# 1 DNA: Structure and function

## *Key concepts*

- The genetic information is contained within nucleic acids
- DNA is a double-stranded antiparallel helix
- Base pairing (A to T and G to C) holds the two strands of the helix together
- DNA replication occurs through the unwinding of the DNA strands and copying each strand
- The central dogma of molecular biology:
  - DNA makes RNA makes protein
- Transcription is the production of an RNA copy of one of the DNA strands
- Translation is decoding of an RNA molecule to produce protein

COPYRIGHTED MATERIAL

Every organism possesses the information required to construct and maintain a living copy of itself. The basic concepts of heredity and, as a consequence, genes can be traced back to 1865 and the studies of Gregor Mendel – discussed by Orel (1995). From the results of his breeding experiments with peas, Mendel concluded that each pea plant possessed two **alleles** for each gene, but only displayed a single **phenotype**. Perhaps the most remarkable achievement of Mendel was his ability to correctly identify a complex phenomenon with no knowledge of the molecular processes involved in the formation of that phenomenon. Hereditary transmission through sperm and egg became known about the same time and Ernst Haeckel, noting that sperm consists largely of nuclear material, postulated that the nucleus was responsible for heredity.

*Analysis of Genes and Genomes* Richard J. Reece
© 2004 John Wiley & Sons, Ltd
ISBNs: 0-470-84379-9 (HB); 0-470-84380-2 (PB)

## 1.1 Nucleic Acid is the Material of Heredity

The idea that genetic material is physically transmitted from parent to offspring has been accepted for as long as the concept of inheritance has existed. Both proteins and nucleic acid were considered as likely candidates for the role of the genetic material. Until the 1940s, however, many scientists favoured proteins. There were two main reasons for this. Firstly, proteins are abundant in cells; although the amount of an individual protein varies considerably from one cell type to another, the overall protein content of most cells accounts for over 50% of the dry weight. Secondly, nucleic acids appeared to be too simple to convey the complex information presumed to be required to convey the characteristics of heredity. **DNA** (deoxyribonucleic acid) was first isolated in 1869 by the Swiss chemist Johann Frederick Miescher. He separated nuclei from the cytoplasm of cells, and then isolated an acidic substance from these nuclei that he called nuclein. Miescher showed that nuclein contained large amounts of phosphorus and no sulphur, characteristics that differentiated it from proteins. In what proved to be a remarkable insight, he suggested that 'if one wants to assume that a single substance... is the specific cause of fertilization then one should undoubtedly first of all think of nuclein'.

In 1926, based on the idea that DNA contained approximately equal amounts of four different groups, called nucleotides, and by determining the type of linkage that joined the nucleotides together, Levene and Simms proposed a tetranucleotide structure (Figure 1.1) to explain the chemical arrangement of nucleotides within nucleic acids (Levene and Simms, 1926). They proposed a very simple four-nucleotide unit that was repeated many times to form long nucleic acid molecules. Because the tetranucleotide structure was relatively simple, it was widely believed that nucleic acids could not provide the chemical variation expected of the genetic material. Proteins, on the

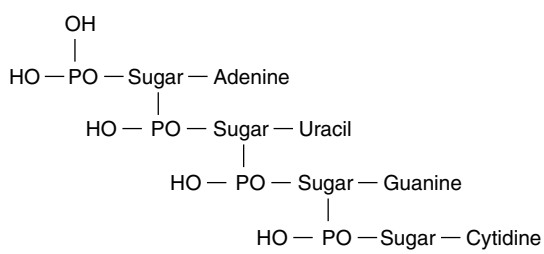



**Figure 1.1.** The tetranucleotide model for nucleic acid structure proposed by Levene and Simms in 1926. At the time that this model was proposed, it was thought that plant and animal nucleic acid might be different, and the differences between DNA and RNA were not fully understood

other hand, containing 20 different amino acids, could provide the basis for substantial variation.

In 1928, Frederick Griffith performed experiments using several different strains of the bacterium *Streptococcus pneumoniae* (Griffith, 1928). Some of the strains used were termed virulent, meaning that they caused pneumonia in both humans and mice. Other strains were avirulent, and did not cause illness. Virulent and avirulent strains are morphologically distinct in that the virulent strains have a polysaccharide capsule surrounding the bacterium and form smooth, shiny-surfaced colonies when grown on agar plates. Avirulent bacteria lack the capsule and produce rough colonies on the same plates. The smooth bacteria are virulent because the polysaccharide capsule means that they are not easily engulfed by the immune system of an infected animal, and thus are able to multiply and cause pneumonia. The rough bacteria that lack the polysaccharide capsule do not have this protection and are consequently readily engulfed and destroyed by the host immune system.

Griffith knew that only living virulent bacteria would produce pneumonia when injected into mice. If heat-killed virulent bacteria were injected into mice, no pneumonia would result, just as living avirulent bacteria failed to produce the disease when similarly injected. Griffith's critical experiment (Figure 1.2) involved the injection into mice of living rough bacteria (avirulent) combined with heat-treated smooth bacteria. Neither cell type caused death in mice when they were injected alone, but all mice receiving the combined injections died. The analysis of blood of the dead mice revealed a large number of living smooth bacteria when grown on agar plates. Griffith concluded that the heat-killed smooth bacteria were somehow responsible for converting the live avirulent rough bacteria into virulent smooth ones. He called the phenomenon **transformation**, and suggested that the **transforming principle** might be some part of the polysaccharide capsule or some compound required for capsule synthesis, although he noted that the capsule alone did not cause pneumonia.

In 1944, Oswald Avery, Colin MacLeod and Maclyn McCarty published their work to show that the molecule responsible for the transforming principle was DNA (Avery, MacLeod and McCarty, 1944). They began by culturing large quantities of smooth *Streptococcus pneumoniae* cells. The cells were harvested from cultures and then heat-killed. Following homogenization and several extractions with detergent, they obtained an extract that, when tested by co-injection with live rough bacteria, still contained the transforming principle. Protein was removed from the extract by several chloroform extractions and polysaccharides were enzymatically digested and removed. Finally, precipitation of the resultant fraction with ethanol yielded a fibrous mass that still retained the ability to induce transformation of the rough avirulent cells. From the

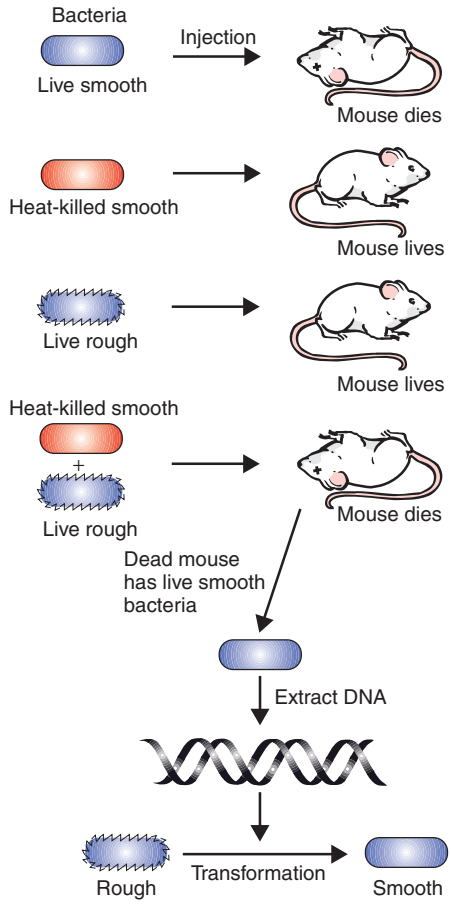



**Figure 1.2.** The key features of Griffith's experiments, combined with data of Avery, MacLeod and McCarty to identify DNA as the transforming principle from *Streptococcus pneumoniae*. Griffith noted that injecting live smooth bacteria into mice led to the formation of pneumonia and eventually to the death of the mouse. Heat-treating the bacteria before injection did not result in the formation of the disease. Non-virulent bacterial strains, which did not cause the disease on their own, could be transformed by co-injection with heat-treated virulent bacteria. Avery, MacLeod and McCarty identified DNA as the transforming principle

original 75 L culture of bacterial cells, the procedure yielded 10–25 mg of the 'active factor'. Further testing established beyond a reasonable doubt that the transforming principle was DNA. The fibrous mass was analysed for its nitrogen/phosphorus ratio, which was shown to coincide with the ratio expected for DNA. In order to eliminate all probable contaminants from their final extract, they treated it with the proteolytic enzymes trypsin and chyomtrypsin, and then digested it with an RNA digesting enzyme called

ribonuclease. Such treatments destroyed any remaining activity of proteins and RNA, but still retained the transforming activity. The final confirmation that DNA was transforming principle came by digesting the extract with deoxyribonuclease, which destroyed the transforming activity.

The second major piece of evidence supporting DNA as the genetic material was provided by the study of the infection of the bacterium *Escherichia coli* by one of its viruses, bacteriophage T2. Often simply referred to as **phage**, the virus consists of a protein coat surrounding a core of DNA (see Figure 1.3). In the early 1950s, little was known about the early steps of phage infection. The phage was known to be adsorbed to the surface of the bacteria, after which there was a latent period of approximately ten minutes before infectious virus particles started to be made, ultimately leading to host cell lysis and phage release. Alfred Hershey and Martha Chase reasoned that if they knew the fate of the phage protein and the nucleic acid at the beginning of the infection process, they would understand more about the nature of those early steps.

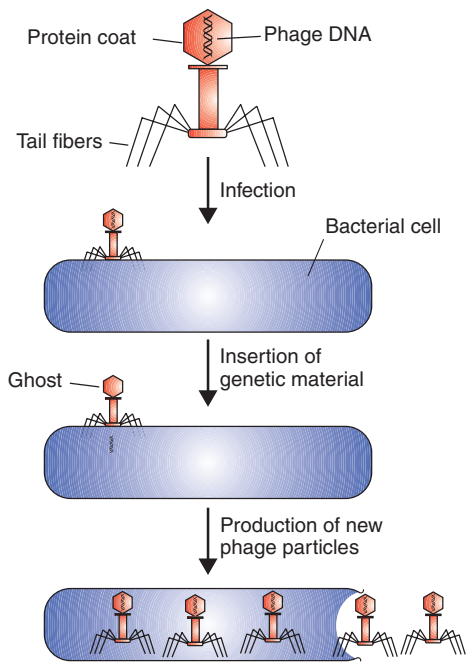



**Figure 1.3.** The life cycle of T2 phage. During the course of infection, the bacteriophage adheres to the surface of the *Escherichia coli* cell. The genetic information, but not the whole phage particle, is inserted into the bacterium, where it is replicated. The phage 'ghost', which lacks the genetic material, remains at the bacterial surface. Once the newly synthesized phage particles are produced, bacterial cell lysis occurs and the phage particles are released into the surrounding medium

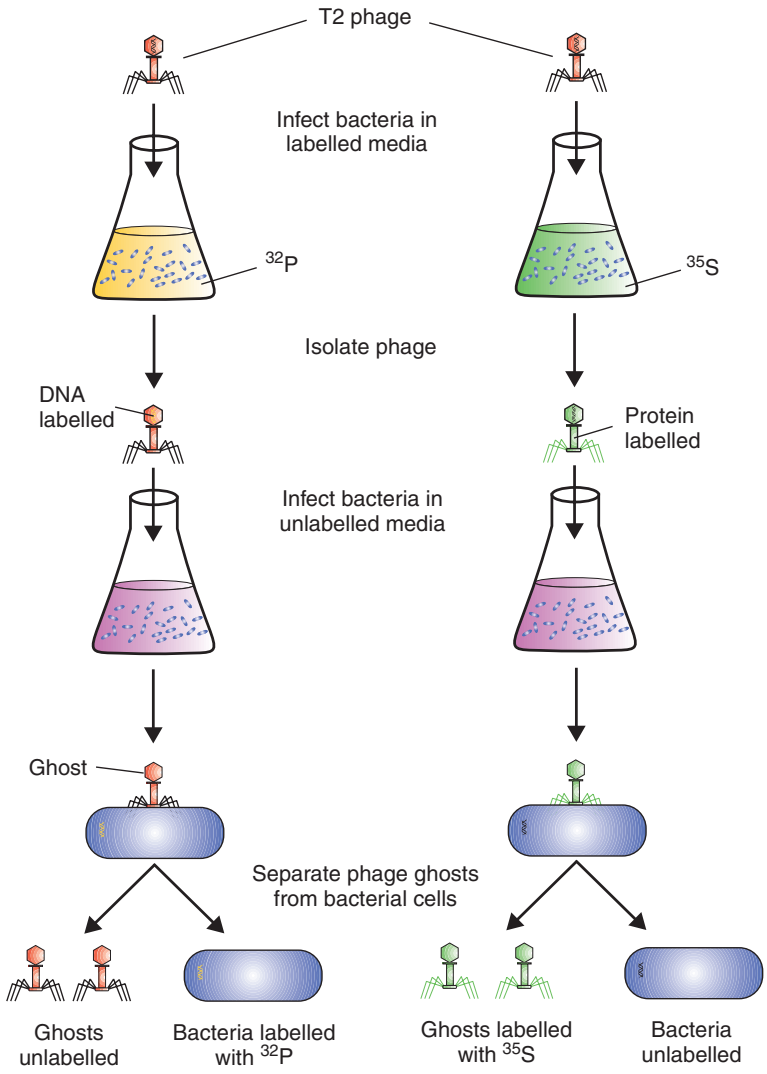



**Figure 1.4.** The Hershey–Chase blender experiment to show that nucleic acid was the genetic material. Hershey and Chase grew T2 bacteriophages on bacteria whose media contained either $^{32}$P (to label the phosphorus of nucleic acid) or $^{35}$S (to label the sulphur of proteins – the side chains of the amino acids methionine and cysteine both contain sulphur). They used their radio-labelled bacteriophages to infect a new culture of unlabelled bacteria. After a brief incubation, the bacteria were harvested by centrifugation and put into a blender to shear the bacteria away from the phage particles attached to their surface. They found that, when the DNA was labelled, the label was transferred to the bacterial cell, while the labelled protein remained with the phage ghosts. They concluded, therefore, that the material of heredity – i.e. the material passed on to make new offspring – was nucleic acid

They used the radioisotopes $^{32}P$ and $^{35}S$ to follow the molecular components of the phages during infection (Figure 1.4). Because DNA contains phosphorus but not sulphur, $^{32}P$ effectively labels DNA, and because proteins contain sulphur but not phosphorus, $^{35}S$ labels protein. If *E. coli* cells are grown in the presence of $^{32}P$ or $^{35}S$ and then infected with T2 virus, the newly synthesized phages will have either a radioactively labelled DNA core or a radioactively labelled protein coat, respectively. These labelled phages can be isolated from the medium of infected cultures and used to infect other unlabelled bacteria. Hershey and Chase labelled the T2 phages with either $^{35}S$ or $^{32}P$, and allowed them to adsorb to unlabelled bacteria. The cells were then separated from the unadsorbed material by centrifugation. The cells were resuspended and the suspension was blended to separate the phage 'ghosts' from the infected bacteria. The ghosts should not contain the genetic material, which needs to be replicated inside the host bacteria. Hershey and Chase found that ~80% of the $^{35}S$ label was removed from the bacteria whereas only ~20% of the $^{32}P$ label was removed. They concluded that 'most of the phage DNA enters the cell, and a residue containing at least 80% of the sulphur-containing protein of the phage remains at the cell surface'. This work, together with that of Avery, MacLeod and McCarty, provided overwhelming evidence that DNA was the molecule responsible for heredity. Curiously, however, Hershey and Chase seemed somewhat skeptical about their findings, concluding their paper by saying that the '... protein has no function in phage multiplication, and that DNA has some function' (Hershey and Chase, 1952).

