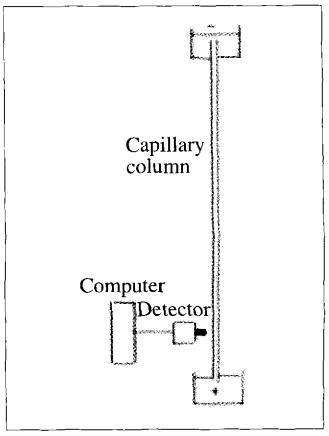


Figure 5.20 Schematic diagram of the equipment used in automated DNA sequencing

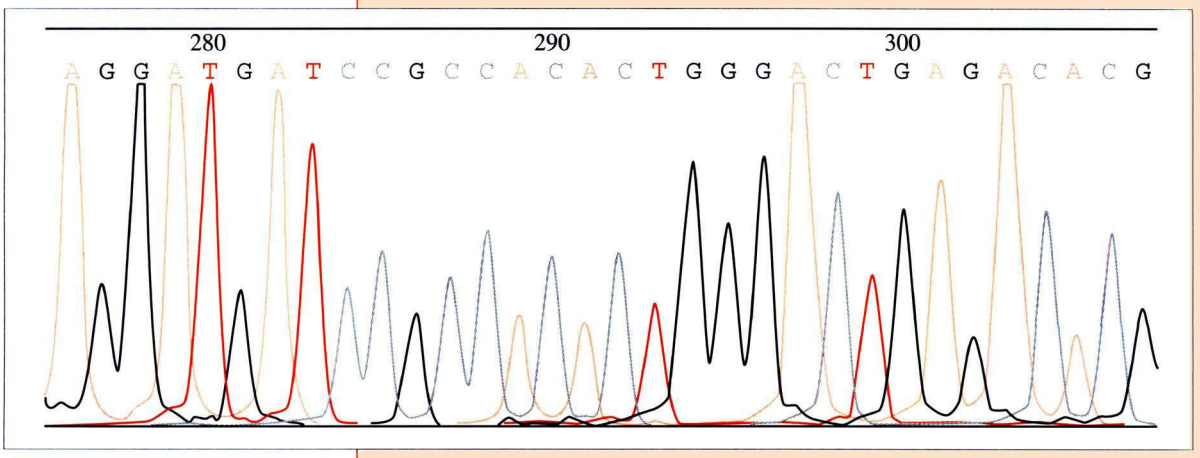


Figure 5.21 A small section of the results from automated DNA sequence analysis. The peaks are the output from the fluorescence detector at each of four wavelengths corresponding to the emission maxima of the fluorophors used. The corresponding residue is given above each peak (courtesy of Dr Renata Culak)

The most advanced machines can handle up to 384 samples in a single run, and can produce 99% reliable sequences up 1000 residues long. The reason why longer sequences can be determined compared with manual methods is that the long fragments that bunch together at the top of the gel become separated as they migrate completely through the capillary. It is perfectly feasible for a laboratory operating such a machine to sequence up to a million bases per day!

It is this sort of technology, combined with the development of robotic instruments that can also deal with sample preparation and handling, which has made genome sequencing a practical proposition. Indeed, genome sequencing has become almost a routine matter. To date, the sequences of more than 100 bacterial genomes have been determined, as have those of several eukaryotic organisms. Reference to some of the interesting results that have come from this work has been made elsewhere in this text, but space precludes any more extensive description. Interested readers can explore the matter further using the internet. For example, in the home page of the NCBI there is a link to a section on genomic biology which has a wealth of information, as does the website of **The Institute for Genomic Research** (TIGR) at http://www.tigr.org/.

5.9 Computer Applications in DNA Chemistry

Modern DNA chemistry would be impossible without the use of computational techniques. We have already seen (Box 5.8) that computers are used to collect, store and analyse the output from automated sequencing experiments. They are also indispensable for analysing the

results of shotgun sequencing and recognizing the overlaps that allow fragments to be assembled into complete structures. For example, the first bacterial genome to be sequenced was that from *Haemophilus influenzae*.[14] The genome is 1.83 Mbp long. It was fragmented by mild sonication and the sequences of about 20,000 clones were determined. These sequences were assembled using a computer into 140 larger fragments (called **contigs** from contiguous sequences). The analysis required 30 h of computer time. Finally, the positions of the contigs in the genome were determined by using a hybridization method. Large fragments of the genome were cloned in λ. Synthetic oligonucleotides from the ends of the contigs were used to screen the library. If oligonucleotides from two different contigs hybridized to the same clone, then it could be assumed that the contigs were adjacent in the genome. The point is that analysis of the experimental data would have been impossible without the aid of computers. Suites of software are available to carry out all the processes of data acquisition and analysis (for example, Lasergene from DNAStar Inc., or the Staden Package from the Medical Research Council, UK). These are mainly of interest to professionals in the field and will not be discussed further here.

Sonication means subjecting the sample to a brief burst of ultrasonic radiation. This physically breaks the DNA molecule. It is a preferred method for fragmenting genomes, because it results in random breaks and increases the probability of finding overlaps between fragments.

The other main uses for computers in this field are for the storage of sequence data and for its analysis. These activities constitute part of what has become known as **bioinformatics**. Central to bioinformatics has been the development of **databases**. These are repositories of, among other things, sequence information and programs to analyse the data. The data are very highly structured and annotated, with links provided to other relevant information, such as citations to where the particular sequencing results were published. There are many databases to choose from, but one of the most useful is at the **National Center for Biotechnology Information (NCBI)** in the USA (http://www.ncbi.nlm.nih.gov), and this is the one to which reference is usually made in what follows. There is a vast amount of information in this database, and the only way to get an appreciation of it is to go in and look around.

The main repository of DNA sequence data is **GenBank**, which on 15 April 2003 contained 31,099,264,455 bases, from 24,027,936 reported sequences. By the time you read this, there will be much more. It is not too easy to appreciate just how much DNA sequence information that is, and how uninformative the raw data are without computational methods to analyse them. Some feel for the problem might be obtained by reference to Figure 5.22, which gives the sequence of the plasmid pBR322[15] (GenBank accession number J01749). This is a very small molecule by DNA standards, but it still requires more than a page to print it. If we assume that 3000 bases can be printed per page, then the human genome sequence would require 1,000,000 pages and would be the most boring set of books ever produced. Storage on computer is the only feasible option.

```
   1 ttctcatgtt tgacagctta tcatcgataa gctttaatgc ggtagtttat cacagttaaa
  61 ttgctaacgc agtcaggcac cgtgtatgaa atctaacaat gcgctcatcg tcatcctcgg
 121 cacacgtcacc ctggatgctg taggcatagg cttggttatg ccggtactgc cgggcctctt
 181 gcgggatatc gtccattccg acagcatcgc cagtcactat ggcgtgctgc tagcgctata
 241 tgcgttgatg caatttctat gcgcacccgt tctcggagca ctgtccgacc gctttggccg
 301 ccgcccagtc ctgctcgctt cgctacttgg agccactatc gactacgcga tcatggcgac
 361 cacacccgtc ctgtggatcc tctacgccgg acgcatcgtg gccggcatca ccgggcgcac
 421 aggtgcggtt gctggcgcct atatcgccga catcaccgat ggggaagatc gggctcgcca
 481 cttcgggctc atgagcgctt gtttcggcgt gggtatggtg gcaggccccg tggccgggggg
 541 actgttgggc gccatctcct tgcatgcacc attccttgcg gcggcggtgc tcaacggcct
 601 caacctacta ctgggctgct tcctaatgca ggagtcgcat aagggagagc gtcgaccgat
 661 gcccttgaga gccttcaacc cagtcagctc cttccggtgg gcgcggggca tgactatcgt
 721 cgccgcactt atgactgtct tctttatcat gcaactcgta ggacaggtgc cggcagcgct
 781 ctgggtcatt ttcggcgagg accgctttcg ctggagcgcg acgatgatcg gcctgtcgct
 841 tgcggtattc ggaatcttgc acgccctcgc tcaagccttc gtcactggtc ccgccaccaa
 901 acgtttcggc gagaagcagg ccattatcgc cggcatggcg gccgacgcgc tgggctacgt
 961 cttgctggcg ttcgcgacgc gaggctggat ggccttcccc attatgattc ttctcgcttc
1021 cggcggcatc gggatgcccg cgttgcaggc catgctgtcc aggcaggtag atgacgacca
1081 tcagggacag cttcaaggat cgctcgcggc tcttaccagc ctaacttcga tcactggacc
1141 gctgatcgtc acggcgattt atgccgcctc ggcgagcaca tggaacgggt tggcatggat
1201 tgtaggcgcc gccctatacc ttgtctgcct ccccgcgttg cgtcgcggtg catggagccg
1261 ggccacctcg acctgaatgg aagccggcgg cacctcgcta acggattcac cactccaaga
1321 attggagcca atcaattctt gcggagaact gtgaatgcgc aaaccaaccc ttggcagaac
1381 atatccatcg cgtccgccat ctccagcagc cgcacgcggc gcatctcggg cagcgttggg
1441 tcctggccac gggtgcgcat gatcgtgctc ctgtcgttga ggacccggct aggctggcgg
1501 ggttgccttа ctggttagca gaatgaatca ccgatacgcg agcgaacgtg aagcgactgc
1561 tgctgcaaaa cgtctgcgac ctgagcaaca acatgaatgg tcttcggttt ccgtgtttcg
1621 taaagtctgg aaacgcggaa gtcagcgccc tgcaccatta tgttccggat ctgcatcgca
1681 ggatgctgct ggctaccctg tggaacacct acatctgtat taacgaagcg ctggcattga
1741 ccctgagtga tttttctctg gtcccgccgc atccataccg ccagttgttt accctcacaa
1801 cgttccagta accgggcatg ttcatcatca gtaacccgta tcgtgagcat cctctctcgt
1861 ttcatcggta tcattacccc catgaacaga aatccccctt acacggaggc atcagtgacc
1921 aaacaggaaa aaaccgccct taacatggcc cgctttatca gaagccagac attaacgctt
1981 ctggagaaac tcaacgagct ggacgcggat gaacaggcag acatctgtga atcgcttcac
2041 gaccacgctg atgagcttta ccgcagctgc ctcgcgcgtt tcggtgatga cggtgaaaac
2101 ctctgacaca tgcagctccc ggagacggtc acagcttgtc tgtaagcgga tgccgggagc
2161 agacaagccc gtcagggcgc gtcagcgggt gttggcgggt gtcggggcgc agccatgacc
2221 cagtcacgta gcgatagcgg agtgtatact ggcttaacta tgcggcatca gagcagattg
2281 tactgagagt gcaccatatg cggtgtgaaa taccgcacag atgcgtaagg agaaaatacc
2341 gcatcaggcg ctcttccgct tcctcgctca ctgactcgct gcgctcggtc gttcggctgc
2401 ggcgagcggt atcagctcac tcaaaggcgg taatacggtt atccacagaa tcaggggata
2461 acgcaggaaa gaacatgtga gcaaaaggcc agcaaaaggc caggaaccgt aaaaaggccg
2521 cgttgctggc gtttttccat aggctccgcc cccctgacga gcatcacaaa aatcgacgct
2581 caagtcagag gtggcgaaac ccgacaggac tataaagata ccaggcgttt ccccctggaa
2641 gctccctcgt gcgctctcct gttccgaccc tgccgcttac cggatacctg tccgcctttc
2701 tcccttcggg aagcgtggcg ctttctcata gctcacgctg taggtatctc agttcggtgt
2761 aggtcgttcg ctccaagctg ggctgtgtgc acgaaccccc cgttcagccc gaccgctgcg
2821 ccttatccgg taactatcgt cttgagtcca acccggtaag acacgactta tcgccactgg
2881 cagcagccac tggtaacagg attagcagag cgaggtatgt aggcggtgct acagagttct
2941 tgaagtggtg gcctaactac ggctacacta gaaggacagt atttggtatc tgcgctctgc
```

```
3001 tgaagccagt taccttcgga aaaagagttg gtagctcttg atccggcaaa caaaccaccg
3061 ctggtagcgg tggttttttt gtttgcaagc agcagattac gcgcagaaaa aaaggatctc
3121 aagaagatcc tttgatcttt tctacggggt ctgacgctca gtggaacgaa aactcacgtt
3181 aagggatttt ggtcatgaga ttatcaaaaa ggatcttcac ctagatcctt ttaaattaaa
3241 aatgaagttt taaatcaatc taaagtatat atgagtaaac ttggtctgac agttaccaat
3301 gcttaatcag tgaggcacct atctcagcga tctgtctatt tcgttcatcc atagttgcct
3361 gactccccgt cgtgtagata actacgatac gggagggctt accatctggc cccagtgctg
3421 caatgatacc gcgagaccca cgctcaccgg ctccagattt atcagcaata aaccagccag
3481 ccggaagggc cgagcgcaga agtggtcctg caactttatc cgcctccatc cagtctatta
3541 attgttgccg ggaagctaga gtaagtagtt cgccagttaa tagtttgcgc aacgttgttg
3601 ccattgctgc aggcatcgtg gtgtcacgct cgtcgtttgg tatggcttca ttcagctccg
3661 gttcccaacg atcaaggcga gttacatgat cccccatgtt gtgcaaaaaa gcggttagct
3721 ccttcggtcc tccgatcgtt gtcagaagta agttggccgc agtgttatca ctcatggtta
3781 tggcagcact gcataattct cttactgtca tgccatccgt aagatgcttt tctgtgactg
3841 gtgagtactc aaccaagtca ttctgagaat agtgtatgcg gcgaccgagt tgctcttgcc
3901 cggcgtcaac acgggataat accgcgccac atagcagaac tttaaaagtg ctcatcattg
3961 gaaaacgttc ttcggggcga aaactctcaa ggatcttacc gctgttgaga tccagttcga
4021 tgtaacccac tcgtgcaccc aactgatctt cagcatcttt tactttcacc agcgtttctg
4081 ggtgagcaaa aacaggaagg caaaatgccg caaaaaaggg aataagggcg acacggaaat
4141 gttgaatact catactcttc ctttttcaat attattgaag catttatcag ggttattgtc
4201 tcatgagcgg atacatattt gaatgtattt agaaaaataa acaaataggg gttccgcgca
4261 catttccccg aaaagtgcca cctgacgtct aagaaaccat tattatcatg acattaacct
4321 ataaaaatag gcgtatcacg aggccctttc gtcttcaaga a
```