## 1.2 Structure of Nucleic Acids

As we have seen in the above section, the material of heredity is DNA, or more correctly nucleic acids. In most organisms, the hereditary material is DNA. However, a number of viruses use RNA (ribonucleic acid) as the building block for their genome. DNA and RNA are polymeric molecules made up of linear chains of subunits called nucleotides. Each nucleotide has three parts: a nitrogenous base, a five-carbon-atom sugar and a phosphate group (Figure 1.5). The combination of base and sugar is termed a **nucleoside**, while the base–sugar–phosphate is called a **nucleotide**. Since they contain the sugar 2′-deoxyribose, the nucleotides of DNA are termed deoxyribonucleotides, while those of RNA, which contain the sugar ribose, are known as ribonucleotides. The nucleotide bases can be either a double-ringed **purine** or a single-ringed **pyrimidine**. DNA and RNA are both built up from two purine containing nucleotides and two pyrimidine containing nucleotides. The purines of both DNA and RNA are the same – **adenine** (A) and **guanine** (G). The pyrimidine

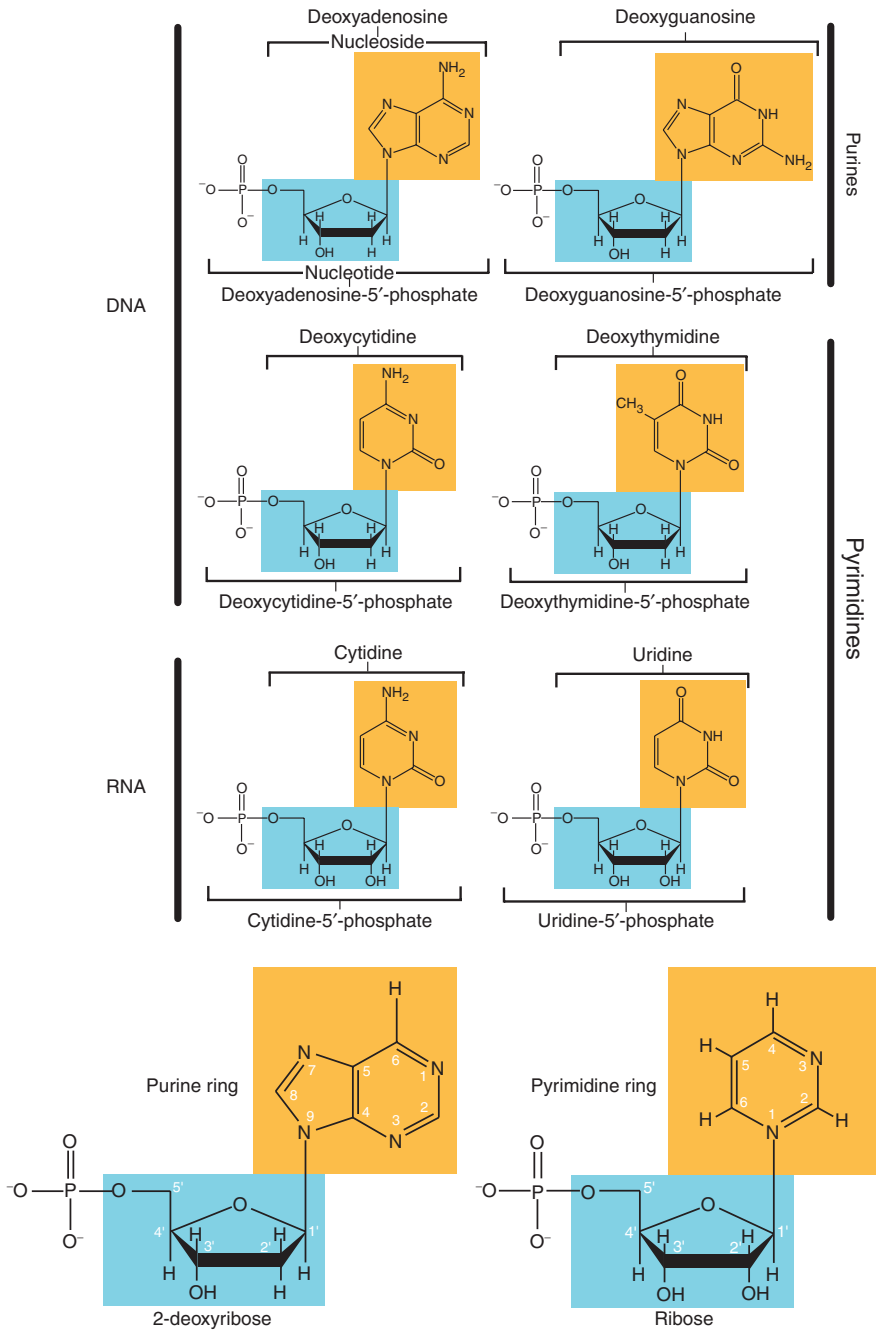



**Figure 1.5.** The structures of the purines and pyrimidines found in nucleic acids. The nitrogenous bases are highlighted in orange and the sugar groups are highlighted in blue. Beneath is the numbering system used throughout this text. The atoms of the purine ring are numbered from 1 to 9, and those of the pyrimidine ring are numbered from 1 to 6. The atoms of the sugar are numbered from 1′ to 5′

**cytosine** (C) is also found in both nucleic acids, while the pyrimidine **thymine** (T) is limited to DNA, being replaced by **uracil** (U) in RNA.

The numbering system for nucleotides that is used extensively through this text is shown in Figure 1.5. Each of the carbon and nitrogen atoms in both the pyrimidine and purine rings is numbered from 1 to 6, or 1 to 9, respectively. The carbon atoms of the sugar ring – either ribose or deoxyribose – are numbered from 1′ to 5′ (spoken as 1-prime to 5-prime). Thus, 2′-deoxyribose lacks a hydroxyl group attached to the 2′ carbon of the sugar ring. Individual nucleotides are connected to each other in both DNA and RNA through sugar–phosphate bonds that connect the hydroxyl group on the 3′ carbon of one nucleotide with the phosphate group on the 5′ carbon of another nucleotide. See Figure 1.6. Two nucleotides connected to each other are called a **dinucleotide**, three are called a **trinucleotide** and numerous nucleotides connected in a long chain is termed a **polynucleotide**.

In the early 1950s, the chemist Erwin Chargaff was performing experiments to address the chemical composition of nucleic acids, and he realized that nucleic acids did not contain equal proportions of each nucleotide. Chargaff isolated DNA from a number of organisms, both prokaryotic and eukaryotic (Chargaff, Lipshitz and Green, 1952; Chargaff *et al*., 1951; Zamenhof, Brawerman and Chargaff, 1952). He hydrolysed the DNA into its constituent nucleotides by treatment with strong acid, and then separated the nucleotides by paper chromatography. His experiments showed that the relative ratios of the four bases were not equal, but were also not random. The number of adenine (A) residues in all DNA samples was equal to the number of thymine (T) residues, while the number of guanine (G) residues equalled the number of cytosine (C) residues (Table 1.1). Chargaff's rules state that for any given species

- A = T and G = C
- sum of the purines = sum of the pyrimidines
- the percentage of (C + G) does not necessarily equal the percentage of (A + T).

These findings opened the possibility that it was the precise arrangements of nucleotides within a DNA molecule that conferred its genetic specificity, but the fundamental significance of the A = T and G = C relationships was not full realized until the three-dimensional structure of DNA was solved. As we will see later, in DNA A always pairs with T and G always pairs with C.

Between 1940 and 1953, many scientists were interested in solving the structure of DNA. X-ray diffraction as a method of determining protein structure was becoming an established technique. X-ray diffraction involves

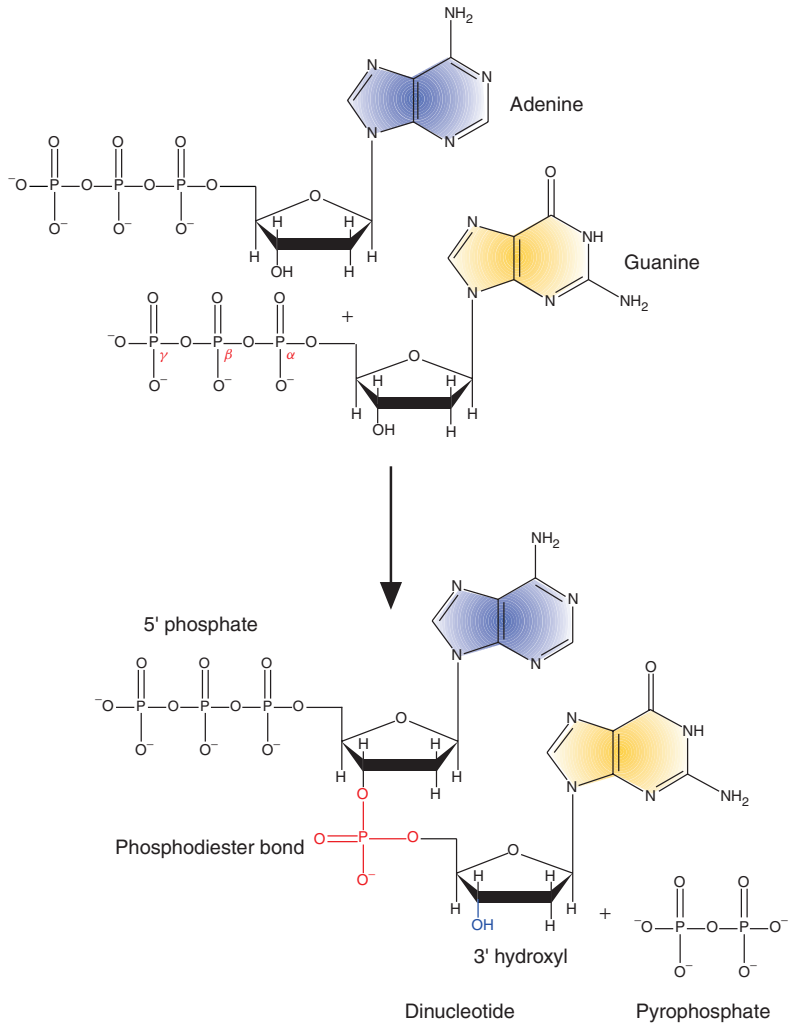



**Figure 1.6.** The joining of nucleotides. The joining of an adenine and a guanine nucleotide. The phosphates on the sugar ring of guanine are designated as $\alpha$, $\beta$ or $\gamma$. In the formation of the dinucleotide, pyrophosphate (representing the $\beta$ and $\gamma$ phosphates) is lost and the phosphodiester bond links the 3′ hydroxyl to the phosphate on the 5′ carbon atom of the sugar. DNA molecules invariably have a free 5′ phosphate and 3′ hydroxyl

firing a beam of X-rays at a regular array of molecules – either a crystal or a fibre. When the X-rays hit an atom in the array they will be diffracted, and the diffracted beams are detected as spots on X-ray film. Analysis of the diffraction patterns yields information about the structure and shape of the molecules in the array. As early as 1938 William Astbury applied the technique to fibres of DNA. By 1947, he had detected a periodicity (or repeating unit) within DNA of

**Table 1.1.** Chargaff's rules. The ratios of individual nucleotides isolated from DNA of various sources. While the ratios of purine:purine and pyrimidine:pyrimidine vary widely, the ratio of purine:pyrimidine was found to be a constant unity

| Organism | A to G | T to C | A to T | G to C | Purines: pyrimidines |
|---|---|---|---|---|---|
| Ox | 1.29 | 1.43 | 1.04 | 1 | 1.1 |
| Human | 1.59 | 1.75 | 1 | 1 | 1 |
| Hen | 1.45 | 1.29 | 1.06 | 0.91 | 0.99 |
| Salmon | 1.43 | 1.43 | 1.02 | 1.02 | 1.02 |
| Sea urchin | 1.83 | 1.80 | 1.02 | 1.00 | 1.01 |
| Wheat | 1.22 | 1.18 | 1 | 0.97 | 0.99 |
| Yeast | 1.67 | 1.92 | 1.03 | 1.2 | 1 |
| *Hemophilus influenzae* | 1.75 | 1.54 | 1.06 | 0.93 | 1.01 |
| *Escherichia coli* | 1.05 | 0.95 | 1.09 | 0.99 | 1 |
| *Serratia marcescens* | 0.76 | 0.63 | 1.03 | 0.85 | 0.92 |
| *Bacillus schatz* | 0.68 | 0.58 | 1.07 | 0.9 | 0.96 |

0.34 nm. Between 1950 and 1953, Rosalind Franklin obtained improved X-ray data from highly purified samples of DNA. Her work confirmed the 0.34 nm periodicity, and suggested that the structure of DNA was some sort of helix. Franklin, however, did not propose a model for the structure of DNA. Rather, Linus Pauling and Robert Corey used Franklin's data, together with that of others, to propose that DNA was a triple helix with the phosphates near the centre of the axis and the bases on the outside (Pauling and Corey, 1953).

## 1.3 The Double Helix

Franklin noted that DNA fibres could give two distinct types of diffraction pattern depending upon how the samples were prepared and stored. The first (termed Structure A) was composed of fibres that were relatively dehydrated, while the second (Structure B) was prevalent over a wide variety of conditions. She noted that the change from Structure A to Structure B was reversible, depending on the levels of sample hydration (Franklin and Gosling, 1953). It is thought that the B-form of DNA is the biologically significant conformation. Other forms of DNA (the right-handed A form and the left-handed Z form) certainly do exist under certain conditions, and may play significant roles in certain cellular processes. For example, a family of proteins that bind specifically to Z-DNA has recently been described (Schwartz *et al.*, 2001). Here, however, we will concentrate on the properties and interactions of B-form DNA.

In 1953, James Watson and Francis Crick attempted to build molecular models of DNA and realized that the Pauling–Corey structure was incorrect, with some atoms having to be closer together than was possible. By combining Franklin's X-ray diffraction patterns with Chargaff's rules, Watson and Crick proposed the, now famous, double-helix model in 1953 (Watson and Crick, 1953a). This model, shown in Figure 1.7, has the following major features, some of which have been updated slightly from the original model in the light of high-resolution crystal X-ray diffraction data.

(a) Two long polynucleotide chains coiled around a central axis, forming a right-handed double helix – this means that the turns are clockwise when looking down the helical axis.

(b) The two chains are antiparallel; that is, each chain has a specific orientation, and these run in opposite directions.

(c) The bases of both chains are flat structures, lying perpendicular to the axis. They are 'stacked' on one another, 0.34 nm apart, and are located on the inside of the helix.

(d) The nitrogenous bases of opposite strands are paired to one another by hydrogen bonds.

(e) Each complete turn of the helix is 3.4 nm long. This means that just over ten bases from each strand (10.4 bp) form one complete turn of the helix.

(f) Along the molecule, alternating larger major grooves and smaller minor grooves are apparent.

(g) The double helix measures approximately 2 nm in diameter.

The pairing of the nitrogenous bases in the centre of the helix is the most significant feature of the model by Watson and Crick. However, several other features are also important to understand the double helix.

### 1.3.1 The Antiparallel Helix

The antiparallel nature of the two polynucleotide chains is a key part of the double helix. Given the constraints of the bond angles of the bases and sugar phosphates, the double helix could not be constructed easily if both chains ran parallel to each another. One chain of the helix runs in the 5′ to 3′ orientation, and the other chain runs in the 3′ to 5′ orientation. This is illustrated in Figure 1.8. The 5′ and 3′ nomenclature is derived from the numbering system of the sugar ring that we saw in Figure 1.5. By convention, DNA sequences are

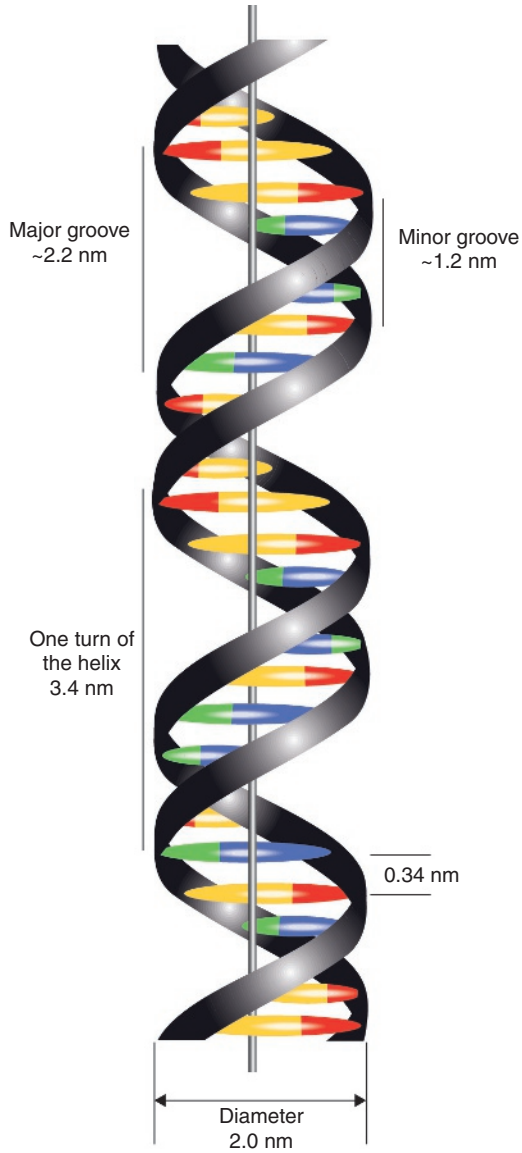



**Figure 1.7.** The Watson and Crick model of DNA

written in the 5′ to 3′ direction. This means that a single DNA chain begins with a free phosphate group on the 5′ carbon of a deoxyribose ring. Additional nucleotides are joined to the chain through phosphodiester bonds, which link the hydroxyl group on the 3′ carbon atom of one sugar with the phosphate on the 5′ carbon atom of an adjoining sugar. The chain terminates in a free hydroxyl group on the 3′ carbon atom of the last sugar.

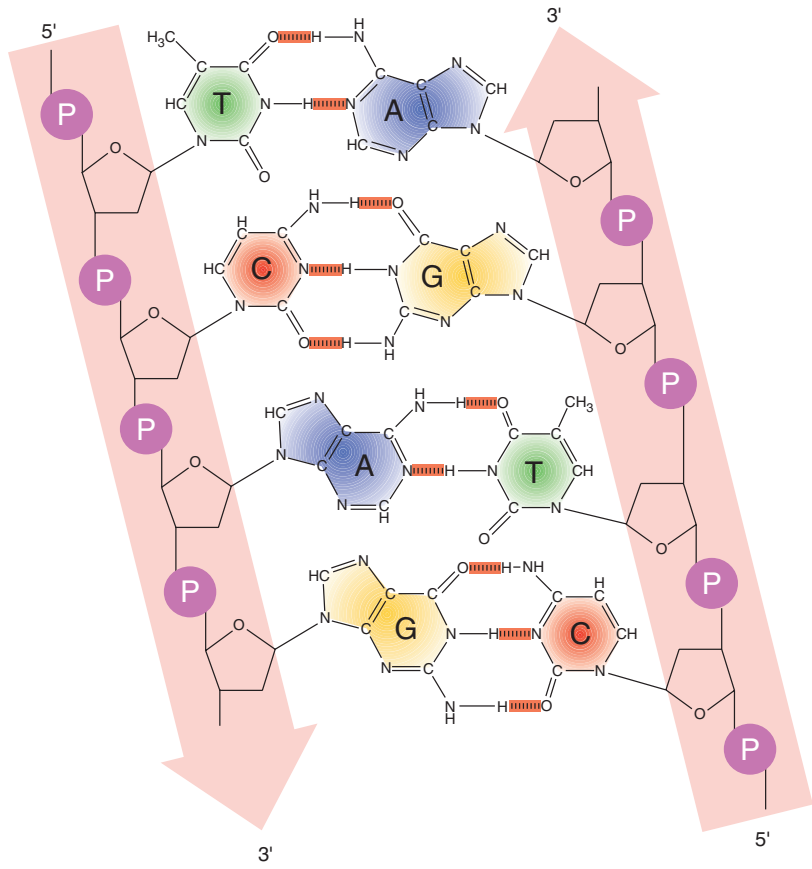



**Figure 1.8.** DNA base pairing and complementation. The two chains of the helix, arrowed in the 5′ to 3′ direction, are antiparallel. The bases on one strand of the helix are complementary to those on the opposite strand, A always base pairs with T and G always base pairs with C

## 1.3.2 *Base Pairs and Stacking*

The bases of both DNA chains are flat structures that lie approximately perpendicular to the helical axis. The bases themselves are stacked upon each other. The arrangement is best illustrated by inspection of a computer-generated model of high-resolution crystal X-ray diffraction data (Figure 1.9). It can be noted that the base pairs are not all perpendicular to the helical axis, and that some show **propeller twist**, where the purine and pyrimidine pair do not lie flat but are twisted with respect to each other, like the blades of a propeller (Dickerson, 1983). The pairing of a purine (A or G) with a pyrimidine (T or C) within the helix is important for the integrity of the helix.