Figure 5.22 The base sequence of the plasmid pBR322 linearized by cleavage with *Eco*RI. Sequences coloured *red* are the restriction sites for *Bam*HI and *Bgl*I (see Figure 5.3). For the significance of the residues in bold, see Worked Problem 5.4

There are many questions that can be asked about a DNA sequence such as that in Figure 5.22. Obvious ones are what restriction sites are present and what open reading frames does it contain? Neither of these questions can be answered by visually examining the sequence, but it easy to write computer programs to do it. The single *Bam*HI site and the three *Bgl*I sites (refer to Figure 5.3) are shown in red in Figure 5.22. Their positions were obtained from the website http://www.fermentas.com/techinfo/NucleicAcids/mappbr322.htm (see Problem 5.3). There are other websites which allow you to submit a DNA sequence and then return the positions of restriction sites of your choice. An example is located at the University of Massachusetts (http://web.umassmed.edu/bioapps/rsites.html).

A word of warning. All the website addresses quoted in this book were current at the time of writing, but they do sometimes change. If this happens, then an attempted log on is often redirected to the new address. Failing that, it is usually possible to find what you want by doing a GOOGLE search. For example, the site giving restriction analysis of pBR322 was found by searching on "pBR322 restriction sites", and the one located at the University of Massachusetts by searching on "restriction site analysis". Any tool you wish to use can be found in this way.

An extremely important application of computers is finding **open reading frames** (or **orfs**, as they are often called) in DNA sequences. The computer reads along the sequence until it finds an initiation codon, and then continues until it reaches a stop codon. This stretch of sequence is potentially a coding sequence. The process is continued until the end of the DNA molecule. It is then repeated, starting at the second residue in the chain rather than the first; that is, it checks the second reading frame.

Finally, for the strand initially selected, the third potential reading frame is examined. The whole process is then repeated for the other DNA strand.

Worked Problem 5.4

Q The plasmid pBR322 contains a gene conferring tetracycline resistance (Figure 5.3). Locate it in the sequence given in Figure 5.22.

A The first thing to do is to search for open reading frames. There is a tool for doing this at the NCBI website.

Log on to the website, and in the Home Page, click on *Tools* (in the panel on the left hand side of the screen) and then in the new page click on *ORF Finder*. In the box *Enter GI or Accession Number* enter *J01749* (the accession number for pBR322). Click on *OrfFind*. A new page opens with the results as shown in Figure 5.23 (note the colour has been changed).

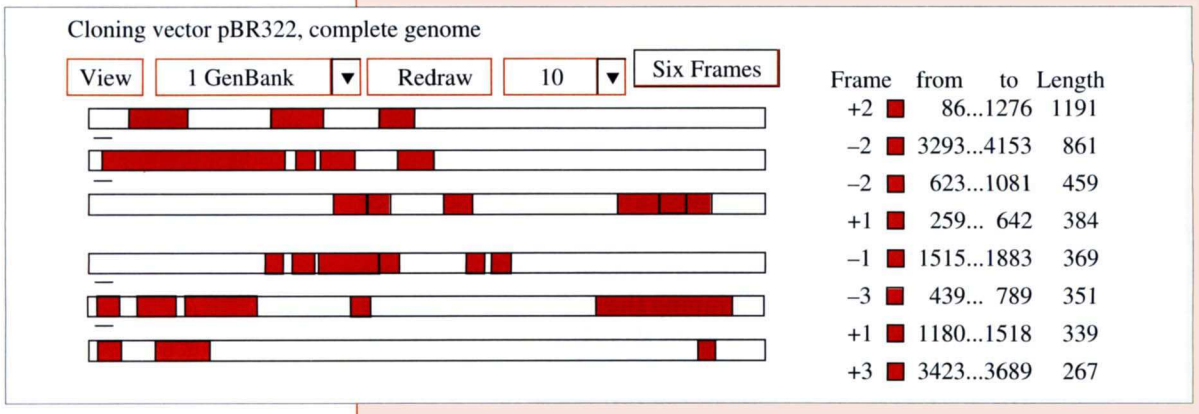


Figure 5.23 Output from the program OrfFind (note that the list on the right-hand side has been truncated to show only the longer orfs)

Clicking on *View* will take you to a page giving the base sequence and other information about entry J01749. The bars on the left of the screen show the positions of the orfs (coloured red here) in the six reading frames. The list of the right of the screen shows the length of the orfs in decreasing order, their positions in the sequence, and the reading frame in which they occur. The long orf at the beginning of the second line looks a promising candidate – it is roughly matches the position and length shown in Figure 5.3. We will show below that it is indeed the coding sequence for the tetracycline resistance gene. The sequence is shown in bold in Figure 5.22.

It is essential to note that this simple approach is applicable only to prokaryotes. With eukaryotes, we have the complicating problem of split genes (see Section 4.2). To identify protein coding regions of the genome we need ways of identifying the intron/exon boundaries. The ways of doing this are still being worked out, and we will not consider the matter further. Suffice it to say that it involves recognizing particular patterns in the DNA sequence, which again is a process which can only be done using computers.

Once open reading frames have been found, the question arises as to which proteins they code for; that is, what is the function of that protein. The answer to this question is provided (if it can be provided) by **database searching**. The predicted amino acid sequence is compared with those of all of the proteins in a database looking for sequence similarities; that is, for regions where identical amino acids occur in the two proteins, or where the amino acids are of similar types. This is usually done using a program called **BLAST** (for *B*asic *L*ocal *A*lignment *S*earch *T*ool[16]). In the simplest situation (see Worked Problem 5.5) the sequence is found to be identical to that of a protein already known, in which case the identification is obvious.

Worked Problem 5.5

Q Confirm that the open reading frame identified in Worked Problem 5.4 does indeed code for the tetracycline resistance protein.

A Clicking on the coloured bar tentatively identified as the required open reading frame opens a new page which is similar to that shown in Figure 5.23, but which also contains the predicted amino acid sequence of the protein coded (not shown here). The page also offers the opportunity to search a protein sequence database using BLAST with the selected sequence as the query. Clicking on *BLAST* runs the program and produces an output showing a large number of proteins with sequences identical to that of the query. They are all tetracycline resistance proteins from a variety of organisms. The protein from pBR322 is included in the list (accession code YTEC32). The identity of the product of the orf is, therefore, confirmed.

In other cases, the protein might be found to be related to, rather than identical with, a protein that had been previously described. For example, consider the amino acid sequence labelled Query in Figure 5.24. It was derived from an open reading frame in the genome of the bacterium *Helicobacter pylori*. A BLAST search of the SwissProt database

(see Worked Problem 5.6 for how this is done) revealed that the sequence was related to those of several proteins that had already been identified as the enzyme aspartate aminotransferase from a variety of bacterial sources. Figure 5.24 shows the alignment of the query sequence with that of the enzyme from *Bacillus subtilis*. The two sequences show 48% of identical residues (indicated by the entries in the middle line of the comparison), and in many cases where the residues are different, they are of similar types (indicated by + signs in the middle line). This high degree of similarity shows that the protein from *H. pylori* is also aspartate aminotransferase, the differences in amino acid sequence compared with the enzyme from *B. subtilis* having arisen by evolutionary changes since the time that the two organisms diverged from a common ancestor.

```
P53001
AAT1_BACSU Aspartate aminotransferase (Transaminase A)(ASPAT) [Bacillus subtilis]

Identities = 189/392 (48%), Positives = 274/392 (69%), Gaps = 7/392 (1%)

Query: 1   MPYSSKVQSLSESATIAISTLAKELKSQGKDILSFSAGEPDFDTPQAIKDAAIKALNDGF 60
           M  + +V +L+ S T+AI+  AKELK+ G D++   AGEPDF+TPQ I DAA++++N+G
Sbjct: 1   MKLAKRVSALTPSTTLAITAKAKELKAAGHDVIGLGAGEPDFNTPQHIIDAAVRSMNEGH 60

Query: 61  TKYTPVAGIPELLKAIAFKLKKENNLDYEPSEILVSNGAKQSLFNAIQALIGEGDEVVIP 120
           TKYTP  G+ EL  +IA K K++ N++Y+PS+I+V  GAK +L+   Q ++ E DEV+IP
Sbjct: 61  TKYTPSGGLAELKNSIAEKFKRDQNIEYKPSQIIVCTGAKHALYTLFQVILDEEDEVIIP 120

Query: 121 VPFWVTYPELVKYSGGVSQFIQTDEKSHFKITPKQLKDALSPKTKMLILTTPSNPTGMLY 180
            P+WV+YPE VK +GG   +++  E++HFKI+P+QLK+A++ KTK +++ +PSNPTG++Y
Sbjct: 121 TPYWVSYPEQVKLAGGKPVYVEGLEENHFKISPEQLKNAITEKTKAIVINSPSNPTGVMY 180

Query: 181 SKAELEALGEVLKDTKVWVLSDEIYEKLVYKG-EFVSCAAVSEEMKKRTITINGLSKSVA 239
           ++ EL ALGEV  +  + ++SDEIYEKL Y G + VS A +S+ +K++T+ ING+SKS +
Sbjct: 181 TEEELSALGEVCLEHDILIVSDEIYEKLTYGGKKHVSIAQLSDRLKEQTVIINGVSKSHS 240

Query: 240 MTGWRMGYAASKDKKLVKLMSNLQSQCTSNINSITQMASIVALEGLVDKEIETMRQAFEK 299
           MTGWR+GYAA  +  ++K M+NL S  TSN  SI Q  +I A  G   + +E MR+AFE
Sbjct: 241 MTGWRIGYAAGSE-DIIKAMTNLASHSTSNPTSIAQYGAIAAYNG-PSEPLEEMREAFEH 298

Query: 300 RCHLAHAKINAIEGLNALKPDGAFYLFIN---IGSLCG-GDSMRFCHELLEKEGVALVPG 355
           R +  +AK+  I G + +KP+GAFYLF N       CG  D   F   LLE+E VA+VPG
Sbjct: 299 RLNTIYAKLIEIPGFSCVKPEGAFYLFPNAKEAAQSCGFKDVDEFVKALLEEEKVAIVPG 358

Query: 356 KAFGLEGYVRLSFACSEEQIEKGIERIARFVK 387
             FG    VRLS+A S + +E+ IERI RFV+
Sbjct: 359 SGFGSPENVRLSYATSLDLLEEAIERIKRFVE 390
```

Figure 5.24 Identification of a protein by similarity searching

Function can be assigned to an unknown protein from sequence similarities that are much more remote than that in Figure 5.24, as demonstrated in Worked Problem 5.6.