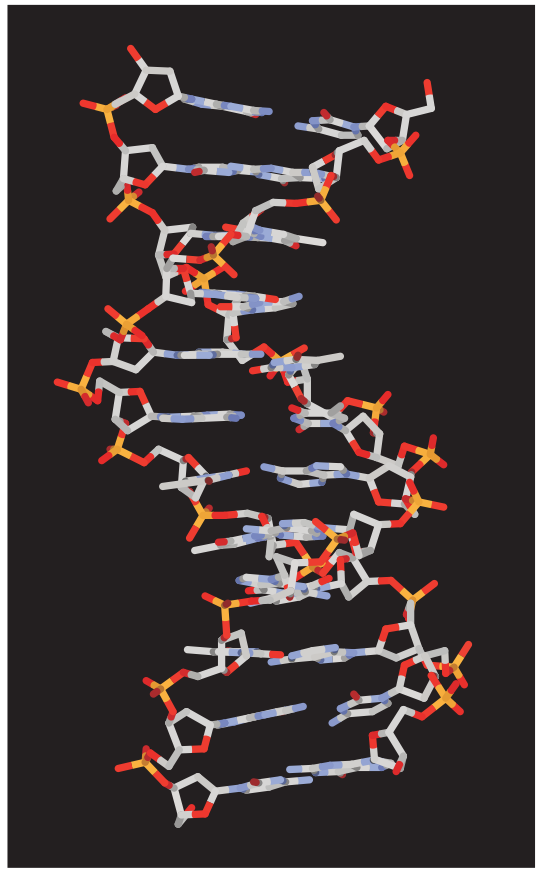


**Figure 1.9.** Computer generated model of DNA. The structure of double-stranded B-form DNA as derived from high-resolution X-ray diffraction of DNA crystals. Oxygen atoms are coloured red, phosphorus is orange, carbon is white and nitrogen is blue

The constant length of the purine–pyrimidine pairing would be disrupted if purine–purine (too large) or pyrimidine–pyrimidine (too small) pairings occurred. The purine–pyrimidine pairs are said to complement each other, and the two strands of a single DNA molecule are thus **complementary** to one another. Thus, if the sequence 5′-ATGATCAGTACG-3′ occurs on one strand of the DNA, the other strand must have the sequence 5′-CGTACTGATCAT-3′. These two sequences are complementary to each other:

```
Strand one: 5′-ATGATCAGTACG-3′
            ||||||||||||
Strand two: 3′-TACTAGTCATGC-5′
```

As we will we see in many of the subsequent chapters, the ideas of complementation between two strands of DNA form the basis of many genetic engineering experiments. The pairing of two DNA strands is very specific. Precise matches between two DNA strands, like those shown above, are highly stable and readily form helices. As we will see later, two DNA strands that are not precisely complementary to one another, but where there is still a high degree of complementation, retain the ability to interact with each other.

### 1.3.3 Gaining Access to Information with the Double Helix without Breaking it Apart

How can the information held within the sequence of DNA be read without having to unravel the double helix? The invariant nature of the sugar–phosphate backbone would seem to provide an almost impenetrable barrier to 'reading' the DNA base sequence. The grooves along the helical axis do, however, provide a mechanism whereby the bases can be distinguished from one another. As we can see in Figure 1.7, DNA is composed of alternating major and minor grooves along its axis. This is a result of the glycosidic bonds that attach a base pair to its sugar rings not lying directly opposite each other across the helical axis. As a result, the two sugar–phosphate backbones of the double helix are not equally spaced along the helical axis, and the grooves that form between the backbones are not of equal size. The major groove is wide (~0.22 nm) and shallow, while the minor groove is narrow (~0.12 nm) and deep. The floor of the major group is composed mainly of nitrogen and oxygen atoms that belong to the unique portions of each base pair. In contrast, the floor of the minor groove is filled with nitrogen and oxygen atoms that are generally common to either the purines or to the pyrimidines. Thus, the potential of the major groove for interactions shows a much greater dependence on base sequence than that of the minor groove. This finding led to the speculation that DNA sequence-specific binding proteins recognize DNA by forming hydrogen bonds predominantly to specific groups positioned within and along the major groove. One of the most common ways in which proteins can recognize specific DNA sequences is by the insertion of a protein $\alpha$-helix into the major groove of DNA. The $\alpha$-helix, originally postulated by Pauling and Corey in 1951, is a protein secondary structure motif in which a right-handed helix is formed by amino acids on a polypeptide chain (Pauling *et al.*, 1951). Each amino acid in the helix occupies a vertical distance of 0.15 nm, and there are 3.6 amino acid residues per turn of the helix (see Appendix 1). The diameter of the polypeptide backbone in an $\alpha$-helix is approximately 0.5 nm; however, the amino acid side chains project away from the helical axis. This results in a protein $\alpha$-helix

being able to fit almost exactly into the major groove of double-stranded DNA. The amino acid side chains that project away from the $\alpha$-helix are able to form hydrogen bonds with the DNA bases in the major groove. The type of protein–DNA interaction is shown in Figure 1.10.

### 1.3.4 Hydrogen Bonding

The Watson and Crick model of DNA structure predicts that the two polynucleotide chains are held together by non-covalent hydrogen bonds rather than by covalent interactions. This raises several important questions – what is a hydrogen bond, and is it sufficiently strong to maintain the integrity of the double helix?

To address the first question, a hydrogen bond is a weak electrostatic interaction between a covalently bonded hydrogen atom and an atom with

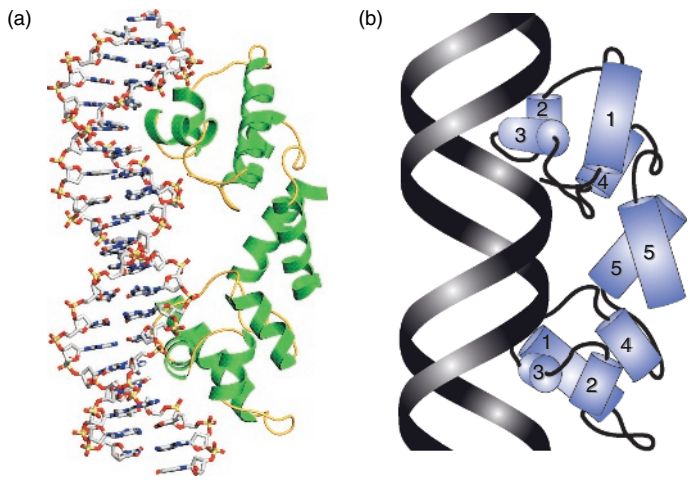



**Figure 1.10.** The interaction between the λ-repressor of bacteriophage λ and DNA. (a) Computer generated model of the interaction between DNA, shown with the same colouring scheme as in Figure 1.9, and λ-repressor, whose $\alpha$-helices are shown in green and are connected by amino acids that do not adopt secondary structure. Two molecules of the λ-repressor (a dimer) interact with a 17 bp segment of double-stranded DNA. (b) The five helices of λ-repressor. Each monomer of the λ-repressor DNA binding domain has five helices, numbered 1–5 from the amino terminal end of the protein. Helix 3 lies in the major groove and the side chains (not shown) extend to the edges of the major groove and make contacts with the DNA bases. Helices 2 and 3 form a 'helix–turn–helix' DNA binding motif that is found in many DNA binding proteins. Helix 3 – the recognition helix – forms DNA sequence-specific contacts in the major groove, while helix 2 – the stabilization helix – interacts non-specifically with the DNA backbone to provide stability to the DNA–protein interaction

an unshared electron pair – a characteristic of covalently bonded oxygen and nitrogen atoms. The hydrogen atom assumes a partial positive charge, while the unshared electron pair assumes a partial negative charge. These opposite charges are responsible for the weak chemical attraction.

In general terms, a chemical bond is an attractive force that holds atoms together. The spontaneous formation of a bond between two free atoms involves the release of internal energy of the unbonded atoms and its conversion to another energy form, e.g. heat. The strength of a particular bond is measured by the amount of energy released upon the formation of the bond. The stronger the bond, the more energy is released. The change in energy ($\Delta G$) that accompanies bond formation is used to describe the strength of a bond. The value of $\Delta G$ is calculated according to the following equation:

$$\Delta G = -RT \ln K_{eq}$$

where $R$ is the universal gas constant (8.314 J $K^{-1}$ $mol^{-1}$), $T$ is the absolute temperature and ln $K_{eq}$ is the natural log of the equilibrium constant between the bonded and the unbonded forms of the molecule. Covalent bonds are short (0.095 nm) and are very strong, with $\Delta G$ values in the range of $-100$ to $-500$ kJ $mol^{-1}$. Hydrogen bonds on the other hand, are longer (approximately 0.3 nm) and are much weaker, with $\Delta G$ values in the range of $-10$ to $-30$ kJ $mol^{-1}$.

As found in the double helix, adenine forms two hydrogen bonds with thymine, and cytosine forms three hydrogen bonds with guanine (see Figure 1.8). While a single hydrogen bond is itself very weak, the 2500 or so hydrogen bonds that hold together every kilobase of DNA provide an extraordinary level of stability to the helix. This also means, as we will see below and in later chapters, that an AT base pair is less stable in thermodynamic terms than a GC base pair.

## 1.4 Reversible Denaturing of DNA

Double-stranded DNA is an immensely stable molecule. In a dehydrated sample, it can survive virtually intact for thousands of years. Fragments of DNA have been isolated from Egyptian mummies that date back some 3000 years. In such samples, the double helix is still intact, but the DNA is fragmented. That is, rather than existing as a chain of millions of nucleotides joined as a contiguous unit, the DNA is composed of short, 500–1000 base pair (bp) fragments, with some of the covalent linkages forming the phosphodiester backbone having been broken.

A major consequence of the non-covalent forces that hold the double helix together is that the two constituent strands of DNA may be separated

(denatured) simply by heating. The thermal energy provided by heating a DNA sample will break the relatively weak hydrogen bonds connecting the two strand of the helix, but will not affect the covalent linkages that hold each strand together (Lin and Chargaff, 1966). The separation of the two DNA strands in a helix is accompanied by changes in the physical properties of DNA. One of these changes is the way in which DNA absorbs UV light. DNA absorbs light at a wavelength of 260 nm due to the presence of alternating single and double bonds in the DNA bases (Figure 1.5). Each of the four nucleotide bases has a slightly different absorption spectrum, and the overall absorption spectrum of DNA is the average of them. A pure DNA solution appears transparent to the eye, and absorption does not become measurable until approximately 320 nm. Moving further into the UV region, there is an absorption peak at about 260 nm, followed by a dip between 220 and 230, and then the solution becomes essentially opaque in the far UV. A solution of double-stranded, native DNA, with a concentration of 0.05 mg/mL, has an absorbance (or optical density, OD) of about 1.0 at the 260 nm peak.

When a DNA helix is denatured to become single strands, e.g. by heating, the absorbance increases. The same 0.05 mg mL double-stranded DNA solution that we used above will have an absorbance of 1.5 when in the single-stranded form. This type of absorbance data is often expressed in the form shown below:

$$1.0\ A_{260\ \mathrm{nm}}\ \text{double-stranded DNA} = 50\ \mu\text{g/mL}$$

$$1.0\ A_{260\ \mathrm{nm}}\ \text{single-stranded DNA} = 33\ \mu\text{g/mL}$$

The increase in absorbance as double-stranded DNA becomes single stranded, called the **hyperchromic effect** (Thomas, 1993), reveals the interaction between the electronic dipoles in the stacked bases of the double helix and is shown diagrammatically in Figure 1.11. The stacking of the base pairs in double-stranded DNA has the effect of dampening the absorption of individual nucleotides due to the competing interactions of adjacent base pairs in the stack. In the single-stranded form, no such competition occurs and consequently single-stranded DNA absorbs light to greater extent than its double-stranded counterpart.

When a DNA sample is heated, the helix begins to separate, in a process sometimes referred to as **melting**. Since AT base pairs have only two hydrogen bonds holding them together, regions of DNA with a high level of A and T residues will separate first. As the temperature continues to increase, more and more of the DNA will assume a single-stranded nature, until eventually complete strand separation occurs. The temperature at which half of the DNA is single stranded is called the **melting temperature** ($T_m$). The melting temperature

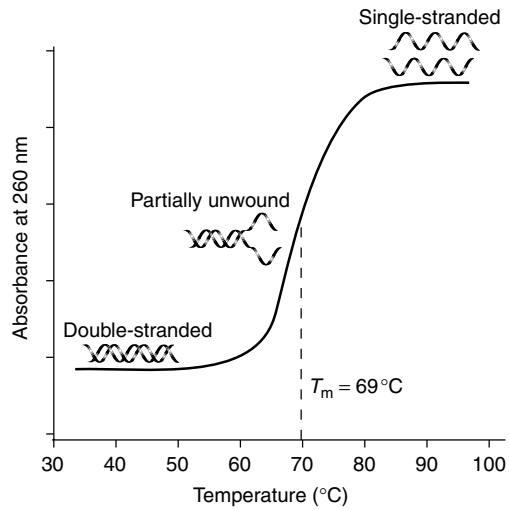



**Figure 1.11.** The effect of heating double-stranded DNA. As a double-stranded DNA molecule is heated above the ambient temperature, the non-covalent hydrogen bonds that hold the two strands together begin to break. This is accompanied by an increase in absorbance at 260 nm. At higher temperatures – above 90 °C – the two DNA strands will completely separate from each other. This process is reversible. If the temperature is reduced relatively slowly, the two complementary strands of DNA will reform to produce a correctly paired double-stranded DNA molecule. The melting temperature ($T_m$) is the temperature at which half the strands are separated

of all DNA molecules is not the same, and depends upon the length of DNA, and the proportion of GC and AT base pairs that it contains. Generally, short DNA molecules have low $T_m$ values, as do those that contain a high proportion of AT base pairs. The melting temperature of a long DNA molecule can be calculated according to the following equation:

$$T_m = 16.5(\log[Na^+]) + 0.41(\%GC) + 81.5\,^\circ C$$

where $[Na^+]$ is the concentration of sodium ions in the DNA solution and (%GC) is the percentage of GC residues in the duplex. Thus, for a DNA molecule that contains 40% GC residues (typical for a mammalian genome) in a solution of 50 mM NaCl, the melting temperature is 76.4 °C. As we will see in later chapters, short DNA molecules (15–30 bases long) are often used in genetic engineering experiments. For short regions of complementation, the equation above breaks down, and the following estimation (sometimes called the Wallace rule) is used to determine the $T_m$ (Wallace *et al.*, 1979):

$$T_m = 2(AT) + 4(GC)$$

That is, every AT base pair in the duplex contributes 2 °C of stabilization the double helix, while every GC pairing contributes 4 °C of stabilization. Therefore, an oligonucleotide of 20 bases of single-stranded DNA containing five A residues, six T residues, three C resides and six G residues will have an annealing temperature to its complementary DNA sequence of approximately $2(11) + 4(9) = 58$ °C. The importance of the annealing temperature will become apparent when we look at polymerase chain reaction (PCR) and nucleic acid hybridization in later chapters.

The heating of double-stranded DNA to produce the single-stranded version is an entirely reversible process. Single-stranded DNA made by heating a duplex will reform when the temperature is reduced. As the temperature falls, the thermal energy that was used to break the hydrogen bonds between the strands is reduced, and random collisions between the complementary strands will result in their re-association. Providing that the temperature is reduced relatively slowly (1–2 °C per second), complete duplex formation will result. The two complementary single-stranded DNA molecules will come together and reform the exact duplex that was present before the sample was heated. If the temperature is dropped rapidly, for example by plunging a DNA sample at 94 °C directly into ice, correct base pairing will not occur and the DNA will remain in a relatively single-stranded form.

## 1.5 Structure of DNA in the Cell

Different types of nucleic acid are used to form the genome of organisms depending on the organism itself. For example, viruses have a genome composed of either double-stranded DNA, single-stranded DNA or RNA, depending on the type of virus. The chromosomes of most eukaryotes are composed of single linear double-stranded DNA molecules. The genomes of prokaryotic organisms are generally composed of a circular DNA molecule. That is, rather than having free 5′- and 3′- ends, the ends are joined to each other to form a continuous ring of double-stranded DNA. A number of extra-chromosomal DNA molecules, called **plasmids**, are found in prokaryotic cells. As we will see in Chapter 3, understanding how cells deal with these plasmids has played a pivotal role in advances in molecular biology and genetic engineering. Plasmid DNA molecules are usually closed circles of either single-stranded or double-stranded DNA.

When looking at the structure of DNA as presented in Figure 1.7, it is easy to get the impression that the double helix is a rather inflexible solid rod. This, however, is simply not the case. Electron microscopy images of DNA molecules (Figure 1.12) indicate that DNA is more string-like than rod-like, and will wrap around itself to form a variety of irregular structures. Inside a eukaryotic

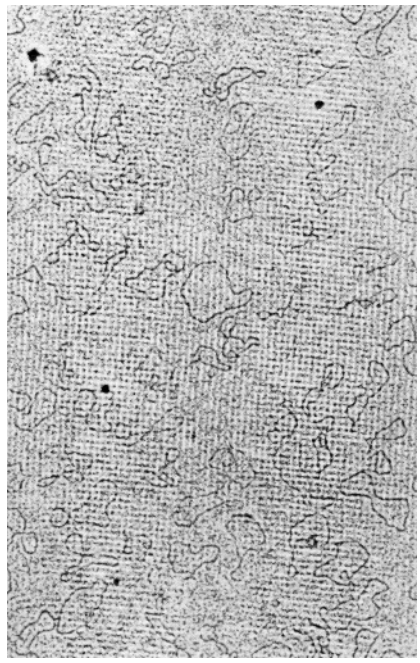


**Figure 1.12.** Images of double-stranded DNA generated by electron microscopy. Circular DNA isolated from polyoma virus shows many crossings of the DNA double strands, indicating that the DNA is supercoiled. Reproduced from Vinograd *et al.* (1965)

cell, DNA is associated with a vast array of proteins – for example, proteins required for its replication, its transcription into RNA and its packaging within the cell. Many of these proteins wrap DNA around themselves, or in other ways constrain or bend the DNA molecule. For short linear DNA molecules, this type of constraint is not a major problem. Stresses placed in one part of a DNA molecule by, for example, twisting the double helix can be relieved by untwisting another part of the DNA. The free ends of the linear DNA allow relatively free rotation of the DNA strands. For circular DNA molecules, however, these stresses can prove extremely problematic. Since the DNA has no free ends, it cannot simply untwist to counteract a twist elsewhere in the molecule. Twisting DNA molecules results in the formation of DNA **supercoils** (Figure 1.13).