Worked Problem 5.6

Q A piece of DNA from the mushroom *Armillaria mellea* was sequenced and translated into the corresponding protein. The protein sequence obtained (164 residues) was as follows:

```
AGPDFDLDYRTYPQSSENICYSWFCNNGPHSVAPDRTHAAAHRASNS
CGNVNPNRCSIRV
GHVSGYQCDEWPWANSNAGGANAATRCIPTADNTGSGSQWGNFINNR
GSQAVGYVLQDNV
VFATIEISNIPTTAEFCKGVLGTAITATMCRQVANGQPYLQRIG
```

Carry out a BLAST search of the SwissProt data base in an attempt to assign a function to this protein. The SwissProt database and the required program can be found at the website of the **Swiss Institute of Bioinformatics** (SIB), address http://ca.expasy.org.

A Log on to the SIB website. On the home page, go to *Proteomics and Sequence Analysis Tools* and select *Similarity Searches [BLAST]*. In the new page that opens up, type or paste the sequence given into the sequence box. (It is best to type the sequence in advance using a word processor, copy it, and paste into the box; this saves time. If you are using a sequence that is already deposited in a database as the query, then all that is necessary is to enter the database ID or accession number in the box – the program will retrieve the sequence for you). In the section *Choose the appropriate BLAST program and database* select *Swiss-Prot*. Leave *Blastp* as the default program (this searches a protein sequence against a protein database; other varieties of blast search nucleic acid sequences against nucleic acid sequence data bases, translated nucleic acids against protein databases, and so on). You may enter your E-mail address if you want the results sent to you, but this is unnecessary when running BLASTP. Finally, click on *RUN*. An edited version of the output is shown in Figure 5.25. This was obtained choosing *Plain Text* as the output format; HTML is prettier, but not so easy to reproduce.

When making a copy of the output from any sequence comparison program using a word processor, you should use **Courier New** as the font. The reason is that, in this font, all the characters are of the same size. With fonts for which this is not the case, the sequence alignment is lost.

```
BLASTP 2.2.5 [Nov-16-2002]


Reference: Altschul, Stephen F., Thomas L. Madden, Alejandro A. Schaffer,
Jinghui Zhang, Zheng Zhang, Webb Miller, and David J. Lipman (1997),
"Gapped BLAST and PSI-BLAST: a new generation of protein database search
programs",  Nucleic Acids Res. 25:3389-3402.

Query= Submission (164 letters)

Database: XXswissprot 131,155 sequences; 49,821,033 total letters


                                                                  Score    E
Sequences producing significant alignments:                       (bits)  Value

sp!P42983!NUCB_BACSU Sporulation-specific extracellular nuclease...    31   0.94
sp!O54898!CCAG_RAT Voltage-dependent T-type calcium channel alph...    30   1.6
sp!P12667!NUCA_BACSU DNA-entry nuclease (EC 3.-.-.-) (Competence...    30   2.7

>sp!P42983!NUCB_BACSU Sporulation-specific extracellular nuclease precursor (EC 3.-.-.-)
          [NUCB] [Bacillus subtilis]
         Length = 136

 Score = 31.2 bits (69), Expect = 0.94
 Identities = 14/38 (36%), Positives = 17/38 (44%)

Query: 65  GYQCDEWPWXXXXXXXXXXXTRCIPTADNTGSGSQWGN 102
           GY  DEWP              R +  +DN G+GS  GN
Sbjct: 83  GYDRDEWPMAVCEEGGAGADVRYVTPSDNRGAGSWVGN 120
```

Figure 5.25 Edited output from a BLAST sequence search

The first point to note is that the SwissProt database, at the time this search was done, contained 131,155 sequences with a total of 49,821.033 letters (residues). The essential results start with a list of sequences producing significant alignments (only the first three have been retained in this edited output). Each entry gives the identity of the matching protein, and a measure (the E value) of the significance of the match. Values around 1 or less are usually of interest. Next come partial sequence alignments for the matches; only the first has been kept. The results show the part of the query sequence for which the best match was obtained. The match starts at residue 65 of the query and aligns with the sequence of protein P42983 from residue 83. The string of Xs in the query sequence indicates that the residues occurring in that part of the protein are of the most commonly occurring type (Ala, Ser, Asn, Gly) and so they have not been taken into account in the analysis; a match between the two sequences in that region might well occur by chance.

The relationship between the two sequences may not look too convincing. It is somewhat more obvious if a full sequence alignment is carried out. This can be done using the program CLUSTAL W,[17] which is also available through the SIB website. The result of the alignment is shown in Figure 5.26. A region of quite substantial similarity between the sequences can now be seen in the central section of the comparison.

```
                10        20        30        40        50        60
                 |         |         |         |         |         |
Query      -----AGPDFDLDYRTYPQSSENICYSWFCNNGPHSVAPDRTHAAAHRASNSCGNVNPNR
P42983     MKKWMAGLFLAAAVLLCLMVPQQIQGASSYDKVLYFPLSRYPETGSH-IRDAIAEGHPDI
           **  :           .::*  :   ::  :   .  ..:.:*   :: .: :*:
               70        80        90       100       110       120
                |         |         |         |         |         |
Query      CSIR--------------VGHVSGYQCDEWPWANSNAGGANAATRCIPTADNTGSGSQWG
P42983     CTIDRDGADKRREESLKGIPTKPGYDRDEWPMAVCEEGGAGADVRYVTPSDNRGAGSWVG
           *:*                :   .**: **** * .: ***.* .* :..:** *:**  *
               130       140       150       160       170       180
                |         |         |         |         |         |
Query      NFINNRGSQAVGYVLQDNVVFATIEISNIPTTAEFCKGVLGTAITATMCRQVANGQPYLQ
P42983     NQMSS-------YPDGTRVLFIVQ-----------------------------------
           * :..       *    .*:* .
```

Figure 5.26 Full alignment of the amino acid sequences in Figure 5.25

The point of all this is that the matching protein in the database is a nuclease; that is, an enzyme that hydrolyses DNA. Hence our unknown protein is almost certainly a nuclease. Subsequent experimental evidence obtained with the purified protein showed that this is indeed the case.[18]

In this case, the two enzymes do not do exactly the same thing, but both of them are nucleases. This reflects the fact that proteins in living organisms fall into families with related functions. Members of these families have arisen by evolution of function from one or a small number of primitive proteins in our ancient ancestors. As a result of this family relationship, the amino acid sequences of the members are related even if, as in the case in Worked Problem 5.6, the relationship is very distant. It is still sufficient, however, to provide the clue to what the unknown protein does.

In summary, then, the approach to assigning function to a protein whose amino acid sequence has been revealed by genome sequencing is to compare the sequence with those of all known proteins and to look for similarities. Similarities may allow definite, or at least tentative,

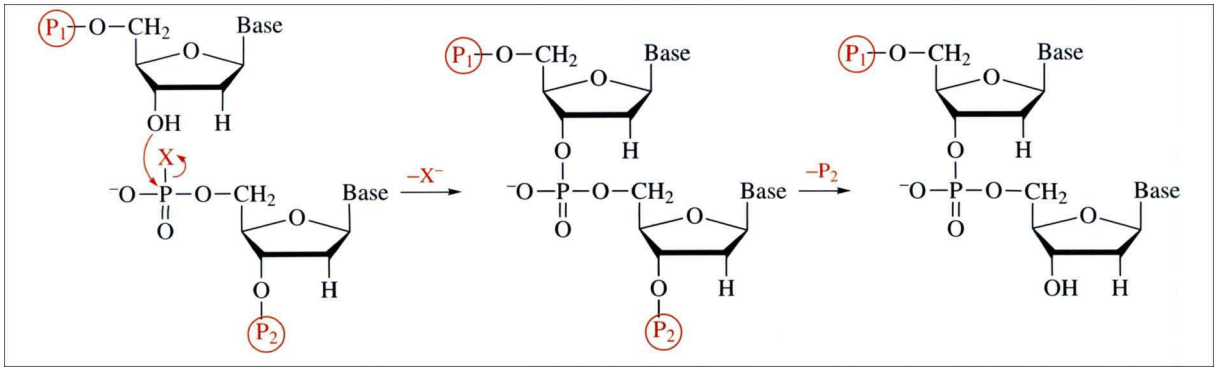
assignment of function. Unfortunately, this does not always work. We still have only a partial knowledge of the complete range of protein structures and functions, and it is usually the case that a substantial fraction of the proteins coded by a newly sequenced genome cannot be identified in this way. There is still a long way to go before knowledge of the sequence of a genome will allow us to completely catalogue the proteins that an organism can make.

5.10 Chemical Synthesis of Oligonucleotides

This section describes specifically the synthesis of oligodeoxyribonucleotides. Synthesis of small RNA species follows the same general principles but is more complicated because of the OH group on the 2′-position of the ribose ring. Although chemical synthesis of these molecules is important in some applications, we will not pursue it further here.

It should be clear from what has been said above that many of the modern methods of DNA analysis require the availability of synthetic oligonucleotides. Techniques such as PCR or dideoxy sequencing would be impossible without them. We will, therefore, finish this review of the tools of DNA chemistry with an account of how these molecules are made.

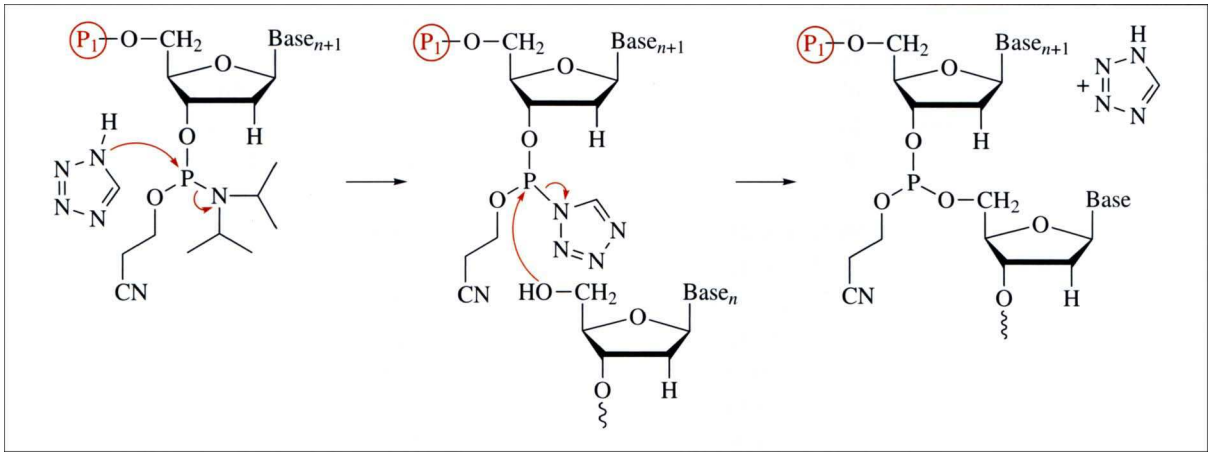
The general approach to synthesis of oligonucleotides was developed by Khorana and his group in the 1970s. The essence of the method that they used is outlined in Scheme 5.9. The 5′-OH of a deoxynucleoside was protected with a group P_1 to eliminate its nucleophilic character. The 3′-OH of a deoxynucleotide was similarly protected with a second group P_2, and the phosphate activated with a good leaving group X. Reaction of these two species produced the desired phosphodiester linkage and yielded a protected dinucleotide. Subsequent removal of protecting group P_2 then allowed for extension of the polynucleotide from the 3′-end. It was important that the two protecting groups could be removed under different conditions so that P_1 stayed in place when P_2 was removed; that is, the 5′-OH of the dinucleotide continued to be protected. Note that three of the four bases also required protection; this is discussed below.