In 1965 Jerome Vinograd and his colleagues suggested that closed-circular DNA molecules could adopt a 'twisted circular form' (Vinograd *et al.*, 1965). Such a form would result if, before joining the ends of a linear duplex DNA into a closed circle, one end was twisted relative to the other to introduce some strain into the molecule. Such coiling of the DNA helix upon itself is called **supercoiling**. Supercoiling can only be introduced into or released from

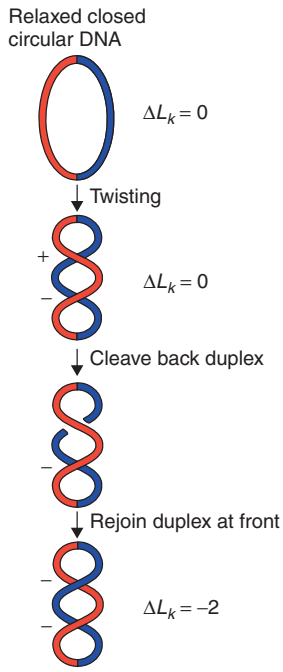



**Figure 1.13.** The formation of DNA supercoiling. A relaxed closed-circular DNA molecule containing no supercoils is shown. The two halves of the molecule have been arbitrarily coloured red and blue to aid the visualization of DNA strands crossing over each other. A twist in the molecule will result in the formation of a positive (+) and a negative (−) supercoil. This process does not result in a change in the linking number. If, however, the DNA backbone is broken and subsequently resealed after a segment of DNA has passed through the break, then the linking number is reduced by 2. The molecule is now negatively supercoiled

a closed-circular DNA molecule by breaking at least one of the phosphodiester backbones.

The level of supercoiling within a particular DNA molecule can be described by its linking number ($L_k$). This number corresponds to the number of double-helical turns in the original linear molecule. The process of introducing supercoils into a DNA molecule reduces or increases the number of helical turns trapped when the circle is formed, and results in changes in the linking number to the final closed-circular species. The linking difference ($\Delta L_k$) gives an indication of the overall levels of supercoiling within a molecule (Figure 1.13). Interested readers should refer to dedicated texts for a fuller description of DNA topology and its associated problems (Bates and Maxwell, 1993).

Closed-circular DNA isolated from prokaryotic cells is negatively supercoiled. That is, the DNA is relatively underwound in comparison to its linear

state. Although eukaryotic cells contain linear DNA molecules, the length of DNA within a chromosome means that constraints placed in one part of the DNA molecule cannot be easily relieved by transferring the strain to one of the ends. An appropriate analogy here is to think about how you would unkink a twisted garden hosepipe. You would transfer the kink to the end of the pipe by untwisting it. At the end, the rotation of the pipe itself will result in the release of the kink. This type of action to release a kink or twist in a linear DNA molecule would be relatively straightforward if the DNA were short. Long chromosomal DNA molecules, however, suffer from a number of difficulties that make this type of constraint release impractical. The sheer length of DNA involved would make it difficult to move twists from the middle of the DNA to the ends, and the association of the DNA with proteins will provide a barrier to this type of movement. Coming back to the hosepipe analogy, another way to remove the kinks (although less practical in the garden) would be to cut the pipe to create new ends near the site of the twist. The pipe could then be untwisted locally and resealed to remove the kink and reform the pipe. It is in this fashion that cells deal with strains in DNA. They possess enzymes, called **DNA topoisomerases**, which make cuts in DNA molecules – either single- or double-stranded breaks in the phosphodiester backbone – to allow local twisting and untwisting of the DNA chain. The topoisomerase enzyme then reseals the DNA backbone to reform the DNA molecule.

## 1.6 The Eukaryotic Nucleosome

Each cell within our body contains a huge amount of DNA. The different chromosomes of the human genome contain approximately $3.2 \times 10^9$ base pairs of DNA. Since we are diploid organisms, having two sets of each chromosome, the total amount of DNA in most of our cells totals $6.4 \times 10^9$ base pairs. At 0.33 nm per base pair (Figure 1.7), this corresponds to an overall length of approximately 2.1 m. How can this fit into a nucleus measuring just 5–10 μm across? The answer is that the DNA is highly compacted. It is associated with a number of proteins that results in the wrapping of DNA into **nucleosomes**. During interphase, the genetic material (together with its associated proteins) is relatively uncoiled and dispersed throughout the nucleus as **chromatin**. When mitosis begins, the chromatin condenses greatly, and during prophase it is compressed into recognizable chromosomes. This condensation represents a contraction in length of some 10 000-fold.

The genetic material when isolated from bacteria and viruses consists of strands of DNA or RNA almost devoid of proteins. In eukaryotes, however, a substantial amount of protein is associated with the DNA to form chromatin.

Electron microscopic observations have revealed that chromatin fibres are composed of linear arrays of spherical particles. The particles occur regularly along the axis of a chromatin strand and resemble beads on a string. These particles, initially referred to as *ν*-bodies (Olins and Olins, 1974), are now called nucleosomes.

The digestion of chromatin with certain nucleases, such as micrococcal nuclease, yields DNA fragments that are approximately 200 bp in length, or multiples thereof (Wingert and Von Hippel, 1968). If the digestion of chromatin DNA were random, then a wide range of fragment sizes would be produced. This therefore demonstrates that the DNA of chromatin consists of repeating units that are protected from enzymatic cleavage. The DNA between the units is attacked and cleaved by the nuclease, and multiples occur where two or more units are joined together.

The proteins associated with DNA in chromatin are divided into basic, positively charged **histones** and less positively charged **non-histones**. Of the proteins associated with DNA, the histones play the most essential structural role. Histones contain large amounts of the positively charged amino acids lysine and arginine (see Appendix 1), making it possible for them to bind through electrostatic interactions to the negatively charged phosphate groups of the DNA nucleotides. There are five different types of histone protein – H1, H2A, H2B, H3 and H4. A nucleosome core particle consists of two copies each of histones H2A, H2B, H3 and H4 to form a histone octamer around which ~150 base pairs of DNA are wrapped in a left-handed superhelix, which completes about 1.7 turns per nucleosome (Figure 1.14). Recent insights into the structure of the nucleosome came in 1997 when Richmond and colleagues were able to solve the X-ray crystal structure of the nucleosome–DNA complex at high resolution (Luger *et al.*, 1997). At this resolution, most atoms of each histone are visible, and the precise path the DNA helix as it encircles the histone octamer can be traced (Figure 1.15). The high-resolution structure also revealed that the amino-terminal ends of some of the histones protrude from the octamer and project away from the nucleosome. The significance of the amino-terminal tails is that they have the potential to interact with adjacent nucleosomes to create nucleosome–nucleosome contacts. We will discuss the significance of the histone tails further when we look at the regulation of gene expression.

Extensive investigation of the structure of nucleosomes has provided the basis for predicting how the chromatin fibre within the nucleus is formed, and how it coils up into the mitotic chromosome. This model is illustrated in Figure 1.16. The 2 nm DNA double helix is initially coiled into a nucleosome core particle that is about 10 nm in diameter. Approximately 200 base pairs of DNA link each core particle to form the 'beads on a string' seen in electron

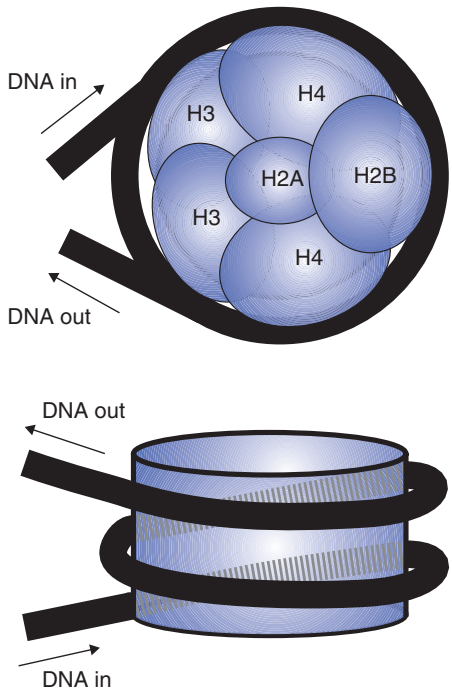



**Figure 1.14.** Wrapping of DNA around the nucleosome core. The nucleosome is composed of two molecules each of histones H4, H3, H2A and H2B. In the representation shown here only a monomer of H2A and H2B can be observed, the other monomers being located at the back of the octamer. Almost 150 bp of DNA wrap around the octamer core, forming approximately two turns

microscopy images. Histone H1, which is not part of the core octamer, may be located at the site where DNA enters and leaves the nucleosome and possibly functions to seal the DNA around the nucleosome.

The formation of nucleosomes represents the first level of packing, whereby the DNA is reduced to about one-third of its original length. In the nucleus, however, chromatin does not exist in this extended form. Instead, the 10 nm chromatin fibre is further packed into a thicker 30 nm fibre, which was originally called a solenoid. It is not clear whether the transition between the 10 nm fibre and the 30 nm fibre represents a physiological event or whether it merely occurs *in vitro* as a consequence of altering the salt concentration. The 30 nm fibre does, however, consist of numerous nucleosomes packed closely together, but the precise orientation and details of the structure are not clear. It has recently been suggested that the 30 nm fibre might adopt a compact helical zig-zag pattern with about four nucleosomes per 10 nm (Beard and Schlick,



**Figure 1.15.** The structure of the nucleosome–DNA complex as solved by X-ray crystallography at high resolution. In both the top and side views of the complex, the strands of the DNA (shown in green and orange) wrap around the histone octamer. The protein core is almost exclusively internal with the exception of the amino-terminal histone tails that extend out from the complex. This figure was kindly provided by Tim Richmond (ETH Zurich) and is reprinted by permission from *Nature* (Luger *et al.*, 1997) copyright (1997) Macmillan Publishers Ltd

2001). The formation of the 30 nm fibre creates a second level of packaging, in which the overall length of the DNA is reduced some two fold.

In the transition to the mitotic chromosome, still another level of packing occurs. The 30 nm fibre forms a series of looped domains that further condense the structure of the chromatin fibre. The fibres are then coiled into the chromosome arms that constitute a **chromatid**, which is part of the metaphase chromosome.

In the overall transition from fully extended DNA helix to the extremely condensed status of the mitotic chromosome, a packaging ratio of about 500:1 must be achieved. The model presented above only accounts for a ratio of about 50:1. The remainder of the packing arises from the coiling and folding of the 30 nm fibre. The tight packing of the DNA into a chromosome presents an enormous challenge to both the replication of DNA and to its **transcription**. We will discuss these issues in the sections below.

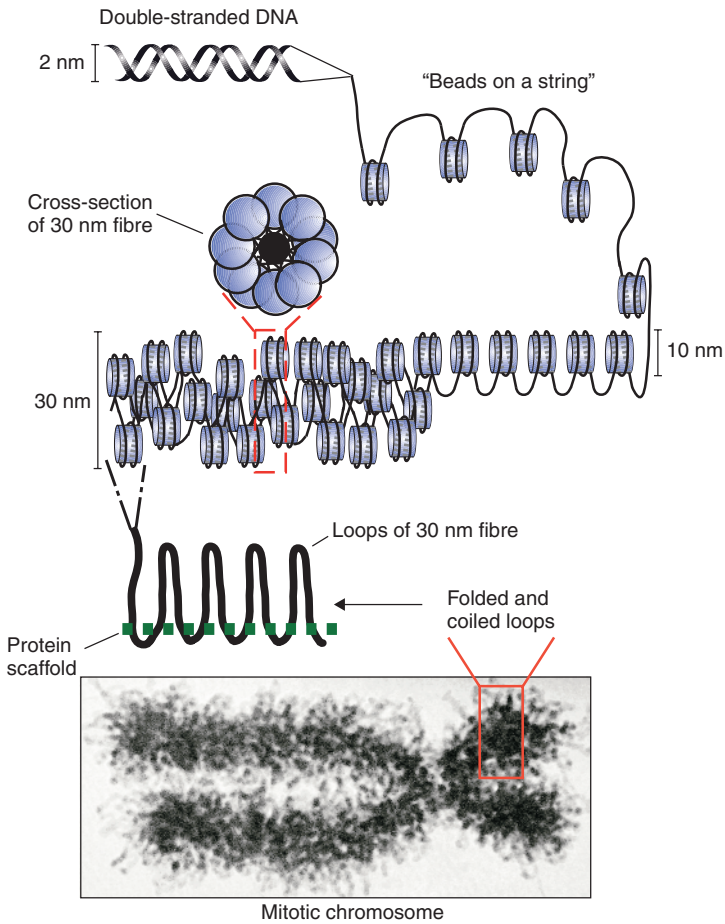



**Figure 1.16.** Packaging of DNA into the eukaryotic chromosome. The 2 nm DNA double helix is wrapped into nucleosomes as illustrated in Figure 1.15. At some point histone H1 enters the complex, possibly at the DNA entry/exit point on the nucleosome. The nucleosomes can form extended 10 nm fibres, which are long arrays of ordered nucleosomes. The nucleosomes can further condense to form a 30 nm fibre. This may be a 'zig-zag' array of nucleosomes. The 30 nm fibre is then wrapped onto a protein scaffold, which can be additionally folded and coiled to form the mitotic chromosome that is observed under the electron microscope

## 1.7 The Replication of DNA

Watson and Crick finished their famous 1953 paper with the statement 'It has not escaped our notice that the specific pairing we have postulated immediately suggests a possible copying mechanism for the genetic material'. It was apparent to them from the arrangement and nature of the bases that each strand

of the DNA double helix could serve as a template for the synthesis of a complementary strand (Figure 1.17). They proposed that if the double helix were unwound, each nucleotide on the parental strands would have an affinity for its complementary nucleotide (Watson and Crick, 1953b). Thus, each of the parental DNA strands could act as a template for the synthesis of a new DNA strand that would result in the production of two identical double-stranded DNA duplexes. In this model, each of the newly synthesized DNA molecules consists of one parental and one newly synthesized DNA strand; hence, the process is known as **semi-conservative replication**.

Two other modes of replication that also rely on the parental strands serving as templates for new DNA synthesis were also considered. In **conservative**

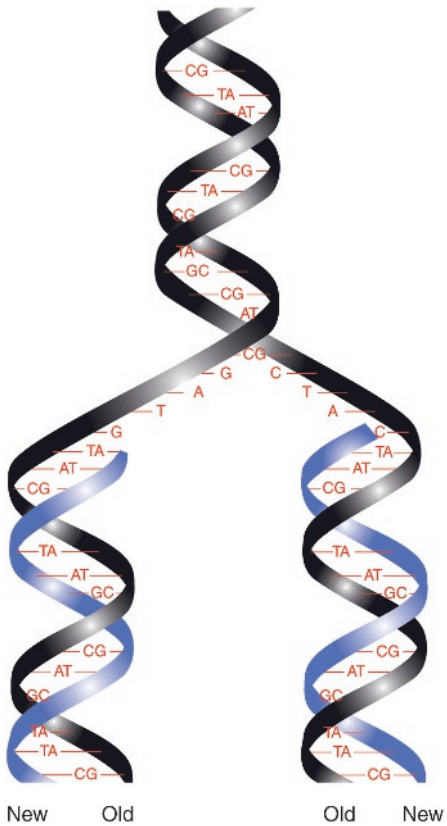



**Figure 1.17.** The replication of double-stranded DNA as suggested by Watson and Crick. The separation of the DNA strands of the parent helix (black) means that each strand can act as a template of the synthesis of a new strand of DNA (blue). This is termed semi-conservative replication, as the daughter DNA molecules contain one old strand and one new strand

**replication**, the parental helix would be opened to reveal the base sequence, and a newly synthesized helix would be produced. The two newly synthesized DNA strands would come together to form a helix and the parental helix would reform. In the second alternative mode, called **dispersive replication**, the parental DNA from one strand could be dispersed into either of the newly synthesized strands, which would then be a mixture of new and old DNA.

Matthew Meselson and Franklin Stahl provided the experimental evidence to suggest that semi-conservative replication was used by bacteria to produce new DNA molecules (Meselson and Stahl, 1958). The results of their classic experiment are depicted in Figure 1.18. They grew *E. coli* cells for many generations in a medium in which the sole source of nitrogen was a heavy – but

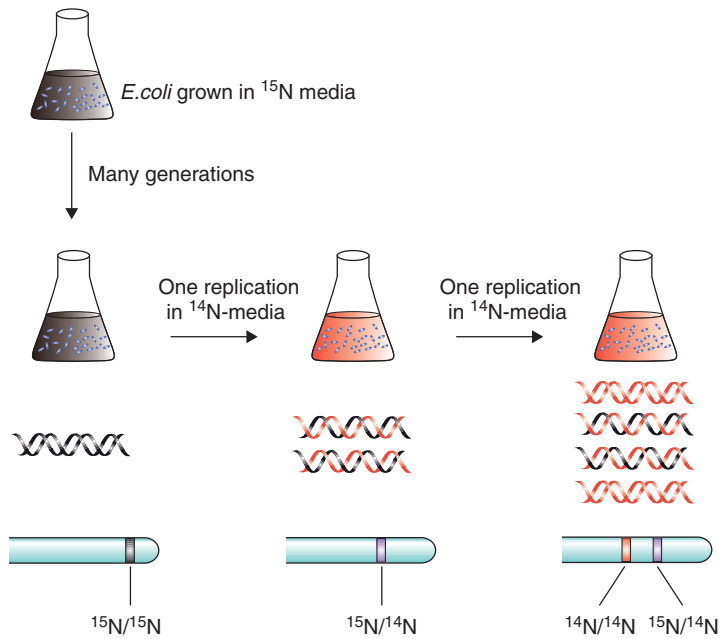



**Figure 1.18.** The Meselson–Stahl experiment. Bacterial cells were grown for many generations in a medium containing the heavy isotope of nitrogen, $^{15}N$. The DNA isolated from such bacteria was isolated and centrifuged through a density gradient. The mobility of the DNA band was noted. The bacteria were then transferred to media containing the normal isotope of nitrogen ($^{14}N$). After one generation, all the DNA isolated from the bacteria ran through the density gradient with a mobility different to that seen previously. After two generations, two DNA species were observed. One had the same mobility as that seen after one generation, and the other ran with the same mobility as DNA isolated from bacteria grown on unlabelled media. In subsequent generations, the intensity of the latter band increased and the band of intermediate density decreased. These data are most readily explained by DNA replication occurring through a semi-conservative mechanism

non-radioactive – derivative of ammonium chloride ($^{15}NH_4Cl$). This meant that the DNA of these *E. coli* cells, into which the $^{15}N$ was incorporated as part of the bases, was heavier than the DNA from cells grown on normal ammonium chloride ($^{14}NH_4Cl$). The normal and heavy forms of DNA could be distinguished after centrifugation through a density gradient of caesium chloride. The heavier DNA will travel further under such conditions than the lighter normal DNA. This provided a means of monitoring the levels of each nitrogen isotope within the bacterial replicated DNA as it was replicated. Meselson and Stahl took their bacteria that contained heavy DNA and transferred them into media containing normal ammonium chloride and then isolated DNA after each successive generation. After a single round of DNA replication on the normal media, they found that the bacterial DNA now possessed an intermediate density between the expected positions for wholly $^{15}N$-containing and wholly $^{14}N$-containing DNA. This result is consistent with semi-conservative replication, but not with conservative replication, where two distinct bands would have been expected. After two cell divisions, Meselson and Stahl did observe the presence of two bands of different densities – the first corresponding to the position of a mixed $^{14}N/^{15}N$ DNA species and the second, of equal intensity, corresponding to the expected position of unlabelled $^{14}N/^{14}N$ DNA. Subsequent rounds of DNA replication resulted in an increased proportion of the $^{14}N/^{14}N$ DNA band.