Scheme 5.9

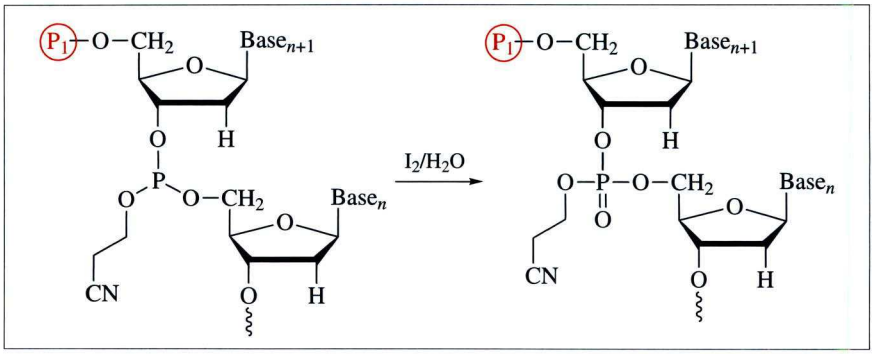
This synthetic method was enormously powerful, and the crowning achievement of Khorana and his co-workers was the complete synthesis of a 126-residue long gene for a tRNA molecule.[19] There have been other significant advances in methodology in the last 20 years, and it is these newer methods that will be described in more detail below. Note, however, that the general approach used is still that pioneered by Khorana.

One of the most significant advances in oligonucleotide synthesis has been in the chemistry of formation of the phosphodiester linkage. This is now almost always done using what is known as **phosphoramidite** chemistry. The technique was introduced by Caruthers and his colleagues and has been reviewed.[20] The reactions involved are outlined in Scheme 5.10.



Scheme 5.10

The chain is built up from the 3′-end, and the next residue is added to the 5′-OH of the growing chain. The derivative used for chain elongation is the protected 2-cyanoethylphosporamidite shown on the left of Scheme 5.10. These compounds are relatively unreactive, but in the presence of tetrazole, reaction occurs to form a more reactive phosphorotetrazolide intermediate. This intermediate then reacts with the 5′-OH of the growing nucleotide chain to form the required phosphorus-oxygen linkage. This is obviously not the desired final product; the phosphorus is in the form of a trivalent phosphite rather than the pentavalent phosphate. After coupling, therefore, the product is reacted with aqueous iodine which oxidizes the phosphorus (Scheme 5.11).



Scheme 5.11

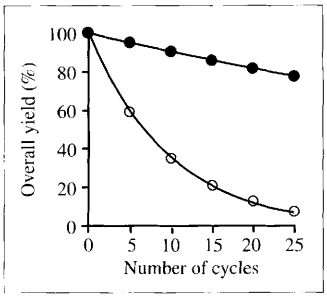
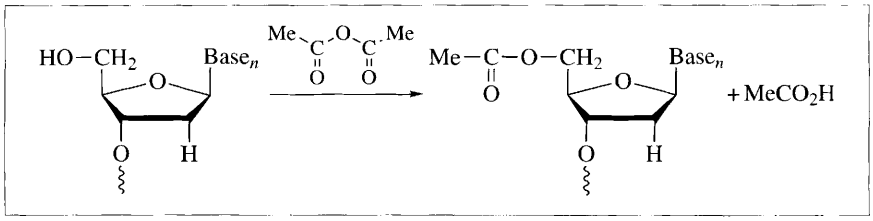


Figure 5.27 Dependence of overall yield on step yield. *Open circles*: step yield 90%; *closed circles*: step yield 99%

This method of making the phosphodiester linkage has several advantages, principal among which is that it can be made to proceed in very high yield (greater than 99%). This is crucial if oligonucleotides of any significant length are to be made. The reason is that overall yield of the product is strongly dependent on step yield. This is illustrated in Figure 5.27, where the overall yield is plotted as a function of the number of synthetic cycles for step yields of 90% and 99%. Clearly a step yield of 90% is unacceptable for synthesis of a 25mer since the overall yield is less than 10%. Even a step yield of 99% results in a final yield of only about 80%, but it is not possible to do better than this.

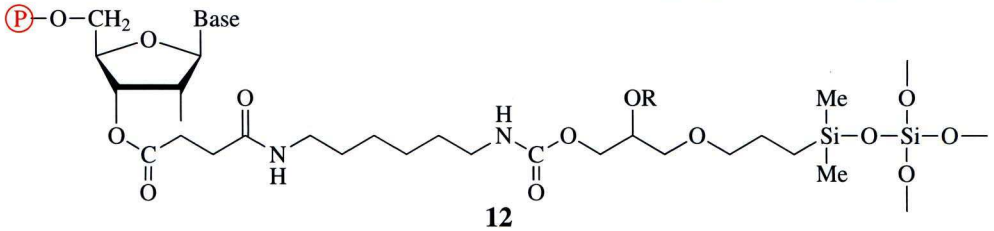
There is another problem associated with incomplete coupling. Any residual free 5′-OH groups can participate in reaction in the following cycle, thus leading to products with internal deletions. This will result in a mixture of products that would be hard to purify. For this reason, at the end of each reaction cycle a **capping** process is carried out. The product is reacted with ethanoic anhydride (acetic anhydride), with the result that any free 5′-OH groups are ethanoylated (acetylated) and not available for subsequent reaction (Scheme 5.12).



Scheme 5.12

The other major advance in oligonucleotide synthesis has been the development of what are known as **solid phase** methods. This means that

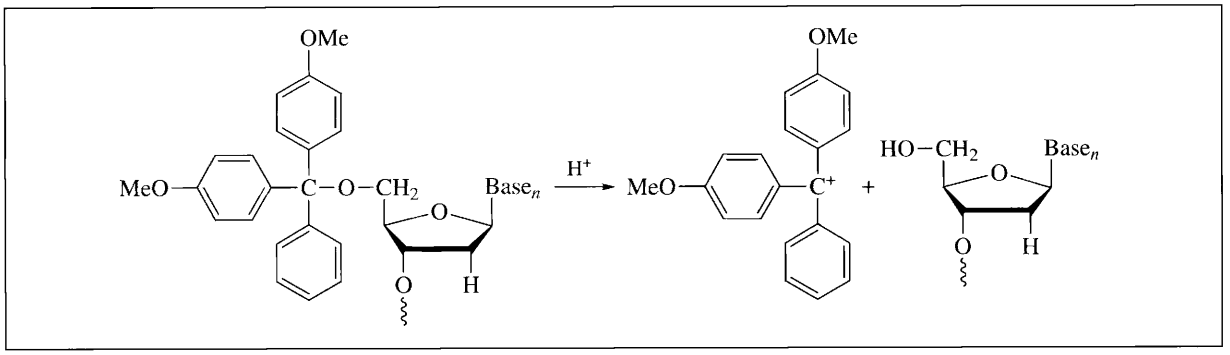
the synthesis is not carried out in free solution, but rather with the growing oligonucleotide chain attached to a solid support, most usually a glass bead. The beads are packed into a column, and all the chemistry is carried out by passing reagents through the column. Similarly, side products of the reactions are removed simply by passing appropriate solvents through the column. Synthesis starts with the protected nucleoside that is to be the 3′-residue attached to the glass support *via*an ester linkage to the 3′-OH group. The structure of a typical starting material is shown in **12**.



12

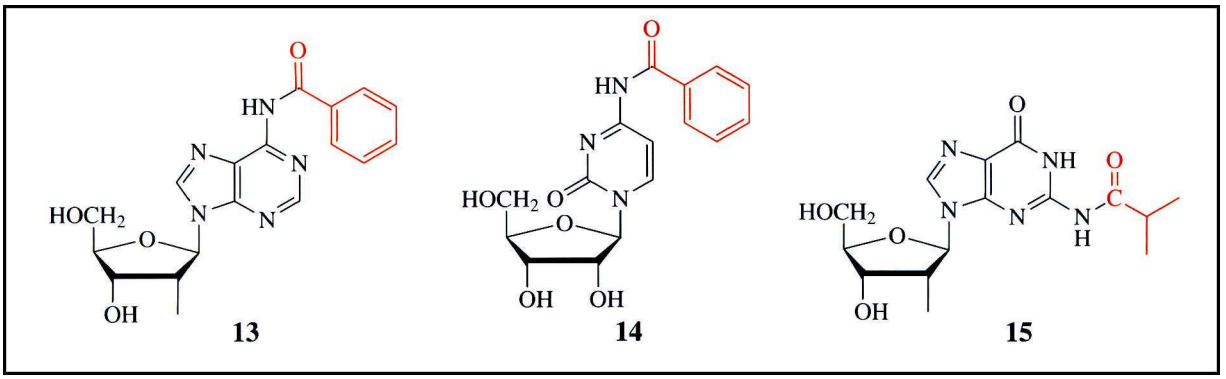
There is a long linker joining the protected nucleoside to the glass. The point of this is to hold the growing chain away from the glass bead to ensure that reaction is not restricted by steric interactions with the support. Note that it is the linkage to the glass that provides protection for the 3′-OH group of the first residue in the chain.

We need to look now at the other protecting groups. As shown in Scheme 5.10, the 5′-OH group of the incoming residue must be protected to prevent it from participating in the coupling reaction. After that reaction has occurred, the protecting group must be removed so that the 5′-OH is free to react in the next synthetic cycle. A major requirement is, then, that the protecting group must capable of being removed under conditions where neither the linkage to the glass support nor the newly formed inter-nucleoside linkage is affected. An ideal blocking group is **dimethoxytrityl**. The structure of a growing chain with a dimethoxytritylated 5′-OH is shown in Scheme 5.13. This group is quite labile and is readily removed by brief exposure to trichloroethanoic acid (trichloroacetic acid) in dichloromethane to yield the dimethoxytrityl cation (Scheme 5.13). An added advantage of using dimethoxytrityl as the protecting group is that the cation is highly coloured, and its release from the protected derivative can be monitored by spectrophotometry to ensure that de-protection is complete.

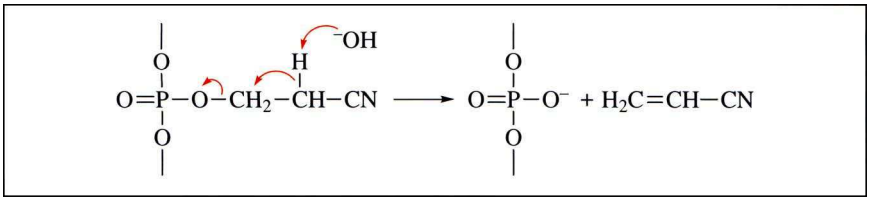


Scheme 5.13

Three of the four bases contain reactive amino groups that require protection during oligonucleotide synthesis. The protecting groups used have to be stable to the mild acid conditions used to remove the 5′-dimethoxytrityl group, but sufficiently labile that they can be removed at the end of the synthetic process. The derivatives usually used are N^6-benzoyladenosine (**13**), N^4-benzoylcytosine (**14**) and N^2-isobutyryl-guanosine (**15**).



When synthesis is complete, the oligonucleotide is removed from the solid support, the bases de-protected, and the 2-cyanoethyl group removed from the phosphate. All of these can be achieved by treatment with aqueous ammonia. Breakage of the link to the glass support and the removal of the base protecting groups are base-catalysed hydrolysis reactions. Removal of the 2-cyanoethyl group is a base-catalysed β-elimination reaction, yielding propenenitrile (acrylonitrile; Scheme 5.14).



Scheme 5.14

Finally, the oligonucleotide may require purification. Even if each step of synthesis proceeds 99% to completion, a 50mer will contain only 60% of the desired product, the remainder being shorter sequences. The best way to achieve purification is by using either polyacrylamide gel electrophoresis or high-performance liquid chromatography.

A great virtue of the solid phase method for oligonucleotide synthesis is that it is amenable to automation, and indeed there are now several machines on the market for automated oligonucleotide synthesis. These machines have a small column into which is loaded the glass support with the 3′-residue attached, and reagent reservoirs to contain the protected and activated nucleosides and other reagents used during the cycle. The desired sequence is then entered into a microprocessor, and the machine does the rest. The cycle time is a matter of only a few minutes and so oligos up to 100 residues long can be synthesized in a day. There are also many companies that offer custom synthesis, and these large centres may be making up to 10,000 oligos per day. Because of the enormous efficiency of the process, prices of synthetic oligos are now very low, typically £0.30 or US $0.50 per base; for many laboratories where the requirement for oligos is only modest, it is more economic to have them custom made rather than to invest in an automated synthesizer.

Box 5.9 DNA Microarrays

The availability of simple and rapid methods of oligonucleotide synthesis has lead to the development of an entirely new approach to analysing gene expression called **DNA microarray** technology (or sometimes **DNA chip** technology). Suppose that we wish to analyse which genes are being expressed in a human tissue at any particular time; that is, to determine which mRNA molecules are being synthesized. Synthetic oligos, typically 25mers, specific for each of the genes in the genome are deposited in minute spots on the surface of a silicon chip. Using robotic methods, up to a million spots can be deposited in precisely determined positions on the surface. If the chip is then incubated with an mRNA extract from the tissue in question, the mRNA molecules will hybridize with the oligos on the chip. The mRNA molecules are fluorescently labelled, so that

examination of the chip after hybridization reveals to which oligos the mRNA species have hybridized, and allows their identification.

This technology has enormous potential applications. For example, the pattern of gene expression varies in many disease states, and so chip technology can be used as a diagnostic tool. Similarly, expression patterns change in response to treatment of some diseases and so chips can be used to monitor the success and progress of treatment.

Box 5.10 DNA Fingerprinting

DNA fingerprinting is a subject which has attracted considerable popular interest in recent years. It is also a technique which makes use of several of the tools that have been described in this chapter. It seems, therefore, an appropriate topic on which to end.

In order to understand how DNA fingerprinting works, we need to know a little more about the organization of the human genome. Of the 3000 Mbp of DNA in the genome, less than 20% is associated with genes and their control elements, and of this only about 10% (that is, 2% of the entire genome) codes for proteins or RNA molecules. The rest is largely in the introns.

It is easy to see that this is about right. The human genome contains about 30,000 genes. Suppose that, on average, there are about 2000 coding residues in each (enough for a protein with about 650 amino acids). That gives 60 Mbp of coding DNA, or 2% of the genome.

What about the remaining 80% of the genome, or **extragenic DNA**, as it is called? About 80% of this, or 60% of the genome, is contained in unique sequences and the rest is in **repetitive sequences**. It is with the latter that we are concerned here. The extragenic DNA is frequently referred to as "junk DNA", because it has no known function. This term should perhaps be used with caution, because the junk may have functions of which we are currently unaware.

The repetitive DNA is of two types known as **tandem repeats** and **interspersed repeats**. In the tandem repeats, a particular nucleotide sequence is repeated many times at a given locus in the genome, and there are many thousands of these loci. So, for example, in a particular type of DNA called **minisatellite DNA** there may be between 10 and 1000 repeat units at a given locus, and the repeat units may be between 10 and 100 nucleotides long. All this gives virtually limitless scope for differences, or **polymorphisms**, between individuals. The main difference is in the number of repeat units at a particular locus (this is consistent with the idea that the DNA is junk since it does not seem to matter how much of it there is). The polymorphism arising from this is called **variable number of**

tandem repeat, or **VNTR**, polymorphism. In the case of interspersed repeats, as the name suggests, the repeated sequences are not arranged in tandem arrays, but rather they are dotted all over the genome interspersed with the unique sequences.

We will deal first with the **single locus** approach to what is properly known as **DNA profiling** (the term fingerprinting is usually reserved for the multilocus method described latter). Consider the very highly simplified locus shown on the left hand side of Figure 5.28. Four possible polymorphisms are shown where there one, two, three or four copies of the repeated element. Downstream is a common section of DNA that can be used as a hybridization site for an oligonucleotide probe. On either side of the tandem repeat there is a cleavage site for a particular restriction enzyme. Digestion of the DNA with the restriction enzyme will lead to four DNA fragments of different lengths and, for this reason, the locus is said to show **restriction fragment length polymorphism** (**RFLP**, pronounced *riflip*). Remember that in practice there may be many hundreds of tandem repeats at a locus and a correspondingly large number of possible restriction fragment lengths.

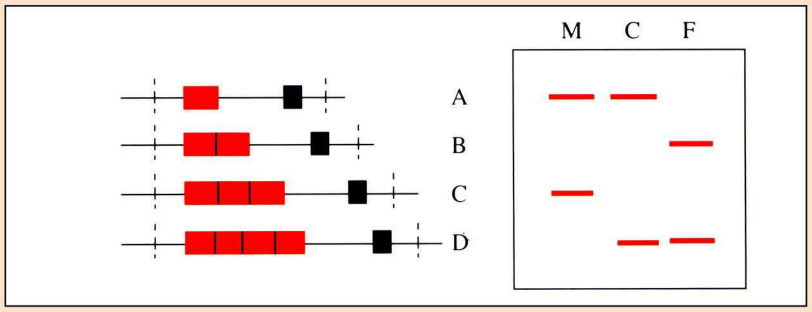


Figure 5.28 VNTR polymorphism and its use in DNA profiling. The left-hand side shows a locus where there are 1, 2, 3 or 4 repeats of a particular sequence (*red boxes*). The *black box* is a site for which an oligo probe is available. The *vertical dashed lines* show the positions of a restriction site. On the right-hand side is a Southern blot showing the profiles for a mother (M), father (F) and one of their children (C)

The fragments can be separated by gel electrophoresis, Southern blotted, and visualized using a radioactive probe. A possible set of results for the members of a family are shown in the right hand side of the figure. The mother shows two bands attributable to polymorphisms D and B (remember that large fragments migrate more slowly than small ones), the father shows two bands attributable to C and A. What would be expected for a child of the family? Clearly, the child will have inherited an allele from its father and one from its mother, so the pattern will contain one band from each of the parental profiles. In the case shown this is D and A.

This is essentially the method used to establish kinship. If the child in the hypothetical example had shown a pattern containing

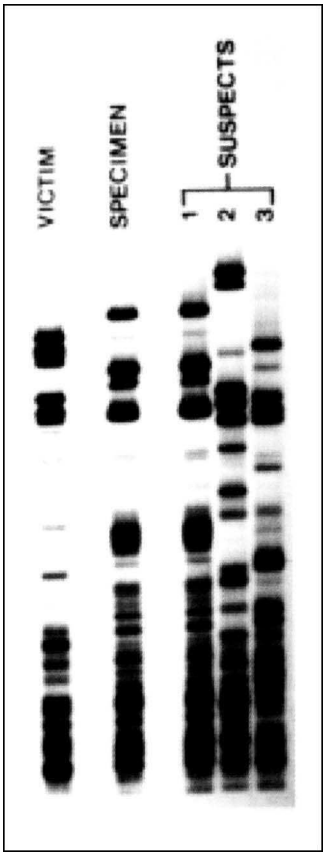


Figure 5.29 DNA fingerprints of a victim, a specimen from the crime scene, and specimens from three suspects (courtesy of Cellmark Diagnostics, Abingdon, UK)

It is important to note that current practice is to digitize DNA profiles and to present the results graphically, rather than as the traditional gel patterns as shown in Figure 5.29. This allows for detailed statistical analysis of the probability that, for example, the profiles from a forensic sample uniquely identifies a crime suspect.

one of the bands from the maternal pattern, but neither of the bands from the pattern of the putative father, then this would constitute strong evidence that the man concerned was not the father. Finding an identical pattern using a single locus, on the other hand, is not adequate positive proof of parentage. The profile from a child might show bands identical to one from each of two randomly selected people purely by chance. What has to be done is to use a combination of two or more probes which are specific for different loci. There is set of five probes in common use for which it has been estimated that the chance of two unrelated individuals giving the same composite profile is 1 in 10^{13} against.

In the method described above, the polymorphisms are detected by using a probe which is specific for a particular locus. However, as stated above, tandem repeats of a particular sequence occur at many loci in the genome, and so if the repeat sequence itself is used as the probe, then many loci can be sampled at the same time. That is, we will have a **multilocus profile**. This is the approach which was originally developed by Alec Jeffreys, and for which he coined the term DNA fingerprinting.[21] The probes that he and his group developed detect repeats with the core sequence GGGCAGGANG, and were found to detect tandem arrays containing between three and 40 repeats at about 20 different positions in the genome. The fingerprints obtained using these probes are, therefore, very rich in information and can produce fingerprints that are completely specific for any individual. A typical example is shown in Figure 5.29.

A sample was obtained from a crime scene and DNA fingerprints obtained from the victim, the sample, and three suspects of the crime. It can be seen that the fingerprints unambiguously identify the perpetrator of the crime as suspect 1.

The original methods used for DNA fingerprinting required a relatively large amount of material. This limitation has, however, been overcome by combining the technique with PCR. Primers are used which amplify the minisatellite DNA and the product can then be used for fingerprinting. If the first few cycles of PCR are carried out under less stringent hybridization conditions than normally used, then a large number of loci will be amplified. This is called the randomly amplified polymorphic DNA (**RAPD**) method, and gives rise to DNA products that produce very specific fingerprints. It is this combination of fingerprinting with PCR which has made it possible to identify an individual from a sample as small as a tiny speck of blood or even the root of a single hair.

Summary of Key Points

1. Separation of DNA molecules can be achieved by electrophoresis through gels of agarose or polyacrylamide. The separation is based on chain length, with smaller molecules migrating more rapidly than large ones.

2. Restriction enzymes cut double-stranded DNA molecules at specific base sequences to produce fragments with either sticky ends or blunt ends. They are used to produce defined fragments of DNA from larger molecules.

3. DNA molecules separated by electrophoresis can be blotted onto mechanically stable membranes, and then identified by hybridization to synthetic oligonucleotides.

4. DNA fragments that have been produced by restriction enzymes can be joined using DNA ligase. The products are called recombinant molecules.

5. Foreign DNA molecules can be inserted into vectors, which are often plasmids, and transformed into bacterial host cells. Individual bacteria can then be grown into colonies called clones. This technique allows both for purification and for amplification of DNA.

6. If cloning of a protein-coding gene is carried out using an expression vector, then the transformed bacterium is able to synthesize the protein.

7. The polymerase chain reaction is a copying technique that provides a simple method for the amplification of small quantities of DNA molecules.

8. DNA molecules are usually sequenced by using the dideoxy chain termination method. The method requires single-stranded DNA, which is produced using cloning vectors based on the genome of phage M13.

9. Large DNA molecules can be sequenced by the shotgun cloning of fragments, and then overlapping the sequenced fragments.

10. Automated DNA sequencing has allowed the determination of the structures of complete genomes.

11. Computers are essential in DNA chemistry. Their applications include analysis of sequence data, storage and analysis of completed sequences, and assignment of function to coding sequences by similarity searching.