The experiments described above suggest that conservative replication does not occur, but could not rule out dispersive replication. Meselson and Stahl, however, dismissed this mode of replication by the analysis of individual single strands of replicated DNA. They heated their DNA (Figure 1.11) and found that individual strands possessed either the density of $^{15}N$- or $^{14}N$-containing single-stranded DNA, but not that of an intermediate density – which would have been expected if a newly synthesized DNA strand contained a mixture of the parental and newly synthesized sequences. Semi-conservative replication has also been shown to occur in eukaryotes (Taylor, Woods and Hughes, 1957), and is now widely accepted as the universal mechanism for DNA replication.

## 1.8 DNA Polymerases

Both bacteria and eukaryotic cells possess multiple DNA polymerase enzymes. *E. coli* contain three such enzymes (Table 1.2). The first to be isolated, now called DNA polymerase I, is primarily involved in DNA repair rather than DNA replication. DNA polymerase II is also involved in DNA repair, while DNA polymerase III is the main replicating enzyme (Kornberg and Baker, 1992). DNA polymerase III is a multisubunit enzyme that functions as a dimer of these

**Table 1.2.** The properties of DNA polymerases I, II and III from *E. coli*

| | Pol I | Pol II | Pol III |
|---|---|---|---|
| Main function | DNA repair and primer removal | DNA repair | DNA replication |
| Size, kDa | 109 | 120 | 140 |
| Structural genes | *polA* | *polB* | *DnaE*, *N*, *Q*, *X*, *Z* |
| Molecules per cell | 400 | | 10–20 |
| 5′ to 3′ polymerization | + | + | + |
| 3′ to 5′ exonuclease | + | + | + |
| 5′ to 3′ exonuclease | + | − | − |

multiple subunits. Many DNA polymerase enzymes have been studied at the structural level, which, combined with extensive biochemical analysis, has led to the elucidation of their molecular mechanism of action.

DNA polymerase enzymes can possess three distinct enzymatic activities.

- All DNA polymerases direct the synthesis of DNA fragments in a 5′ to 3′ direction.
- Many DNA polymerases also possess 3′ to 5′ exonuclease activity. This enables the enzyme to remove nucleotides at the 3′- end of the newly synthesized chain. This is often considered a 'proof-reading activity', which can be used to excise incorrectly incorporated nucleotides before their replacement with the correct bases.
- Some DNA polymerases also have a 5′ to 3′ exonuclease activity. This allows the enzyme to remove sequences that have already been synthesized and, as we will see below, is involved in joining up discontinuous DNA fragments produced on the lagging strand during DNA replication.

DNA polymerases are not able to initiate DNA synthesis on their own. They have an absolute requirement for a free 3′-end onto which the enzyme can add new nucleotides. This means that two primers are required to initiate DNA replication in each direction from the origin – one complementary to each DNA strand to be replicated. As stated above, DNA synthesis only proceeds in a 5′ to 3′ direction. That is, new nucleotides are added on to the 3′ end of an existing polynucleotide chain. This leads to a problem. Both strands of the parent DNA molecule must be replicated, but only one of the strands is orientated in a 5′ to 3′ direction from the origin of replication. This strand, called the

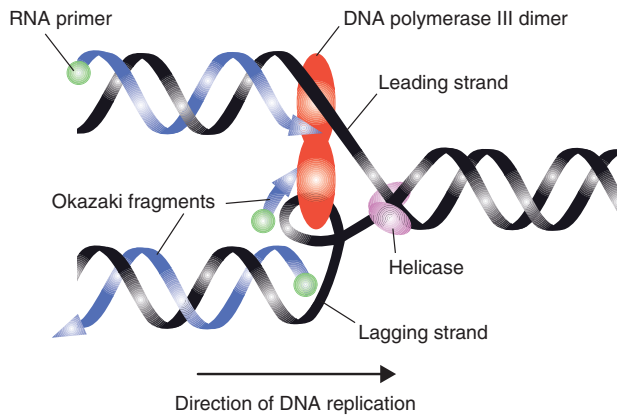



**Figure 1.19.** DNA replication by DNA polymerase III. The strands of the parent DNA molecule (black) are unwound by a helicase at the replication fork. The leading strand and the lagging strand are then fed into a dimer of DNA polymerase III. DNA replication on the leading strand occurs in a 5′ to 3′ direction by extending a single RNA primer. Many primers are required to produce the lagging strand in short sections (Okazaki fragments), which are later joined using DNA ligase

**leading strand**, can be copied in a continuous manner – DNA replication can occur without stopping as the DNA polymerase is working in its favoured orientation. Replication of the other strand, the **lagging strand**, would appear to have to occur in the 3′ to 5′ direction, which is not possible. The lagging strand is replicated in a discontinuous fashion (Figure 1.19). For discontinuous strand synthesis, a series of short DNA segments are produced in a 5′ to 3′ direction (Okazaki *et al.*, 1968) that are then ligated together to produce the intact newly synthesized strand.

An enzyme that we will return to in later chapters is the RNA-dependent DNA polymerase called **reverse transcriptase**. This enzyme is found in certain viruses (called retroviruses) that contain RNA, rather than DNA, as their genetic material. Following infection of the host cell, the viral RNA serves as a template for the synthesis of a complementary DNA (**cDNA**) molecule. The DNA may then be incorporated into the host's genome where, if the DNA is expressed, the retroviral RNA genome will be produced. The ability to use the reverse transcriptase enzyme *in vitro* to produce DNA from RNA will be discussed in Chapter 5.

## 1.9 The Replication Process

How does DNA replication start and finish? As we have already seen, the DNA molecules inside cells are big and the need to tightly control the replication of

DNA sequences is obvious. We can consider the replication process in three phases – initiation, elongation and termination.

- *Initiation of replication*. The initiation of DNA replication does not occur at random sites around the genome, but rather it is initiated at specific points called **origins of replication**. Once DNA synthesis has been initiated, two replication forks, extending in either direction from the origin of replication, proceed to allow the full replication of the genome. Bacteria, such as *E. coli*, have a single origin of replication (called *Ori*C). Eukaryotic cells, on the other hand, have multiple origins of replication which are different from *Ori*C – the yeast *Saccharomyces cerevisiae* has been estimated to have about 300 replication origins, whilst human cells utilize over 20 000 origins during the replication of the genome. *Ori*C is a 245 bp DNA sequence that acts as the binding site for a number of proteins (namely DnaA, B and C). The binding of these proteins promotes the melting (opening) of the DNA helix, a process that is essential so that DNA replicating enzymes can read the base sequence. Replication origins in both prokaryotes and eukaryotes probably serve the same overall function, but the replication origins of prokaryotic cells will not substitute for their eukaryotic counterparts and *vice versa*. Once the helix has been unwound, the DNA polymerase can access the base sequences by 'reading' the base pair hydrogen bonds such that a complementary DNA chain may be synthesized. However, as we noted above, the polymerase can only function if a free 3′ hydroxyl group is present. This hydroxyl group is provided by an RNA primer (which is complementary to the DNA) that is 5–15 nucleotides long. The synthesis of the primer is directed by a form of RNA polymerase (called primase) that does not require a free 3′- end to initiate synthesis. DNA replication then proceeds simultaneously on both strands. Other proteins are also required for DNA replication to occur. The sections of single-stranded DNA produced during replication are stabilized through the binding of single-stranded binding proteins (SSBs), and the DNA is unwound into the polymerase complex with the help of DNA helicases (Figure 1.19). Additionally, topoisomerase enzymes (e.g. DNA gyrase) are required to relieve tension in the helix that results as a consequence of the unwinding process.

- *Elongation*. Given that DNA polymerase III can only replicate in a 5′ to 3′ direction, how can both DNA strands be simultaneously replicated? A mechanism for leading and lagging strand DNA synthesis is shown in Figure 1.19. In this model, DNA polymerase III, which is a dimer, is

able to replicate the leading strand in a continuous fashion. The lagging strand, however, forms a loop so that nucleotide polymerization can occur on both template strands in a 5′ to 3′ direction. Looping will invert the orientation of the template with respect to the enzyme but not the direction of actual synthesis on a lagging strand. After the synthesis of approximately 1000–2000 base pairs, the monomer of the enzyme on a lagging strand encounters a completed Okazaki fragment, at which point it releases the lagging strand. A new loop is then formed with the lagging strand and the process is repeated. The 5′ to 3′ exonuclease activity of DNA polymerases may aid the complete formation of Okazaki fragments. DNA ligase is then required to join the phosphodiester backbone of the Okazaki fragments to form a complete newly synthesized lagging strand. We will discuss the mechanism of action of DNA ligase and its use in genetic engineering in Chapter 2.

- *Termination*. Just as important as the correct initiation of DNA synthesis is its correct termination. For bacteria, such as *E. coli*, which contain a circular genome, if replication initiates bidirectionally from a single point, then termination should occur at a point halfway round the genome. Termination is not a passive process. It occurs at defined DNA sequences (called terminator sequences) that act as binding sites for a protein called Tus. Tus binding to the terminator sequences is highly asymmetric, which allows replication forks travelling in one direction to pass, but blocks DNA replication in the opposite direction (Kamada *et al*., 1996). Little is known about the termination of DNA replication in eukaryotes.

In eukaryotic cells, genome replication must be coordinated with the cell cycle so that two copies of the entire genome are available when the cell divides. The cell cycle is a four-stage process that is based upon microscopic observations of dividing cells (Figure 1.20). These observations showed that dividing cells pass through repeated cycles of metaphase, when nuclear and cell division occurs, and interphase, where few changes can be detected using a microscope. DNA replication occurs during interphase. The four stages of the cell cycle are the following.

- M-phase (mitosis) – the period when the nucleus and the cell divides.
- G1-phase (gap 1) – a growth phase where transcription, translation and other general cellular activities occur.
- S-phase (synthesis) – where DNA synthesis occurs and the genome is replicated.

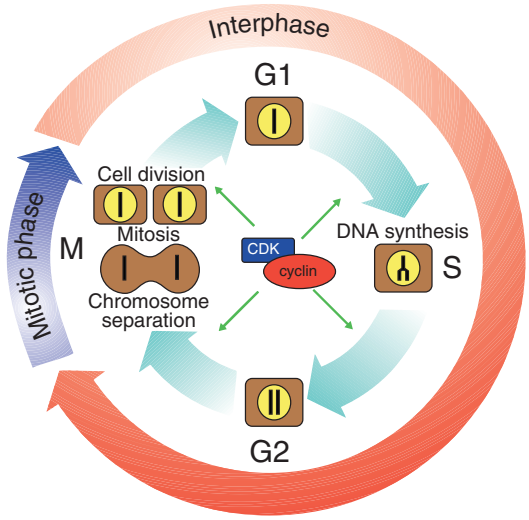



**Figure 1.20.** The eukaryotic cell cycle is split into cell division (mitosis) and the period between divisions (interphase). After a division is complete, the cell grows during the first gap phase (G1). DNA synthesis then occurs (S) before a second gap phase (G2). Finally, the duplicated chromosomes separate and cell division occurs (M). The decisions to proceed with particular phases of the cycle are controlled by proteins – cyclins and cyclin-dependent kinases (CDKs). These are termed checkpoint controls

- G2-phase (Gap 2) – the second interval period. Another growth phase when final preparations for cell division take place.

A human cell in culture takes about 20 hours to progress through one complete cell cycle. Of this, over 9 h will be spent in G1, while S-phase takes about 8 h to complete, and G2 lasts about 2 h. The actual time the cell spends dividing (i.e. in M-phase) is only about 45 min. Clearly, it is important that the S- and M-phases are coordinated so that the genome is completely replicated once, but only once, before cell division occurs. The periods immediately before entry into S- and M-phases are the key **cell-cycle checkpoints**. Proteins, cyclins and cyclin-dependent kinases (CDKs) act to control entry into each phase of the cell cycle. Cyclins are a diverse family of proteins that bind and activate some members of the CDK family. Most cyclins display dramatic changes in concentration during the cell cycle, which help to generate the oscillations in CDK activity that form the foundation of the cell-cycle control. Other checkpoints also exist, primarily within S-phase, to ensure that, for example, DNA damage is repaired before replication can be completed. Lee Hartwell, Paul Nurse and Tim Hunt were awarded the 2001 Nobel Prize in Physiology or Medicine for their discoveries of key regulators of the eukaryotic cell cycle.

## 1.10 Recombination

DNA should never be considered as a static molecule. It is constantly changing, and one mechanism by which this change is brought about is recombination. Recombination is the large-scale rearrangement of a DNA molecule. This type of rearrangement occurs as a consequence of two DNA molecules sharing either extensive regions of similar sequence (**homologous recombination**) or very short regions of homology (**site-specific recombination**). Cells containing a diploid set of chromosomes have plenty of opportunities to find a homologous partner for recombination to occur. Recombination in a bacterial system was first demonstrated independently by Alfred Hershey and Max Delbrück in 1947. They studied the infection of *E. coli* with bacteriophages. If an *E. coli* cell was infected at the same time with two genetically different T2 bacteriophages, the resulting phage population included recombinant phage types as well as the original parental phage types (Hershey, 1947).

Homologous recombination occurs between two DNA molecules that have essentially the same sequence. The genetic information is swapped and mixed up between the two DNA versions such that the recombined DNA molecules are a mixture of the starting ones. Several mechanisms have been proposed to explain the molecular basis of these events. The key to understanding the molecular processes involved in recombination was first articulated in 1964 by Robin Holliday (Holliday, 1964). His model for recombination required that a nick (cleavage at the phosphodiester bond between two nucleotides) occurred at the same site in two homologous DNA molecules (called the donor and recipient DNA). A strand exchange reaction then occurred, followed by sealing of the nicks using DNA ligase. The exchange of donor and recipient DNA strands leads to the formation of a **Holliday junction** – a structure in which the crossed DNA strands connect the donor and recipient DNA molecules. The Holliday junction can undergo branch migration, leading to more recombination between strands from different DNA molecules. Finally, resolution of the junction gives rise to the recombined DNA molecules. Here, however, we will concentrate on the modified mechanism proposed by Meselson and Radding (1975) since the event initiating recombination can be brought about by a single nick in one of the DNA strands rather than earlier models, which relied on two DNA nicks before recombination could occur. The basic model is shown in Figure 1.21(a). The formation of the Holliday junction and its resolution, together with some of the *E. coli* proteins known to be involved in the process, are outlined below.

- The nick is created by the RecBCD endonuclease, which cleaves DNA strands at sequences called chi ($\chi$) sites (5′-GCTGGTGG-3′).

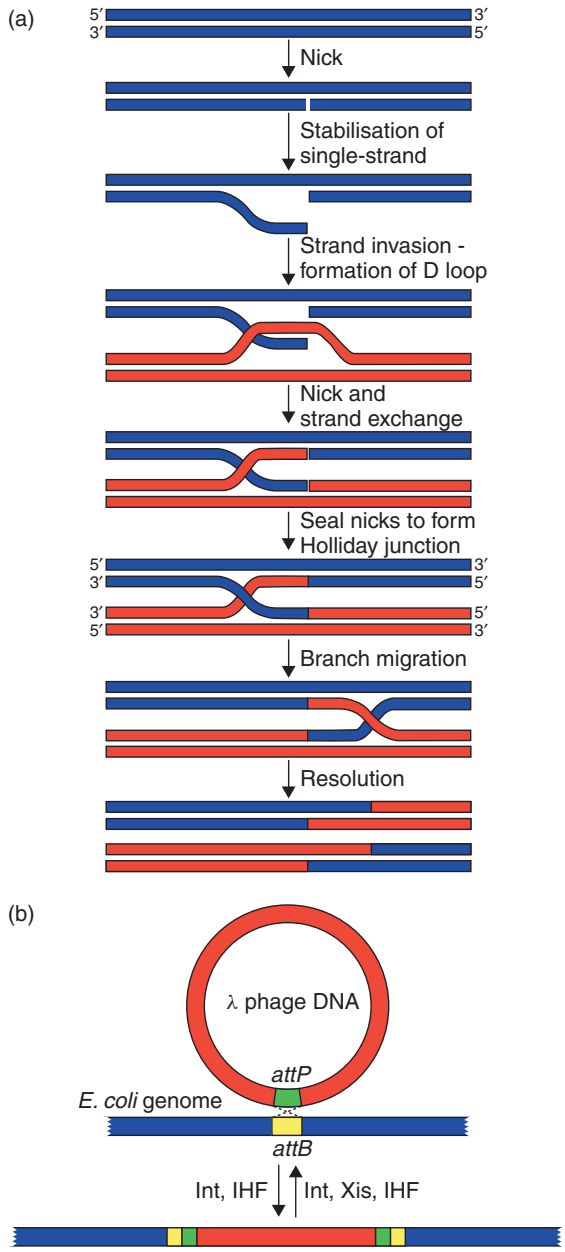



**Figure 1.21.** Homologous and site-specific recombination. (a) The Meselson–Radding model for homologous recombination. (b) The site-specific integration of λ DNA into the *E. coli* genome. See text for details

- Strand invasion is catalysed by the RecA protein, along with single-strand DNA binding protein (SSB), which stabilizes the free single DNA strand.
- The 'D-loop' that is formed after strand invasion is cut, again by the RecBCD endonuclease.
- Strand exchange occurs and the ends are sealed using DNA ligase. This forms the Holliday junction.
- The formation of the Holliday junction is followed by branch migration, catalysed by the proteins RuvA and RuvB.
- The resolution of the Holliday junction requires the RuvC protein, which nicks two of the DNA strands. DNA ligase then seals the strands after they have been cut.

At the end of these processes, the cell will have two molecules that have exchanged DNA. The same process outlined above also takes place in eukaryotes, using similar kinds of enzymatic activity. Recently, the high-resolution structure of the Holliday junction stabilized by the RuvA protein has been solved (Hargreaves *et al.*, 1998).

Unlike homologous recombination, site-specific recombination occurs when a particular DNA, that is not at all homologous to another DNA, has a sequence that acts as a target for the recombination of that other DNA. In prokaryotes, this is the case when bacteriophage λ integrates its DNA into the *E. coli* chromosome, where it will reside in the lysogenic state. We will look at the life cycle of λ in more detail in Chapter 3. This recombination event occurs between sites in the bacteriophage DNA (*attP*) that have a sequence homology with sites in the bacterial DNA (*attB*), even though the two DNA molecules are completely different in all other sequences. This is shown in Figure 1.21(b). The enzyme that catalyses this event is called integrase (Int), along with a host protein called integration host factor (IHF). When the bacteriophage DNA is excised from the *E. coli* genome to begin the formation of new phage particles, a phage protein called Xis is required. This site-specific recombination still requires the base pairing of a short homologous region (the *attP* with the *attB* sites), but does not involve any of the proteins of the homologous recombination pathways.

## 1.11 Genes and Genomes

The genetic information contained within the DNA base sequence directs the production of the proteins and enzymes that build the cell. The nucleic

acids are arranged into units called **genes**, and the whole collection of genes constitutes the **genome**. Most genes are used as templates to produce proteins. The complement of proteins within a particular cell type is distinctive – a hair follicle will produce keratin and a pancreatic $\beta$-cell will produce insulin – but the protein content of a cell can also change dramatically depending upon, for example, the availability of nutrients. This means that in certain cell types, certain genes will be expressed and others will not, and the cell must have the ability to activate certain genes in response to external signals. This raises a number of issues. How is the cell able to discern within its genome what is a gene and what is not, and how is the cell able to turn a particular gene on while at the same time not affecting the expression of other genes?

## 1.12 Genes within a Genome

What is a gene? Surprisingly, there is not a universally accepted answer to this question. In its broadest terms a gene may be considered as a unit of heredity composed of nucleic acid. In classical genetics, a gene is described as a discrete part of a chromosome that determines a particular characteristic. Here, we will use the term gene to describe an open reading frame (ORF – the region of the gene that will be transcribed and translated into a protein sequence) together with its transcriptional control elements (promoter and terminator). Thus a gene is not just the coding sequence itself, but contains the information required for the expression of that coding sequence. Most protein coding ORFs have the same overall format (Figure 1.22). They start with a particular triplet of DNA bases (ATG) and end at stop signal, again a triplet of DNA bases (either TGA, TTA or TAG). The sequence in between will either directly code for the protein (prokaryotes) or be split into a series of **exons** and **introns** (eukaryotes). We will discuss exons and intron later, but with the advent of fully sequenced genomes (Chapter 9), a gene is usually defined as a DNA sequence that has at least 100 codons between the start and the stop signals. This sequence alone, is however insufficient to direct protein production. Each gene requires other DNA sequences surrounding it that act as binding sites for proteins such as the RNA polymerase and for those involved in transcriptional and translational initiation and termination. When combined, the coding portions of the gene and these promoter and terminator sequences form the functional gene, sometimes referred to as a 'transcriptional unit'.

How many genes are present within a genome, how many genes are essential to the cell and how many are expressed at one time? The advent of completely sequenced genomes (Chapter 9) has allowed us to address, at least in part, some of these questions.

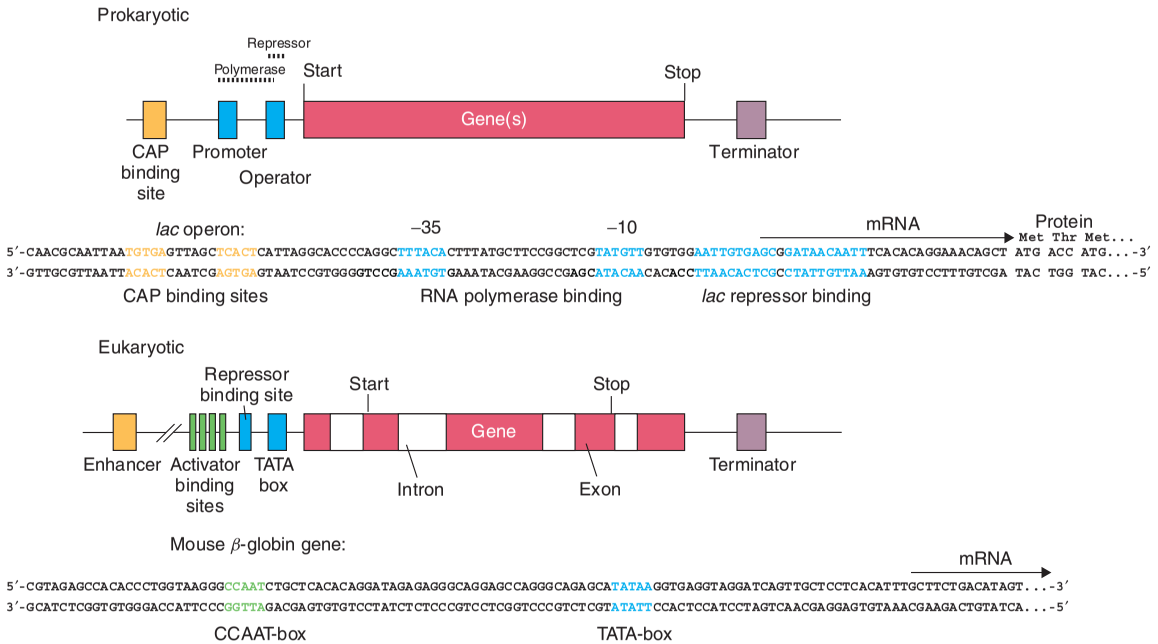



**Figure 1.22.** The architecture of a typical prokaryotic and a typical eukaryotic protein coding gene. In prokaryotes, families of genes required, for instance, to produce all the enzymes of a pathway in prokaryotes are often transcribed together as a polycistronic message. The operon contains the binding sites for the RNA polymerase (promoter) and enhancer elements, such as the cyclic AMP activator protein (CAP). The binding of these proteins triggers transcription. Additionally, the operon may contain the binding sites for repressor proteins, which, when bound to DNA, occlude polymerase binding and thus stop transcription. The transcript ends at terminator sequences at the 3′-end of the operon. Below is shown the sequence of the control region of the *lac* operon. In eukaryotes, single genes are produced and regulated individually. They contain activator and repressor binding sites at the 5′-end of the gene. They also may contain enhancer elements several thousand bace pairs upstream of the gene which are required for its full expression. The activator promotes the assembly of the RNA polymerase II holoenzyme complex at the TATA box, and transcription begins a short distance to the 3′ side of this. Again, termination occurs at specific sites at the 3′-end of the gene. Perhaps the major difference between eukaryotes and prokaryotes is the presence of introns that split the coding sequences (exons). The introns are removed by splicing before the transcript is translated. The DNA sequence of the promoter region of the mouse β-globin gene is also shown

- *How many genes are contained within a genome*? The *E. coli* genome (4.6 Mb in length) contains about 4200 genes. The average length of a gene in *E. coli* is ~950 bp, and the average separation between genes is 118 bp, so the majority of the genome codes for RNA or protein. The yeast *Saccharomyces cerevisiae*, a single-cell eukaryote, has a larger genome size (13.5 Mb) and some 6000 genes. Humans, on the other hand have a genome that is almost 250-fold larger that of yeast (at 3300 Mb) but may have as few as 30 000 genes – representing an increase of only fivefold.

- *How many genes are essential to the organism*? To determine whether a gene is essential for the organism, the function of the gene must be impaired in some way. This is usually achieved by **gene knockouts**, where the gene is removed from the genome and its effects on cell viability determined. Analyses such as these suggest that about half the *E. coli* and yeast genes are required for viability. Many apparently non-essential genes may play specialist roles under conditions not examined in this type of experiment (nutrient starvation for example). The apparently low percentage of essential genes, however, may reflect a functional degeneracy among certain sets of genes.

- *How many genes are expressed at one time*? The genes that are expressed in a particular cell define that cell. Some of the genes expressed in a liver cell must be different to those expressed in a skin cell. Genome-wide analysis of the expression of all genes within fully sequenced organisms such as yeast suggests that about 90% of genes are expressed at any one time, but 80% of these are expressed with very low abundance levels – in the order of 0.1–2 transcripts per cell (Causton *et al*., 2001). The level of expression of individual genes will vary widely. Highly abundant proteins are produced from highly expressed genes, whilst other proteins that may be present at a much lower level (e.g. one copy per cell) are often produced from genes that are expressed at a very low level.

The intermediary between DNA and its encoded protein is RNA. The process of converting DNA sequences into proteins occurs in two steps. The DNA is first **transcribed** into RNA which is subsequently **translated** into the polypeptide sequence of the protein. As we have already seen (Figure 1.5), RNA differs from DNA in that it contains a ribose sugar rather than a deoxyribose sugar. Additionally, uracil replaces thymine as the complementary base to adenine in RNA. Transcription is similar to DNA replication except that only one of the two DNA strands is copied. There are two distinct classes of RNA that are produced by transcription, messenger RNA (mRNA) and structural RNAs

(ribosomal RNA (rRNA) and transfer RNA (tRNA)). mRNA, although by far the least abundant form of RNA – in most cells representing a few percentage points of the total RNA – is the carrier of the genetic information.

## 1.13 Transcription

Transcription is the process by which an RNA copy of one of the strands in the DNA double helix is made. The antisense strand of the DNA directs the synthesis of a complementary RNA molecule. The RNA molecule produced is therefore identical to the sense strand of the DNA – except that it contains U instead of T. There are fundamental differences in the ways in which genes are transcribed in prokaryotes and eukaryotes. Here, it is important to understand the processes involved in each case. Many of the experiments we will look at in later chapters involve the use of eukaryotic cells, but the bacterium *E. coli* still plays a vital role in almost all genetic engineering experiments.

Transcription begins at specific DNA sequences called **promoters**. Like DNA replication, transcription occurs in three phases – initiation, elongation and termination. Initiation of transcription usually occurs to the 3′ side of the promoter, and termination occurs at specific sites downstream of the coding sequence of the gene. At first glance, the overall architecture of a typical prokaryotic gene and a typical eukaryotic gene may appear to be similar (Figure 1.22). However, the controlling region for eukaryotic genes will not function in a prokaryotic cell, and *vice versa*.

Most protein coding genes in prokaryotes are transcriptionally active by default. That is to say, in the absence of other factors, the RNA polymerase can recognize the promoter of a gene, bind to it and produce RNA. Transcriptional control is brought to bear on the gene by repressor proteins that bind to DNA sequences adjacent to the RNA polymerase binding site. DNA binding by the repressor either occludes RNA polymerase binding and/or prevents a bound polymerase from transcribing. The eukaryotic RNA polymerase involved in the production of protein coding genes (pol II) is unable to recognize promoter sequences on its own. Therefore, eukaryotic genes are transcriptionally inactive in the absence of other factors. In both prokaryotes and eukaryotes, transcription is a highly regulated process. Proper timing and levels of gene expression are essential to almost all cellular processes.

### *1.13.1 Transcription in Prokaryotes*

François Jacob and Jacques Monod were the first to elucidate a transcriptionally regulated system (Jacob and Monod, 1961). They worked on the lactose

metabolism system in *E. coli*. Lactose is a disaccharide of galactose and glucose. When the bacterium is in an environment that contains lactose as the sugar source it expresses the following structural genes.

- $\beta$-galactosidase – the enzyme hydrolyses the bond between the two sugars, glucose and galactose. It is coded for by the gene *lac*Z.
- Lactose permease – the enzyme spans the cell membrane and brings lactose into the cell from the outside environment. The membrane is otherwise essentially impermeable to lactose. It is coded for by the gene *lac*Y.
- Thiogalactoside transacetylase – the function of this enzyme is not known. It is coded for by the gene *lac*A.

These three enzymes are located adjacent to each other in the *E. coli* genome (Figure 1.23) and a region of DNA that is responsible for their transcriptional regulation is located just to the 5′ side of these structural genes. This assortment of genes and their regulatory regions is called the ***lac* operon**. In order to metabolize lactose, the structural genes must be expressed and the protein products produced. One could imagine, therefore, a simple **genetic switch** in which the *lac* structural genes are transcriptionally inert in the absence of lactose and on in its presence. There are, however, other factors that influence the transcriptional activity of the *lac* structural genes. The favoured sugar source for *E. coli* is glucose, since it does not have to be modified to enter the respiratory pathway. So, if both glucose and lactose are available, the bacterium turns off lactose metabolism in favour of glucose metabolism. The *lac* operon has the following DNA sequence elements.

(a) Operator (*lac*O) – the binding site for repressor.

(b) Promoter (*lac*P) – the binding site for RNA polymerase.

(c) Repressor (*lac*I) gene – which encodes the Lac repressor protein. This protein binds to DNA at the operator and blocks transcription of the structural genes by RNA polymerase bound at the promoter.

(d) Pi the promoter for *lac*I.

(e) CAP binding site for cAMP/CAP complex.

In the absence of lactose, RNA polymerase binds to the promoter but is prevented from transcribing the structural genes by the *Lac* repressor which is bound to the operator site (Lee and Goldfarb, 1991). Since no RNA can be made, no protein is produced. The repressor itself is produced from the *lac*I

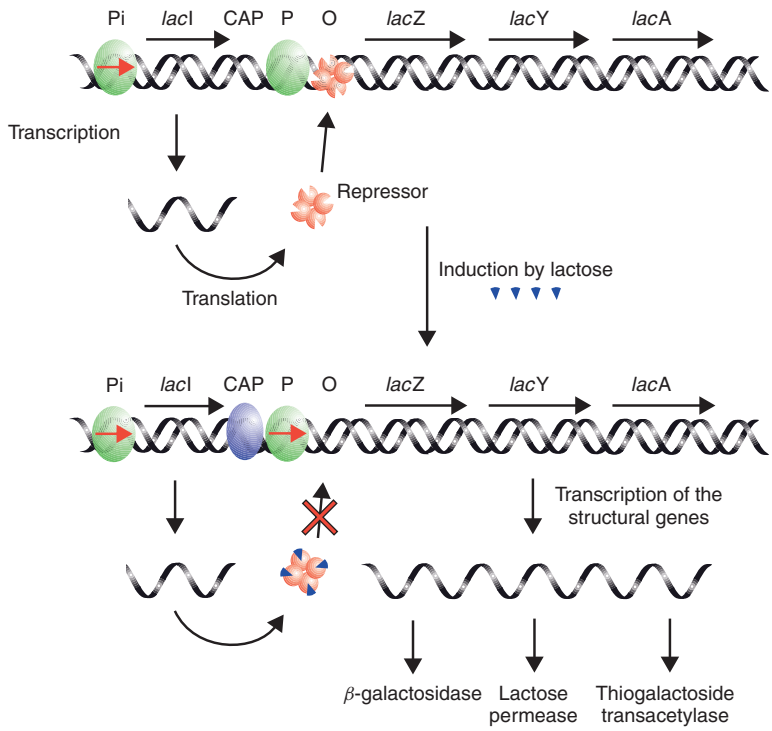



**Figure 1.23.** The activation of the lac operon in *E. coli*. In the absence of lactose, the *lac* repressor binds to the operator (O) and prevents RNA polymerase bound to the promoter (P) from transcribing the structural genes. When lactose enters the cell, it binds to the *lac* repressor and causes a conformational change that inhibits its ability to bind DNA. Consequently, the polymerase bound at P will transcribe the *lac* structural genes (*lac*Z, *lac*A and *lac*Y). Full activation of the *lac* genes only occurs in the absence of glucose when the catabolite activator protein (CAP) binds to the operon and aids the binding of RNA polymerase to P

gene by an RNA polymerase bound to another promoter (Pi). When lactose is present within the cell, it acts as an inducer of the operon. It binds to the *lac* repressor, and induces a conformational change that results in the repressor dissociating from DNA. Now the RNA polymerase bound to the promoter is free to move along the DNA and RNA can be made from the three genes. Lactose can now be metabolized.

A pertinent question to ask here is how the *lac* genetic switch actually gets turned on initially. As stated above, the *lac* structural genes, including the permease required to get lactose into the cell, are switched off in the absence of lactose. It therefore seems impossible to achieve activation, since activation is required to get lactose into the cell. This is, however, an over-simplification. Each of the *lac* structural genes is transcribed at a low, basal, level (approximately five

copies per cell) in the absence of lactose. This means that when the bacterium encounters a source of lactose it is able to transport a few molecules into the cell so that full induction of the *lac* structural genes can occur. When fully induced, approximately 5000 copies of each protein product are present in the cell.