12. Chemical synthesis of oligonucleotides is done using phosphoramidite chemistry. Automation of the method has been essential to produce the large number of oligonucleotides required for applications such as DNA microarray technology.

Problems

5.1. What mixture of oligonucleotides would be required identify the DNA coding for a protein that contained the partial amino acid sequence Met-Tyr-Trp-His-Met? What range of melting temperatures would you predict for the mixture of oligonucleotides? If the next residue in the sequence was Gly, how could the length, and hence the specificity, of the probe be increased without increasing the number of oligonucleotides required?

5.2. Explain why recombinant plasmids can be made from DNA that has been restricted with *Bam*HI and a plasmid that has be opened with *Bgl*II (refer to Table 5.1).

5.3. Locate the exact position of the restriction site for *Pvu*I in the sequence of pBR322 given in Figure 5.22 (you could attempt to do this by eye, but it is much better to use a web resource!).

5.4. Suppose that you wanted to amplify the linearized pBR322 molecule in Figure 5.22 by PCR. What pair of 15mer oligonucleotides would you use to do it?

5.5. Read as much as possible of the DNA sequence from the gel in Figure 5.12. Remember that you will be reading the sequence in the 3′→5′ direction. Note that the bands are of unequal intensities, and the separation is not regular. Take special care when there are runs of the same residue not to miss one.

5.6. The following are the sequences of three fragments from a shotgun DNA sequencing experiment. Assemble them into a single sequence. Allow for the possibility that all three fragments may not

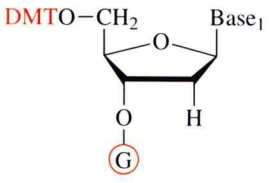
be from the same strand. (Note that for convenience the sequences given are very short with correspondingly small overlaps. In practice, the sequences would be expected to be 300–400 residues long):

5′ tgaagtggtggcctaactacggct
5′ gacagtatttggtatctgcgctctgc
5′ aatactgtccttctactgtagccg

5.7. Refer to Figure 5.23. Confirm that the open reading frame at the right-hand side of the fifth bar corresponds to the coding sequence for the ampicillin resistance gene. Identify the exact position of the coding sequence in Figure 5.22. From which strand of pBR322 is it transcribed?

5.8. Identify the gene whose partial sequence was obtained in Problem 5.5 by doing a BLAST search of GenBank.

5.9. Give a sequence of reactions for the incorporation of the second nucleotide in a synthetic oligo starting from **16**. It is unnecessary to specify the nature of the base. The red sphere represents the glass bead to which the 3′-residue is attached.



16

References

1. S. Olson, *Mapping Human History: Discovering Our Past Through Our Genes*, Bloomsbury, London, 2002.
2. H. O. Smith and K. W. Wilcox, *J. Mol. Biol.*, 1970, **51**, 379.
3. C. M. Lukacs, R. Kucera, I. Schildkraut and A. K. Aggarawal, *Nat. Struct. Biol.*, 2000, **7**, 134.
4. E. M. Southern, *J. Mol. Biol.*, 1975, **98**, 503.
5. D. A. Jackson, R. H. Symons and P. Berg, *Proc. Natl. Acad. Sci. USA*, 1972, **69**, 2904.
6. S. N. Cohen, A. C. Chang, H. W. Boyer and R. B. Helling, *Proc. Natl. Acad. Sci. USA*, 1973, **70**, 3240.

7. P. Balbas, X. Soberon, F. Bolivar and R. L. Rodriguez, *Biotechnology*, 1988, **10**, 5.
8. R. K. Saiki, S. Scarf, F. Faloona, K. B. Mullis, G. T. Horn, H. A. Erlich and N. Arnheim, *Science*, 1985, **230**, 1350.
9. A. M. Maxam and W. Gilbert, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 560.
10. F. Sanger, A. Nicklen and A. R. Coulson, *Proc. Natl. Acad. Sci. USA*, 1977, **74**, 5463.
11. F. Sanger, G. M. Air, B. G, Barrell, A. R. Coulson, J. C. Fiddes, C. A. Hutchinson III, P. M. Slocome and M. Smith, *Nature*, 1977, **265**, 687.
12. S. Anderson, A. T. Bankier, B. G. Barrell, M. H. de Bruijn, A. R. Coulson, J. Drouin, I. C. Eperon, D. P. Nierlich, B. A. Roe, F. Sanger, P. H. Schreier, A. J. Smith, R. Staden and I. G. Young, *Nature*, 1981, **290**, 457.
13. R. M. Lechan, P. Wu, I. M. Jackson, H. Wolf, S. Cooperman, G. Mandel and R. H Goodman, *Science*, 1986, **231**, 159.
14. R. D. Fleischmann, M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick *et al.*, *Science*, 1995, **269**, 496.
15. N. Watson, *Gene*, 1988, **70**, 399.
16. S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller and D. J. Lipman, *Nucleic Acids Res.*, 1997, **25**, 3389.
17. J. D. Thompson, D. G. Higgins and T. J. Gibson, *Nucleic Acids Res.*, 1994, **22**, 4673.
18. V. Healy, S. Doonan and T. V. McCarthy, *Biochem J.*, 1999, **339**, 713.
19. H. G. Khorana, *Science*, 1979, **203**, 614.
20. M. H. Caruthers, A. D. Barone, S. L. Beaucage, D. R. Dodds, E. F. F. Isher, L. J. McBride, M. Matteucci, Z. Stabinsky and J.-Y. Tang, *Methods Enzymol.*, 1987, **154**, 287.
21. A. J. Jeffreys, V. Wilson and L. S. Thein, *Nature*, 1985, **314**, 67.

Further Reading

S. B. Primrose, R. M. Tyman and R. W. Old, *Principles of Gene Manipulation*, 6th edn., Blackwell Science, Oxford, 2002.

T. A. Brown, *Gene Cloning*, 4th edn., Blackwell Science, Oxford, 2001.

M. Singer and P. Berg, *Genes and Genomes*, Palgrave Macmillan, Basingstoke, 1996.

C. Dennis and R. Gallagher (eds.), *The Human Genome*, Palgrave Macmillan, Basingstoke, 2001.

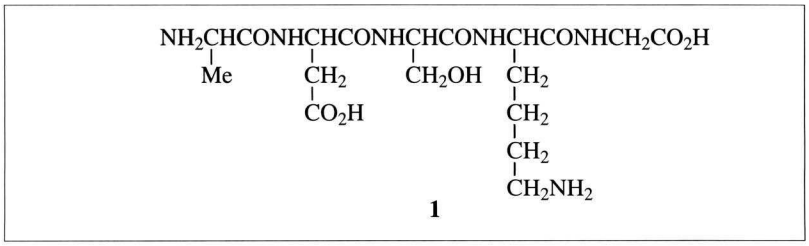
D. E. Crane and M. L. Raymer, *Fundamental Concepts of Bioinformatics*, Benjamin Cummings, San Francisco, 2003.

M. Krawczak and J. Schmidtke, *DNA Fingerprinting*, 2nd edn., BIOS Scientific, Oxford, 1998.

Answers to Problems

Chapter 1

1.1. The structure is shown in **1**.



1

1.2. Taking the residues three at a time and referring to Table 1.3 gives the protein sequence:

Glu-Leu-Val-Ile-Ser-Ile-Ser-Thr-His-Glu-Lys-Ile-Asn-Gly

Re-written in the single-letter code this is ELVISISTHEKING or ELVIS IS THE KING! This is a view that may not be shared (or perhaps understood) by people born in the second half of the 20th century! (There is no protein known whose sequence contains a complete phrase in English).

1.3. In the third reading frame the sequence splits up and translates as follows:

UU|CCA|CAG|UGU|UAU|AUC|CGG|CUG|GG

Pro-Gln-Cys-Tyr-Ile-Arg-Leu

1.4. The possible reading frames are CAG|CAG|CAG, AGC|AGC|AGC and GCA|GCA|GCA. These reading frames code for poly-Gln, poly-Ser and poly-Ala, respectively.

1.5. The three possible reading frames and their translation products are as follows:

AUA|GAU|AGA|UAG|AUA
Ile-Asp-Arg-STOP

A|UAG|AUA|GAU|AGA|UAG
STOP-Ile-Asp-Arg-STOP

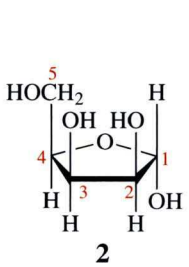
UAG|AUA|GAU|AGA|UAG
STOP-Ile-Asp-Arg-STOP

Every fourth triplet is a termination codon, so only a tripeptide can be made. It is experiments like this that established the nature of the termination codons.

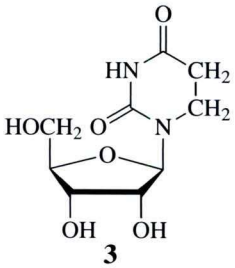
Chapter 2

2.1. The structure is shown in **2**. The OH groups on both C-2 and C-3 are reversed compared with ribose. The OH on C-1 is downwards because this is the α-anomer.