Regulatory control of the *lac* operon is more complicated since it needs to be turned off (repressed) if glucose is present, no matter whether lactose is present or not. The repression of the *lac* operon by glucose is mediated through the CAP site. When levels of glucose (a catabolite) in the cell are high, the production of cyclic AMP (cAMP) is inhibited, so when glucose levels drop, more cAMP forms. cAMP binds to a protein called CAP (catabolite activator protein), which then binds DNA at the CAP site in the *lac* operon. This activates transcription of the *lac* structural genes by increasing the affinity of the promoter for RNA polymerase. This phenomenon is called **catabolite repression**, since glucose inhibits the transcription of the *lac* operon by not allowing its full activation.

Like many operons in prokaryotes, the three *lac* structural genes are transcribed as a single mRNA. This **polycistronic** message is translated as it is produced since there is no physical separation of the DNA from the cytoplasm. Each of the ORFs contains its own ribosome binding site (see below) within the encoded mRNA and is translated independently of the other *lac* genes.

We have seen above how the transcript is initiated, but how does transcription end? Once RNA polymerase has started transcription, the enzyme moves along a DNA template, synthesising RNA, until it meets a **terminator** sequence. Here, the polymerase stops adding nucleotides to the growing RNA chain, releases the completed product and dissociates from the DNA. Termination can be brought about in one of two ways.

- *Rho-dependent termination*. The protein rho binds to the newly synthesized RNA chain and appears to cause termination when the polymerase pauses at certain DNA sequences.
- *Rho-independent termination*. Intrinsic termination sequences may form hair-pin loops in the RNA that promote polymerase dissociation. This type of termination is independent of additional factors and is the most common form of termination in *E. coli*.

### 1.13.2 Transcription in Eukaryotes

The process of activating gene expression in eukaryotes is far more complex than in prokaryotes. There are several differences between eukaryotes and prokaryotes that impinge upon gene activation and RNA processing.

- Eukaryotes possess multiple RNA polymerase enzymes.
- The wrapping of DNA into nucleosomes represents a barrier to transcription.
- The physical separation of the nucleus (where transcription occurs) and the cytoplasm (where translation occurs) in eukaryotes means that mechanisms must exist to protect the mRNAs from degradation in the cytoplasm before they have been translated.
- Eukaryotic genes are **monocistronic**, with each gene being produced as a separate transcript from its own promoter.
- The genes of eukaryotes are not continuous and are split into coding regions and non-coding regions.

We will touch upon each of these issues as we consider how a typical eukaryotic gene may be transcribed, and the subsequent fate of the RNA that is produced.

Eukaryotes contain three different RNA polymerases, each responsible for the transcription for a certain class of genes (Table 1.3). Each of the polymerases is a large complex of proteins that are involved in, amongst other things, regulating polymerase activity. Activation of RNA polymerase II genes (those coding for proteins) is brought about by transcriptional activator proteins binding to DNA sequences within the promoter of a gene (Figure 1.24). In the absence of these activators, the genes are not transcribed. The role of the activator has been the focus of much speculation. However, it appears to act simply as a way of recruiting various protein complexes to a particular promoter. Control of gene expression is therefore targeted through the activator. Activators have at least two separate functions. First, they must be able to bind to DNA so that they can find their target gene within the mass of genomic DNA. Second, they participate in protein–protein contacts necessary to recruit the RNA polymerase and other complexes to the gene through an activation domain. Either of these functions can be modulated to regulate gene expression in response to specific signals.

**Table 1.3.** Properties of eukaryotic RNA polymerases. In older papers, these are sometimes referred to as polymerases A, B and C

| RNA polymerase | Location | Products |
|---|---|---|
| I | Nucleolus | 28S, 18S and 5.8S rRNA |
| II | Nucleus | mRNA, and some snRNA |
| III | Nucleus | tRNA, 5S rRNA, and some snRNA |

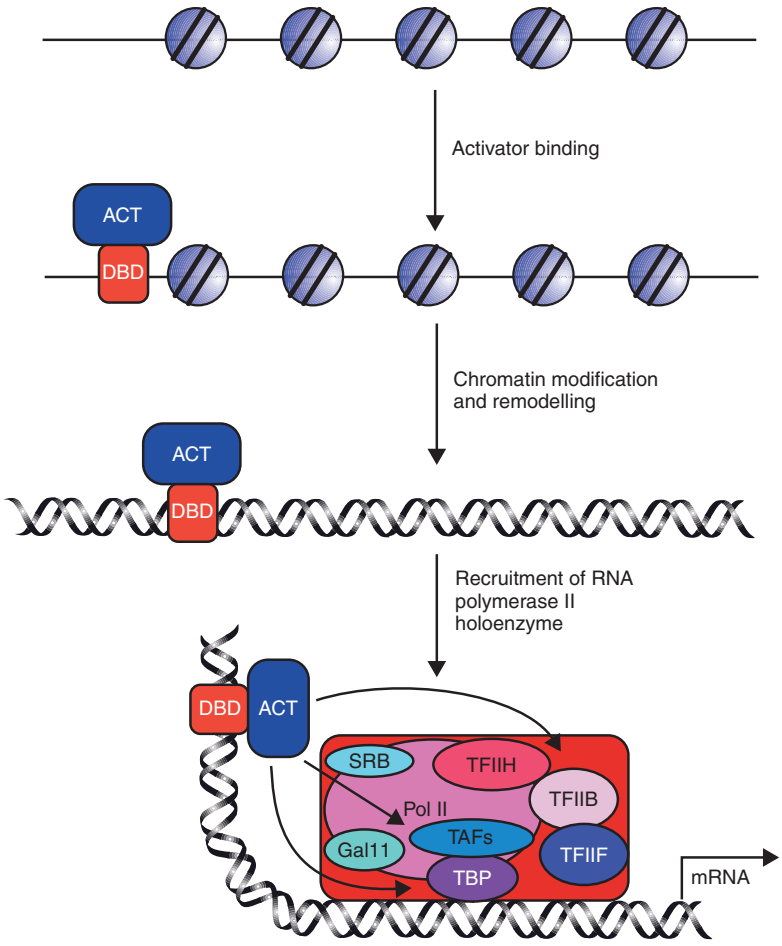



**Figure 1.24.** The multiple roles of transcriptional activators in switching on gene expression in eukaryotes. The binding of the activator protein through its DNA binding domain (DBD) can occur either in a nucleosome-free region, as shown, or on the DNA already bound within a nucleosome. The activation domain of the activator (ACT) recruits chromatin modifying and chromatin remodelling complexes to the gene. The activator then functions to recruit an RNA polymerase II holoenzyme complex to the gene so that transcriptional initiation can occur. The polymerase holoenzyme may exist as one or more discrete sub-complexes within the cell

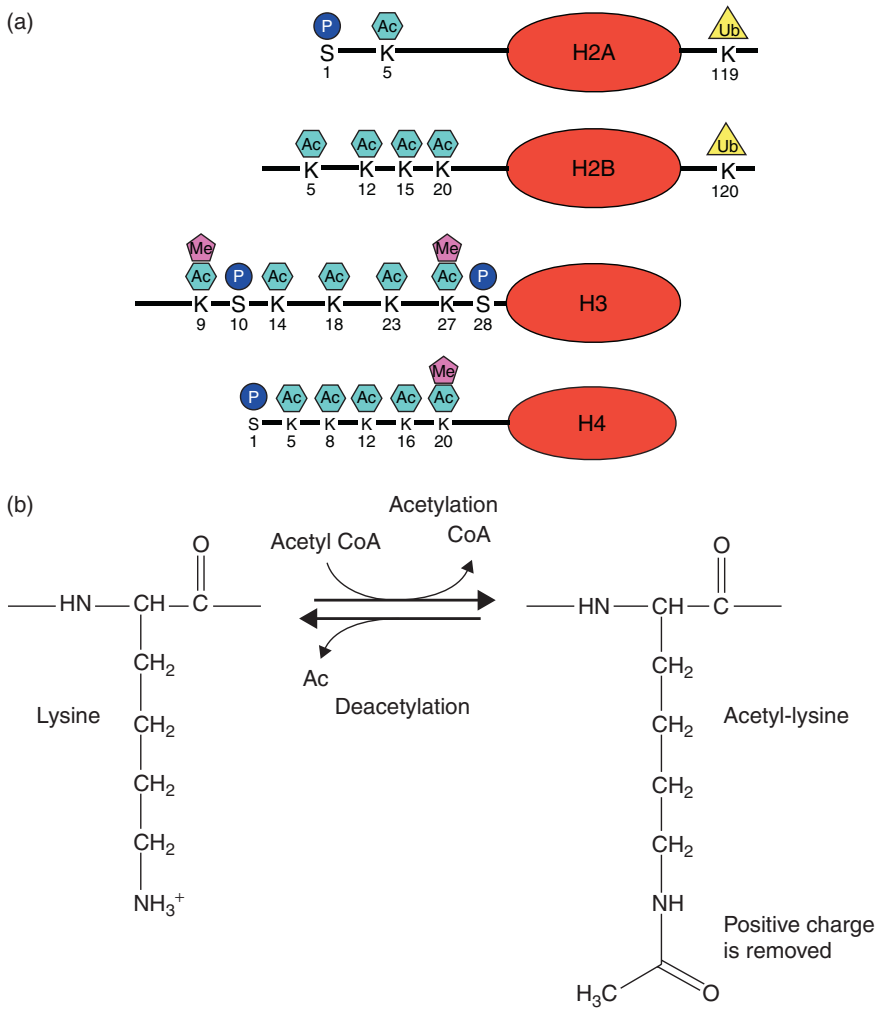
An additional problem for the transcription of eukaryotic genes is that the majority of the DNA is wrapped into nucleosomes, which act as physical barriers to the assembly of the transcriptional complexes. Many activators are able to bind to their cognate DNA recognition sequences whether or not these are contained within nucleosomal DNA (Taylor *et al.*, 1991), while others require that the nucleosomes are removed before they can bind DNA. After

DNA binding, the activator recruits one, or more, of the following complexes to the gene.

- *Chromatin-modifying complexes*. Histone proteins form the core of the nucleosome. However, as can be seen in Figure 1.15, the amino-terminal ends of the histones protrude from the complex. These 'tails' can be modified, by acteylation, phosphorylation or methylation of particular amino acids (Figure 1.25(a)), and the modifications play a significant role in the ability of nearby genes to be transcribed. Although these tails are not needed to maintain the structural integrity of the nucleosome, they do have roles in higher-order chromatin structure and in interactions with non-histone chromosomal proteins. Much attention has been focused recently on the role of histone acteylation in the process of transcription. It has been known for many years that increased levels of histone acetylation at a gene or chromosomal region are associated with transcriptional activity, whereas under-acetylation of histones is observed in non-transcriptionally active regions (Allfrey, Falkner and Mirsky, 1964), but the significance of this observation was not fully realized until it was discovered that transcriptional activators recruit chromatin modifying complexes to the promoters of genes (Bhaumik and Green, 2001; Larschan and Winston, 2001). The acetylation of histones is performed by enzymes called histone acetyltransferases (HATs) while the reverse reaction is catalysed by enzymes call histone deacetylases (H-DACs). The acetylation of lysine residues results in the elimination of positive charge from the protein (Figure 1.25(b)). It has been speculated that in under-acetylated histones (transcriptionally inactive) the histone tails wrap around the DNA with their positive charge being attracted to the negatively charged sugar–phosphate backbone. Acetylated histone tails, on the other hand, may interact with the DNA more loosely so that the histone–DNA complex is less rigid and may be more easily dissociated to allow transcription to occur.

- *Chromatin remodelling complexes*. These complexes, such as the yeast SWI/SNF proteins (pronounced 'switch–sniff', based on the nomenclature used in the original genetic screens that isolated the genes), use energy derived from ATP to move nucleosomes. Their precise mechanism of action is unclear, but they may function as ATP driven motors that move along DNA and disrupt protein–DNA interactions as they move (Pazin and Kadonaga, 1997).

- *RNA polymerase II holoenzyme*. RNA polymerases were first isolated in 1960 (Stevens, 1960; Weiss, 1960). Multiple forms of the enzyme in


(a)
P
Ac
S
K
1
5
H2A
Ub
K
119
Ac
K
5
12
15
20
H2B
120
Me
P
S
9
10
14
18
23
27
28
H3
H4
(b)
Acetylation
CoA
Acetyl CoA
Ac
Deacetylation
HN
CH
C
O
CH2
NH3+
Lysine
NH
H3C
Acetyl-lysine
Positive charge
is removed
(c)
N
HAT
Acetylation
Deacetylation
HDAC

eukaryotic cells were identified by their sensitivity to the bicyclic octapeptide α-amanitin – RNA polymerase I is not affected by this compound; RNA polymerase II is rapidly inhibited at low concentrations of α-amanitin; RNA polymerase III from mammalian cells is inhibited by high levels of α-amanitin. Each of the polymerases is a large (~500 kDa) multi-protein complex typically containing 8–14 individual subunits. Here we will concentrate on RNA polymerase II (pol II) since it is responsible for the transcription of protein coding genes. Pol II is an enzyme comprising 12 subunits encoded, in yeast, by the RPB1 to RBP12 genes. The crystal structure for 10 of the subunits of RNA pol II has been solved recently (Cramer *et al*., 2000; Cramer, Bushnell and Kornberg, 2001) and is shown in Figure 1.26; it gives tantalizing insights into the structural mechanism of RNA production. The largest subunit Rpb1p contains a highly conserved carboxy-terminal repeat domain (CTD) with the consensus sequence Tyr–Ser–Pro–Thr–Ser–Pro–Ser. The yeast CTD includes 26 or 27 repeats and the human CTD has 52 repeats. This sequence can be extensively phosphorylated, which is an important step in converting RNA pol II from a form involved in promoter recognition to a form involved in transcriptional elongation (Hampsey, 1998). We will see later that, besides its role in transcription, the CTD is also critical for co-transcriptional RNA processing events such as capping, polyadenylation and splicing. As stated previously, pol II is incapable of binding DNA on its own. It requires other proteins – called general transcription factors – to achieve this function. There are six principal general transcription factors (TFIIA, IIB, IID, IIE, IIF and IIH) that are essential for gene-specific DNA binding and the formation of a polymerase complex capable of transcriptional initiation. Again, most of the general transcription factors are not single proteins, but are multi-protein complexes – for example TFIID is composed of 13 polypeptides – the TATA-box binding protein (TBP) and 12 TATA-box binding

**Figure 1.25.** Modifications of the histone tails. (a) The amino-terminal 'tails' of the histone proteins may be modified by phosphorylation (P), acetylation (Ac) or methylation (Me) at the serine (S) or lysine (K) residues indicated. The numbers refer to the amino acid positions within each protein. Additionally, histones H2A and H2B can be ubiquitinated (Ub) at their carboxy-terminal ends. The expression of genes contained within or close to nucleosomes containing these modified histones can be drastically altered. Reprinted from *FEMS Microbiology Review*, Vol 23, Pérez-Martín, Chromatin and transcription in *Saccharomyces cerevisiae*, pp. 503–523, Copyright (1999), with permission from Elsevier. (b), (c) The acetylation of the histone tails alters their charge, and results in a 'loosening' of the histone–DNA complex. This allows transcription to occur more readily

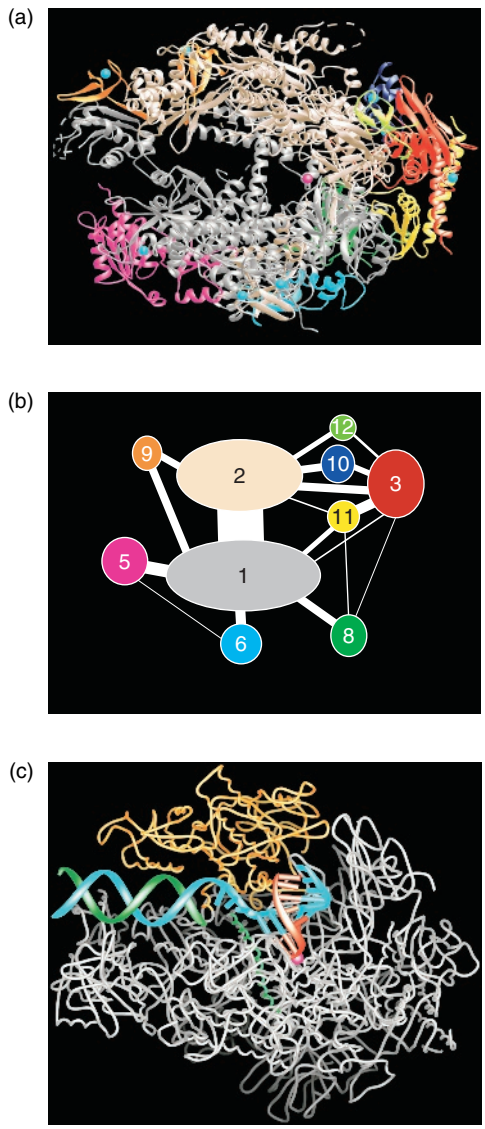



**Figure 1.26.** Architecture of yeast RNA polymerase II. (a) The backbone model for 10 subunits of RNA polymerase as determined by X-ray crystallography. The red sphere in the large cleft at the centre of the molecule represents the magnesium ion at the active site. A diagrammatic arrangement of the subunits of the protein present in the structure is shown in panel (b) and the protein complexed with both double-stranded DNA (blue and green) and RNA (red) is shown in (c). These figures were kindly provided by Patrick Cramer (University of Munich) and are reprinted with permission from *Science* (Cramer *et al.*, 2000; Gnatt *et al.*, 2001). Copyright American Association for the Advancement of Science

associated factors (TAFs) (Lee and Young, 2000). Pol II exists in the cell as a large complex of proteins, termed the holoenzyme. The holoenzyme is composed of SRB (suppressor of RNA polymerase B) proteins that interact with the CTD, Med (mediator) proteins and some of the general transcription factors. The precise composition of the general transcription factors involved with the holoenzyme is still unclear. The SRB and Med proteins appear to act as a regulatory 'glue' that stabilizes interactions between RNA pol II and general transcription factors and may also confer responsiveness to transcriptional activators (Koleske and Young, 1994; Malik and Roeder, 2000). The mediator complex is assumed to enter the initiation complex with RNA pol II and is released at the end of initiation or early in RNA chain elongation (Svejstrup *et al*., 1997). Transcriptional activators interact with a number of the components of the Pol II holoenzyme, including TBP, TAFs, TFIIB and some of the SRB and Med proteins. Whatever interactions occur within the cell, the end result appears to be simply the recruitment of the pol II complex to a promoter that has been freed from nucleosomes. Once the polymerase is bound to the promoter, transcription will be initiated.