2.2. The structure is shown in **3**. C-5 and C66 of the pyrimidine ring are shown explicitly for clarity.

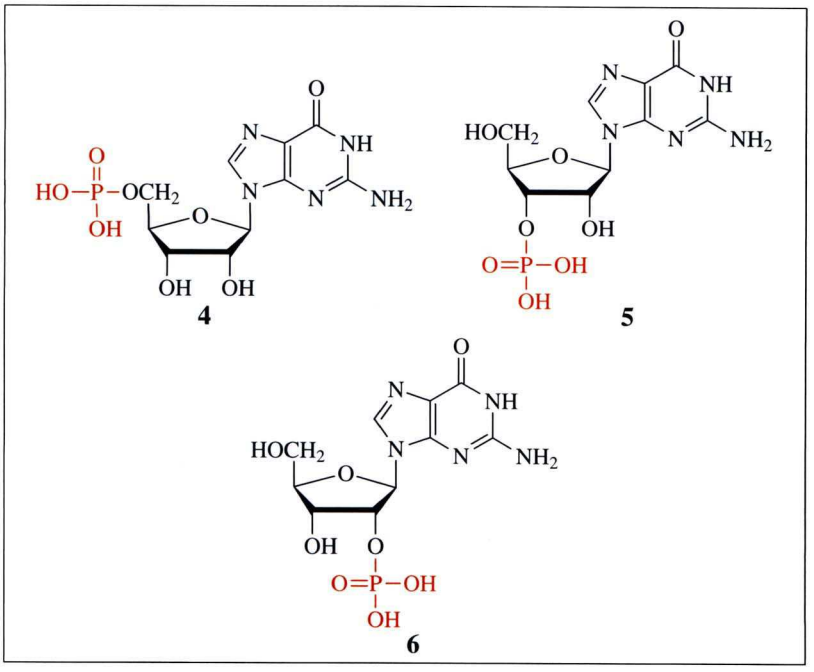


2

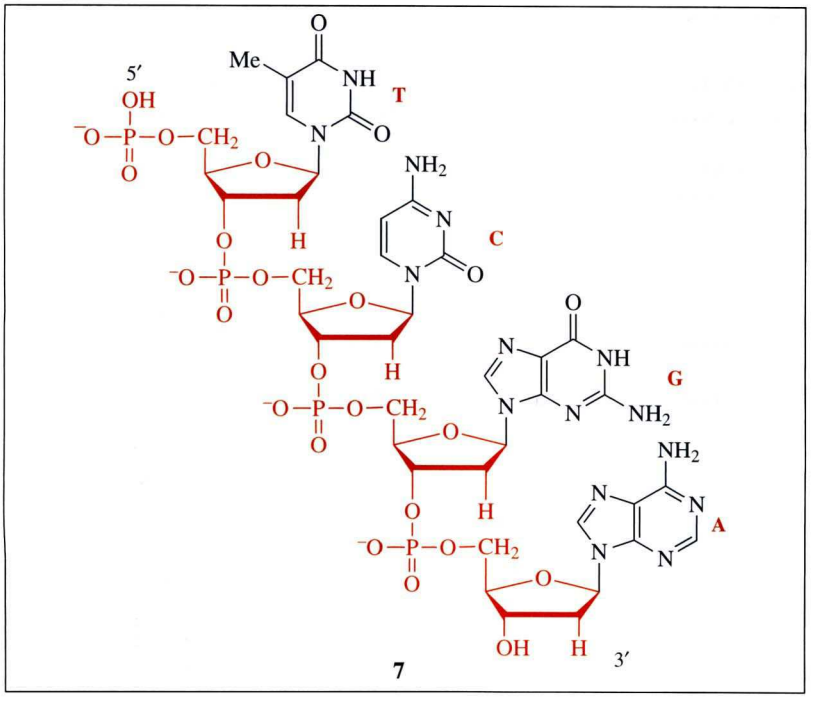


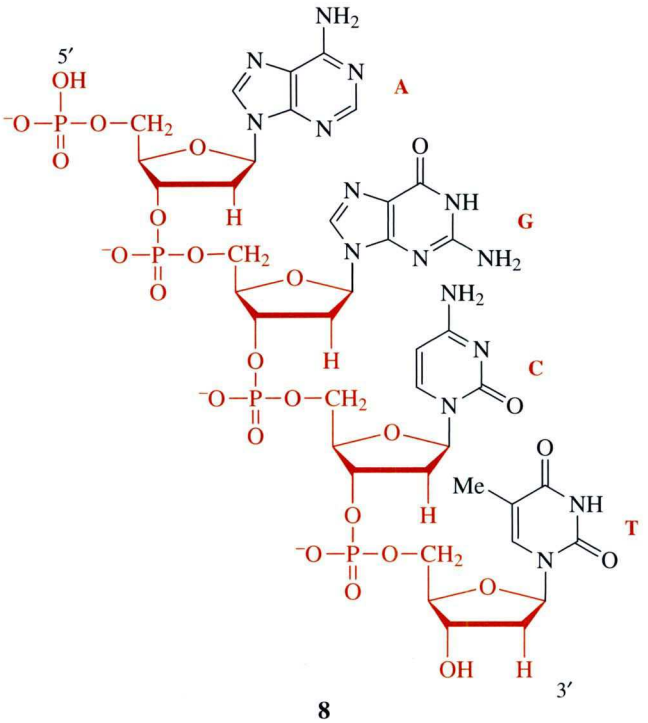
3

2.3. The structures are shown in **4–6**. **4** is guanosine 5′-monophosphate, **5** is guanosine 3′-monophosphate, and **6** is guanosine 2′-monophosphate.



2.4. These must be DNA molecules because they contain T. The structures are in **7** and **8**, respectively.





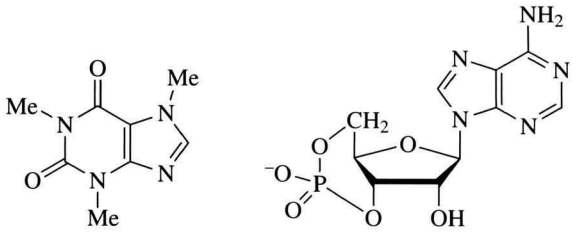
8

2.5. The products formed by the two enzymes are given in Table A1. The enzyme from pancreas gives the same products from both tetranucleotides. The enzyme from spleen produces different products and could be used to distinguish between them.

Table A1 Hydrolysis product of the tetranucleotides in Problem 2.4

	Products	
Enzyme	*pTCGA*	*pAGCT*
Pancreatic	5′-TMP, 5′-CMP, 5′-GMP, 5′-AMP	5′-AMP, 5′-GMP, 5′-CMP, 5′-TMP
Spleen	3′,5′-TDP, 3′-CMP, 3′-GMP, A	3′,5′-ADP, 3′-GMP, 3′-CMP, T

2.6. The structures are shown in **9** and **10**, respectively.



9 **10**

Chapter 3

3.1. The bases can be recognized by their patterns of substitution, and the sequence is ACTTAA. The chain direction is 3′→5′.

3.2. In Worked Problem 3.3 the value of K was obtained as 68 °C. So for a DNA molecule with $(G + C) = 72\%$ we have $72 = 2.44(T_m - 68)$, from which $T_m = 97\,°C$.

3.3. The diameter of the B-form double helix is 2 nm, so the radius is 1 nm and the volume of the molecule in question is $V_{DNA} = \pi \times 1^2 \times 6.53 \times 10^7\ nm^3$. The radius of the cell is 5 μm or 5×10^3 nm, so $V_{cell} = 4\pi \times (5 \times 10^3)^3/3\ nm^3$. Hence:

$$\frac{V_{DNA}}{V_{cell}} = \frac{3 \times 6.53 \times 10^7}{4 \times (5 \times 10^3)^3} \approx 2 \times 10^{-4}$$

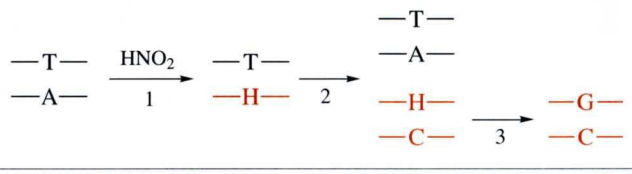
Or, put another way, the DNA from this chromosome occupies only 0.02% of the volume of the cell in spite of having a length 10,000 times greater than the radius of the cell. This emphasizes the degree of compaction achieved when DNA is packaged into chromosomes.

3.4. Approximately 150 bp of DNA is in contact with the histone core. The length of this fully extended would be $150 \times 0.34 = 51$ nm. The total width of the core particle is 10 nm (6 nm for the histone octamer and twice 2 nm for the DNA wrapped around it). So the degree of compaction is five-fold.

3.5. In the second generation the cells contained equal amounts of $^{15}N/^{14}N$ and $^{14}N/^{14}N$ molecules. The $^{15}N/^{14}N$ hybrid in the next generation will produce one $^{15}N/^{14}N$ molecule and one $^{14}N/^{14}N$. The $^{14}N/^{14}N$ molecule in the second generation will produce two $^{14}N/^{14}N$ copies. So in the third generation the ratio of $^{15}N/^{14}N$ to $^{14}N/^{14}N$ molecules will be 1:3.

3.6. The process is shown in Scheme A1. Starting with a piece of DNA which contains a normal A/T base pair, in step 1 the A is converted into H (the base in hypoxanthine). Replication in step 2 leads to one molecule with the normal A/T base pair, and a mutant with an H/C base pair. Subsequent replication (step 3) produces DNA with a G/C base pair.

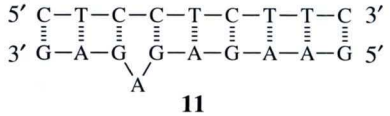
Scheme A1



3.7. The DNA polymerase copies at a rate of about 2000 bases per second and has an error frequency of about 1 in 10^{10}. So the time required to replicate the genome is $\frac{4.8\times10^6}{2\times10^3}$ s, which is 2400 s or 40 min. The chances of an error being made will be about 1 in $\frac{10^{10}}{5\times10^6}$, or 1 in 2000. So an error is likely once in 2000 replications.

Chapter 4

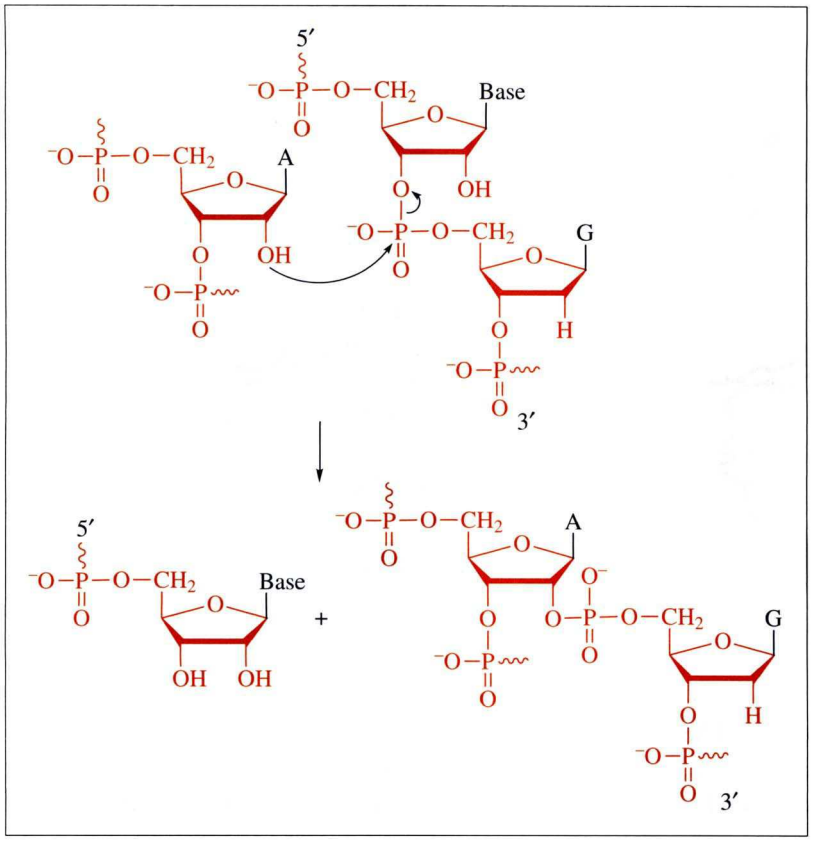
4.1. In the file under the entries "CMPND" the sequences are given as DNA (5′-D(*CP*TP*CP*CP*TP*CP*TP*TP*C)-3′) and RNA (5′-R(*GP*AP*AP*GP*AP*GP*AP*GP*AP*G)-3′). This simply means that the DNA is 5′-CTCCTCTTC-3′ and the RNA is 5′-GAAGAGAGAG-3′. These will form the duplex structure shown in **11**. There is an adenosine bulge in the RNA.



11

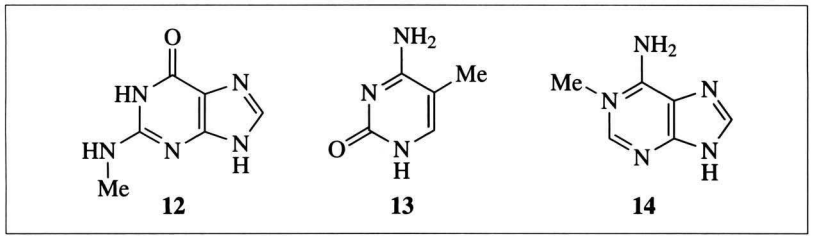
4.2. It is stated in Section 4.2 that the dystrophin gene is about 2.4 Mbp long (the length of the exons is insignificant compared with the introns). The time required to transcribe this gene would be $(2.4\times10^6)/50 = 4.8\times10^4$ s, or 800 min, or somewhat more than 13 h.

4.3. The reactions are shown in Scheme A2. Note in particular the structure of the adenosine derivative at the branch point.



Scheme A2

4.4. The structures are shown in **12–14**, respectively.



12 **13** **14**

4.5. Refer to Table 4.1 for the possible wobble bases in anticodon. The codons for cysteine are 5′-UGU-3′ and 5′-UGC-3′. Both U and C in the wobble position can be read by G so the anticodon sequence is 3′-ACG-5′. In the case of arginine, the three codons 5′-CGU-3′, 5′-CGC-3′ and 5′-CGA-3′ can all be read by 3′-GCI-5′. The codon 5′-CGG-3′ requires 3′-GCC-5′. The remaining two codons are

5′-AGA-3′ and 5′-AGG-3′, both of which can be read by 3′-UCU-5′. Hence three tRNA molecules are required to recognize the six codons.

4.6. The model is shown in Figure A1 (the orientation of the one that you produced may well be different). Note the close base-stacking between wybutosine and the preceding base.

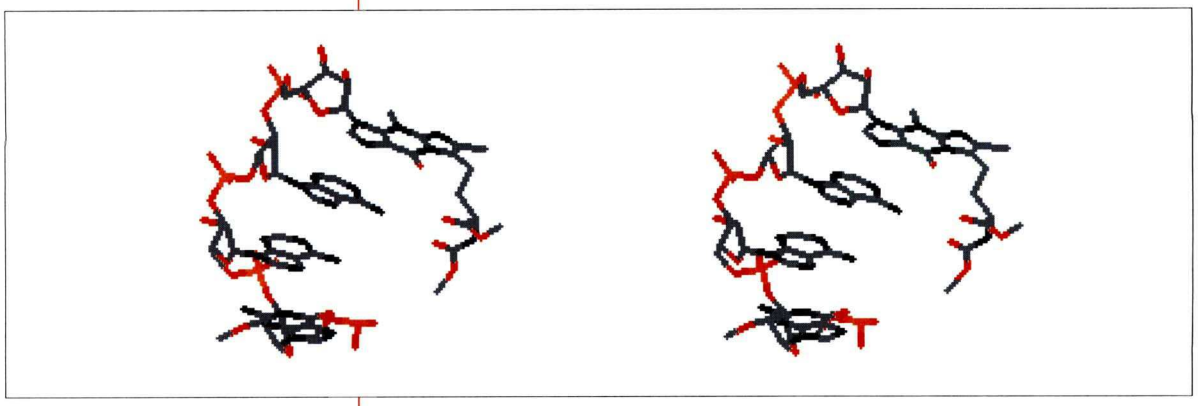
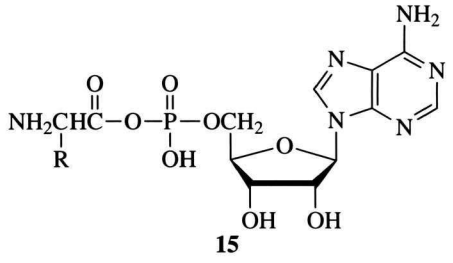


Figure A1 Stereo view of the anticodon in phenylalanyl-tRNA

4.7. The structure is shown in **15**. Not that it is a reactive mixed anhydride. Given that the residue to which the amino acid is attached is the 3′-end of the tRNA, then the base must be an adenine.



15

Chapter 5

5.1. This is a very favourable case because three of the amino acids are coded by single codons and the other two by two codons (see Table 1.3). As a result, a mixture of only four oligos is required. The easiest way to see this is to write the possible coding sequences as shown below:

AUG UA^{U}_{C} UGG CA^{U}_{C} AUG

The oligos required are then ATGTATTGGCATATG, ATGTATTGGCACATG, ATGTACTGGCATATG and ATGTACTGGCACATG. The Us in the RNA codons are, of course, replaced by Ts in the oligos.

Using equation (5.1), the T_m for the first oligo listed can calculated as 40 °C, and that for last as 44 °C. These values are low because of the preponderance of A and T in the sequences.

Glycine is specified by 4 codons, and so including it in the site would increase the number of oligos required to 16. However, by including only the first two bases of the glycine codons (GG) the probe length is increased by two without increasing the number of oligos required.

5.2. Although the recognition sites of *Bam*HI and *Bgl*II are different (GGATCC and AGATCT, respectively), they both cut after the first base and the central four bases are the same. Hence the sticky end produced in both cases has the sequence GATC.

5.3. The easiest way to do this is by logging on to the site http://www.fermentas.com/techinfo/ NucleicAcids/mappbr322.htm. Click on *Restriction sites of the pBR322 DNA*. The new page contains a list of all enzymes that cleave the molecule. The site for *Pvu*I is shown to be at position 3733. Note that this is the position of the first base in the recognition sequence (check it in Figure 5.22).

Alternatively, log on to http://web.umasmed.edu/bioapps/rsites.html. This site requires that the sequence to be searched is entered in the *Sequence Entry* box. You clearly will not want to type it in, so go to the NCBI, retrieve the pBR322 sequence, copy it, and paste it into the box. In the section *Restriction Enzyme Selection,* highlight *Pvu*I, check on *Explicit Pick from Above List,* and then *Submit Sequence*. The result will show that there is one site for *Pvu*I and show its position as 3737. In this case, the position quoted is that of the base after which the cut occurs.

5.4. The best way to see what is required is to write out the complementary sequences at the two ends of the molecule as follows:

5′-TTCTCATGTTTGACA---------TTTCGTCTTCAAGAA-3′

3′-AAGAGTACAAACTGT--------AAAGCAGAAGTTCTT-5′

Reference to Figure 5.11 shows that we require an oligo complementary to the 3′-end of the upper sequence, and this is simply the 5′-end of the lower sequence. One of the required 15mers is, therefore, 5′-TTCTTGAAGACGAAA-3′. Similarly, we require an oligo complementary to the 3′-end of the lower sequence and this will be 5′-TTCTCATGTTTGACA-3′.

5.5. It should be possible to read at least 100 bands. On the (larger) original of this figure, about 140 could be read with reasonable confidence. The sequence read, starting from the prominent band at the bottom of the C track, and reversing it to read 5′→3′ was:

5′-CGTCCCGGCCTTGTACACACCGCCCGTCACACCATGG
GAGTTTGTTGCACCAGAAGTAGGTAGTCTAACCTTAGG
GGGGACGCTTACACGGTGTGGCAGATGACTGGGGTGA
AGTCGTAACAAGGTAGCCGTAGGGAAC-3′

5.6. This is very difficult because the partial sequences are not all from the same chain, as shown by the fact that there are no overlaps from the end of one fragment to the beginning of another. It can be done by first writing down the fragment sequences and their complements:

5′ tgaagtggtggcctaacta**cggct**
3′ acttcaccaccggattgatgccga
5′ *gacagtatt*tggtatctgcgctctgc
3′ ctgtcataaaccatagacgcgagacg
5′ aatactgtccttctactgtagccg
3′ *ttatgacag*gaagatgaca**tcggc**

It can now be seen that the third sequence comes between the first and second, but that it was sequenced from the complementary chain. The complete sequence can now be assembled into:

tgaagtggtggcctaacta**cggct**acactagaag*gacagtatt*tggtatctgcgctctgc

with the overlaps in **bold** and *italics*.

(Note: this problem was intended as a demonstration of the fact that fragment assembly by eye can be very difficult with only three fragments, and gets to be virtually impossible with more than three. It is always done in practice by computer.)

5.7. Follow the procedure in Worked Problem 5.4 to obtain the reading frames of pBR322 as shown in Figure 5.23. Then follow the procedure in worked Problem 5.5, but clicking on the identified orf at the right of the fifth bar. Run the BLAST search. The output gives a very long list of proteins, all of which are β-lactamases (the enzyme responsible for ampicillin resistance); the protein from pBR322 has the accession code AAB59737.1.

The list of orfs in Figure 5.24 gives its position as 3293 to 4153. The codon starting at 3293 is UUA (in RNA terms), which is not an initiator codon. On the other hand, the complementary strand would have a codon 5′-TAA-3′ (or UAA in RNA terms). This is a termination codon. This suggests that the ampicillin resistance gene is read from the other strand. Consistently, the codon starting at 4153 on the complementary strand is 5′-ATG-3′ (AUG in RNA terms), which is the initiation codon. The gene is, therefore, coded by the complementary strand.

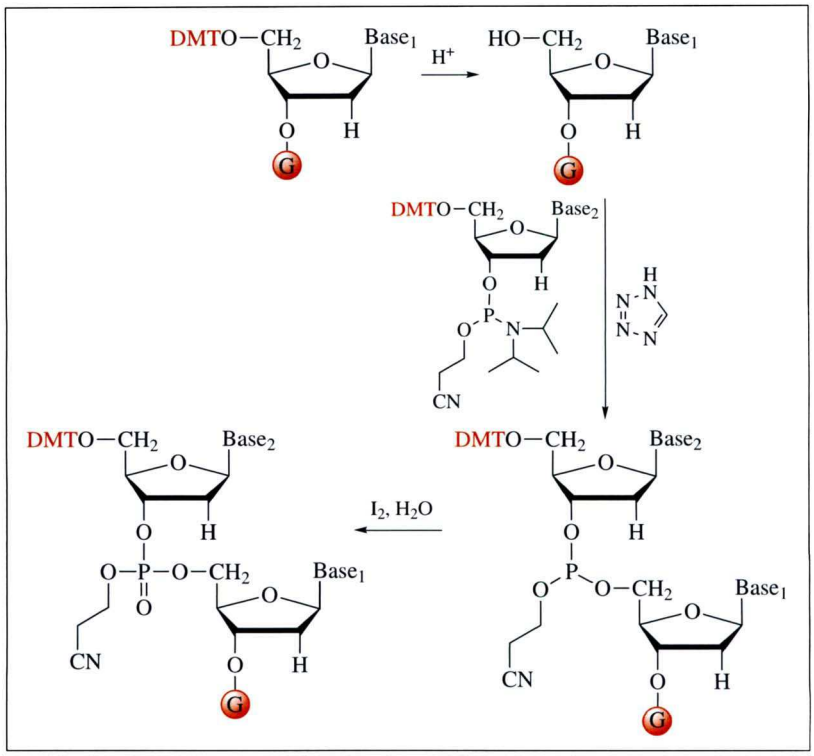
5.8. The sequence in Problem 5.5 was submitted for a BLAST search at the NCBI web site (following the route *Tools→BLAST→ Standard nucleotide-nucleotide BLAST*). A very large number of matches were obtained, all of them 16S RNA molecules. The alignment for the best match is shown in Figure A2.

```
gi 21389187Alkane-degrading soil bacterium MVAB Hex1

16S ribosomal RNA gene, complete sequence.

Query: 7    ggccttgtacacaccgcccgtcacaccatgggagtttgttgcaccagaagtaggtagtct 66
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct: 1378 ggccttgtacacaccgcccgtcacaccatgggagtttgttgcaccagaagtaggtagtct 1437

Query: 67   aaccttaggggggacgctta-cacggtgtggcagatgactggggtgaagtcgtaacaagg 125
             |||||||||||||||||| |||||||||| ||||||||||||||||||||||||||
Sbjct: 1438 aaccttagggggggacgcttaccacggtgtggccgatgactggggtgaagtcgtaacaagg 1497

Query: 126  tagccgtaggg 136
             ||||||||||
Sbjct: 1498 tagccgtaggg 1508
```

Figure A2 Identification of the sequence from Problem 5.5 as part of a 16S RNA gene

The unknown (query) sequence aligns with the 3′-end of the 16S RNA gene as expected. The gap at position 87 may have arisen from a misreading of the gel, as might the difference in bases at position 99.

5.9. The reaction sequence is shown in Scheme A3.



Scheme A3

Subject Index

Nucleic Acids describes the way in which the fundamentally important biological activities of these molecules can be understood in terms of their chemical structures. The book focuses on the chemistry of the deoxyribonucleic acids (DNA) and ribonucleic acids (RNA). However, because nucleic acid chemistry cannot be fully understood without some knowledge of the underlying biology, a substantial amount of the background biology is also included.

Beginning with a concise review of the biological roles of nucleic acids, the text then discusses the components from which they are made, and works up through nucleosides and nucleotides to the covalent structures of the nucleic acids themselves. The double helical structure of DNA and its implications for replication are then described.

This is followed by a detailed treatment of the chemistry of the processes by which the information encoded in DNA is expressed in terms of the amino acid sequences of proteins. The final chapter describes modern tools of DNA analysis and how they have been used in a range of recent applications such as gene cloning, genome sequence analysis, and DNA fingerprinting.

Although targeted specifically at undergraduate chemistry students, Nucleic Acids will also be of interest to undergraduates studying biochemistry.

ISBN 0-85404-481-7

9 780854 044818 >