The interactions between activators and the proteins described above lead to a model for gene activation in which the activator must first bind to DNA. The activator then recruits chromatin modifying and remodelling complexes to remove nucleosomes from the promoter, and finally pol II is brought to the promoter so that transcription can begin. The entire gene need not be devoid of nucleosomes for full transcription to occur. If the promoter is accessible to RNA polymerase and transcription factors, the presence of nucleosomes will not inhibit the elongation of the message. The structure of the nucleosome is such that the positive charges (lysines and arginines) of the histones are placed adjacent to the negative DNA helix. As the polymerase advances, it displaces DNA from the nucleosome and forms a closed loop, while the torsion ahead of the RNA polymerase generates supercoiling of the DNA. This displaces the histone octamer, which keeps contact with the DNA behind the RNA polymerase. In this way, the nucleosome need never lose contact with DNA as the RNA polymerase passes by (Studitsky, Clark and Felgerfeld, 1995). So once the RNA polymerase is transcribing the gene, nucleosomes do not stop it. The important event for transcription-level gene expression is to get the RNA polymerase bound and functioning at the promoter.

The molecular processes involved in transcriptional termination in eukaryotes are relatively poorly defined. Some transcripts terminate over 1000 bp downstream of the 3′ end of the mature mRNA, and appear to end at termination regions rather than at specific sites. The transcript is cleaved from the extending polymerase before the polymerase itself terminates transcription

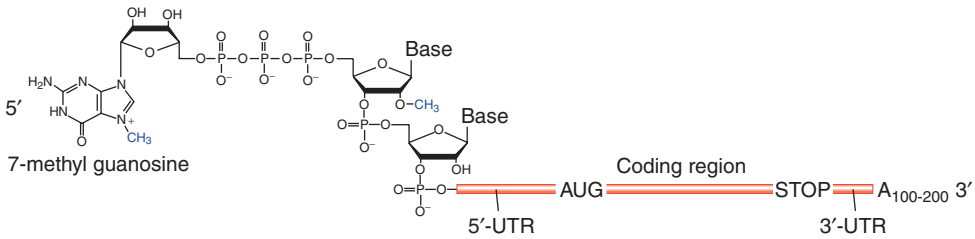



**Figure 1.27.** The structure of a mature mRNA molecule in eukaryotes. After transcription, a 7-methyl guanosine cap is added to the 5′-end of the message, and the ribose sugar of the first, and sometimes second, nucleotide is methylated at the 2′-position. The 3′-end of the transcript is polyadenylated with the addition of 100–200 A residues

(Figure 1.27). The function of the RNA polymerase complex is not finished once the transcript has been cleaved. The polymerase also functions to direct processing of the transcript, such as splicing and polyadenylation (McCracken *et al.*, 1997; Hirose and Manley, 1998). Indeed, recent evidence suggests the enzymes and cellular machinery involved in transcription, RNA processing and translation may be extensively coupled to form 'gene expression factories' that maximize the efficiency and specificity of each stage of gene expression (Maniatis and Reed, 2002).

## 1.14 RNA Processing

Perhaps the most obvious difference between the genes of prokaryotes and eukaryotes is that genes of the later are split into **exons** (coding regions) and **introns** (non-coding regions). In the mid-1970s, Philip Sharp and Richard Roberts independently hybridized a messenger RNA to the DNA from which it had been transcribed, and viewed the resulting hybrids using electron microscopy. They found that RNA hybridization occurred at discontinuous sequences within the DNA. That is, the binding of mRNA to the DNA sequence from which it was derived would result in the formation of loops in the DNA, corresponding to sequences in the DNA that were not present in the mRNA molecule (Chow *et al.*, 1977; Berget, Moore and Sharp, 1977). Walter Gilbert gave names to these various regions. The parts of the gene that are represented in the mRNA, and are therefore part of the expressed region, he called exons. The parts of the gene that are the intervening sequences he called introns. The mechanism by which the precursor RNA (pre-mRNA) is converted into the final mRNA – in other words, how introns are removed – has been the subject of intense scrutiny since their discovery. Before we address this issue, we need to see what happens to an mRNA molecule as it is produced.

Once a gene has been transcribed, the mRNA produced is extensively modified. Almost as soon as mRNA synthesis is initiated, the 5′-end of the message is capped by the addition of a 7-methyl guanosine residue. The cap is added to the 5′-end of the mRNA *via* a 5′–5′ condensation of guanosine to the mRNA to form a structure consisting of 3′–G–5′ppp5′–N-3′p. After cap formation, a methyl group is added to the guanine residue and to the first and/or second adjacent nucleotide (Figure 1.27). The function of the cap is not entirely clear. It has been suggested that the cap, and proteins that bind to it, direct ribosome binding and correct translational initiation. In addition to the cap, the 3′-end of the mRNA is cleaved from the extending chain and then polyadenylated.

RNA transcript cleavage and polyadenylation are directed by a polyA signal within the RNA (Figure 1.28). The core polyA signal for vertebrate pre-mRNAs consists of two recognition elements flanking a cleavage–polyadenylation site. Typically, an almost invariant 5′-AAUAAA-3′ hexamer is found 20–50 nucleotides upstream of a more variable element rich in U or GU residues. Cleavage of the newly formed transcript occurs between these two elements and is coupled to the addition of approximately 200 adenosines to the 3′-end of the 5′ cleavage product. Two protein factors required for this process are the cleavage and polyadenylation specificity factor (CPSF), which binds to the AAUAAA motif, and the cleavage stimulation factor (CStF), which binds the downstream GU-rich element. As the polyA signal is extruded from the polymerase, it is recognized by a subset of cleavage and polyadenylation factors that have already been recruited to the RNA polymerase CTD. Then assembly continues as more factors join the complex. Ultimately cleavage occurs, followed by polyadenylation in which a stretch of ~200 A residues is added to the 3′-end of the message. Polyadenylation only occurs to mRNAs, but not all mRNAs are polyadenylated; e.g. the mRNA species that encode histone proteins are not polyadenylated. The process of polyadenylation is carried out by an enzyme called polyA polymerase (PAP), and not only confers stability on the transcript, but is also required for the movement of the mature mRNA out of the nucleus into the cytoplasm, where translation occurs. As we will see in later chapters, the addition of polyA to mRNAs has important practical consequences. The hybridization of polyT sequences to only mRNA, and not other RNAs, is a vital tool to the genetic engineer.

### 1.14.1 RNA Splicing

There are several different mechanisms of RNA splicing, which function on different RNA species. Here, however, we will concentrate on the mechanism

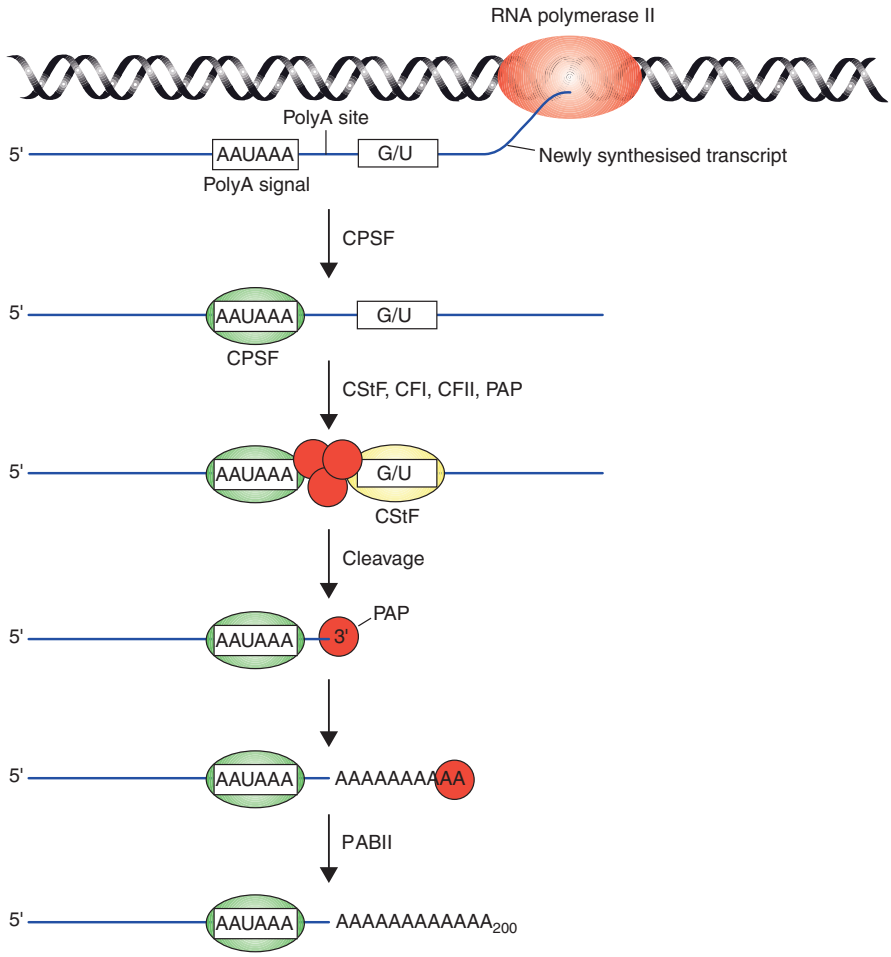



**Figure 1.28.** Cleavage and polyadenylation of the mRNA transcript. The polyA site within the newly formed transcript is shown, together with the signal AAUAAA and the downstream U/GU region. A complex of cleavage and polyadenylation specificity factor (CPSF), cleavage factors I and II (CFI and II) and cleavage stimulation factor (CStF) bind to these sequences. PolyA polymerase (PAP) also joins the complex at this stage. RNA cleavage occurs, and CPSF remains bound to the 5′-end of the cleaved message. The 3′-end is rapidly degraded. PAP begins the synthesis of the polyA tail, resulting in the addition of the first 10 A residues. Finally, polyA binding protein II (PABII) joins the reaction, stimulating the synthesis of polyA and extending the tail to about 200 A residues

of mRNA splicing since understanding it, and its consequences, are most applicable to the genetic engineer. The capped and polyadenylated mRNA is **spliced** to remove the introns and fuse the exons together into a single unit that can be translated. The need to produce precisely spliced products is obvious. Any slippage in the fusing of exons together would have disastrous

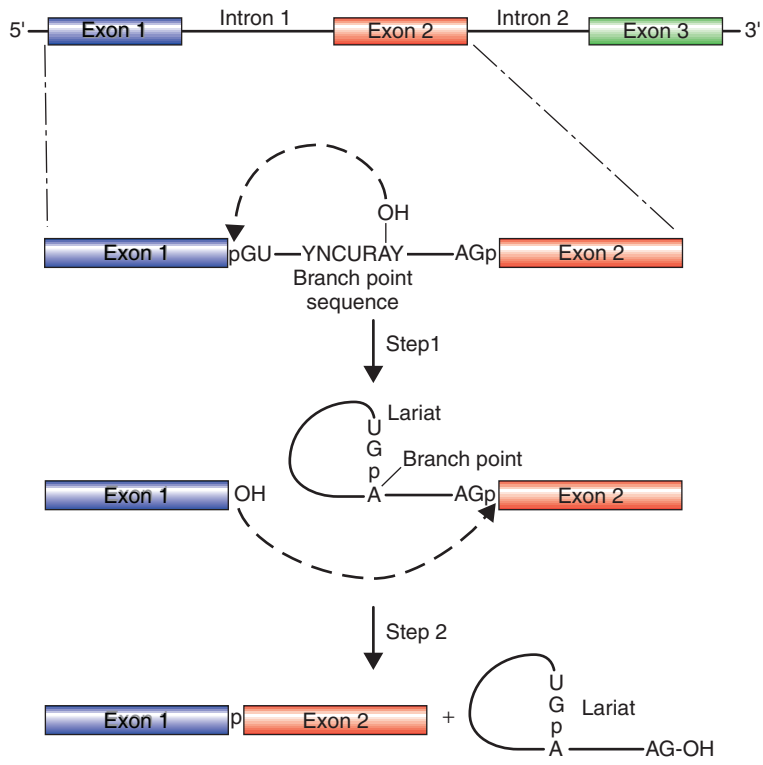



**Figure 1.29.** The splicing of exons to form a mature mRNA. Nuclear pre-mRNA splicing takes place in two distinct mechanistic steps. Step 1 involves a nucleophilic attack by the 2′-OH group of the branch point A residue on the phosphodiester bond of the 5′ exon–intron boundary, displacing the 5′ exon (shown in blue) as the leaving group and giving the lariat structure shown. The second step is a nucleophilic attack of the 3′-end of the 5′ exon on the phosphodiester bond of the 3′ intron–exon boundary, displacing the intron (as a lariat form) and sealing the two exons together (the blue and red)

consequences for the protein that is to be eventually made. How are the boundaries between an intron and an exon marked, and by what molecular mechanism does splicing actually occur? Perhaps surprisingly, there is relatively little RNA sequence conservation at the exon–intron boundary. Most introns begin with the dinucleotide sequence 5′-GU-3′ and end with the dinucleotide sequence 5′-AG-3′ (Figure 1.29). There are, however, other less conserved sequences present within the intron that act as binding sites for complexes that are essential for splicing. These complexes, collectively called the **spliceosome**, contain both RNA and protein components, and are found exclusively in the nucleus. The RNA molecules found within the spliceosome are small (100–200 nucleotides) and are complexed with proteins to form small nuclear

ribonucleoproteins (snRNPs, or snurps). The RNA molecules within these complexes are rich in uridine residues, and consequently the snRNPs have been designated by the names U1, U2, U4, U5 and U6. A model of the overall splicing process catalysed by the spliceosome is shown in Figure 1.29. The mechanism of splicing for exon coupling and intron extrusion depends on two transesterification reactions, which result in the formation of a **lariat** form of the intron and the fused exons. The branching associated with the A residue within the intron in the lariat form contains two phosphodiester linkages to the 2′ and 3′ of the A. The snRNPs play roles both in the catalytic process and in structurally maintaining the two exons in proximity with each other as the splicing reaction proceeds. Readers interested in the mechanism of splicing are directed to specific reviews on the topic (Sharp, 1994).

### 1.14.2 Alternative Splicing

Since each exon in a eukaryotic gene encodes a portion of a protein, you can imagine that it is possible, by altering how the pre-mRNA is spliced, to produce different versions of the mRNA and ultimately, different proteins. Alternative splicing is a widely occurring phenomenon, with recent estimates suggesting that at least 30% of all human genes are subject to this type of processing (Sorek and Amitai, 2001). How do splice variants differ from the original sequence from which they are derived? The physiological activity of proteins produced from splice variants may be the same, opposite or completely different and unrelated. Alternative splicing therefore greatly increases the number of different protein activities that can be generated from a defined set of genes within the genome.

Perhaps one of the most dramatic examples of alternative splicing is found in the genes determining sex in the fruit fly *Drosophila melanogaster*. There are three genes that are involved in the sex determination process: *sxl* (sex lethal), *tra* (transformer) and *dsx* (double sex) (Chabot, 1996). Each of these genes produces a pre-mRNA that has two possible splicing patterns, depending upon whether the fly is male or female. Figure 1.30 shows these three genes and their splicing patterns. For males, the inclusion of two exons (exon3 in *sxl* and exon2 in *tra*) produces mRNA molecules that have termination (stop) codons and results in the formation of inactive proteins. The only active male product is the protein translated from the *dsx* gene. This protein inactivates all female-specific genes. The female, on the other hand, produces mRNAs without the stop codon containing exons. In the case of *sxl* and *tra*, the protein products have a positive effect on the splicing patterns observed, controlling the choice of introns removed in the spliceosome reaction.

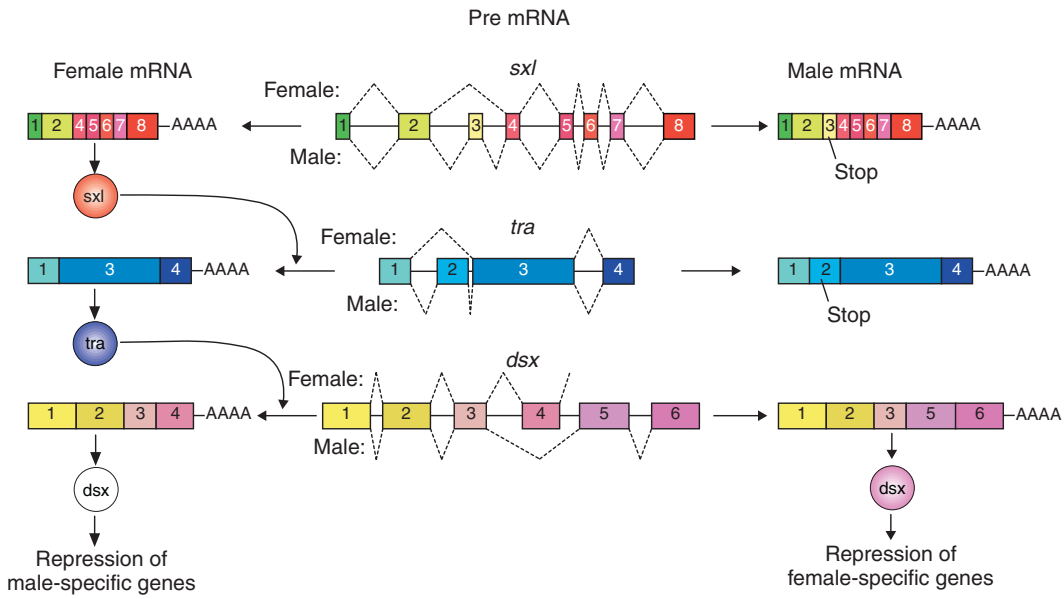



**Figure 1.30.** Alternative splicing during sex determination in *Drosophila*. In the centre of the diagram are the pre-mRNAs for three genes involved in the sex determination process (*sxl, tra* and *dsx*), with the splicing pattern for the female (top) and the male (bottom). The product mRNAs are shown to either side (female on the left, male on the right). In the female, the protein products of the *sxl* and *tra* mRNA control splicing site selection. These proteins are not produced in the male due to the inclusion of exons containing stop codons. Males, however, produce the *dsx* protein, which represses the transcription of other female-specific genes. The female version of the *dsx* protein represses the transcription of male-specific genes

## 1.15 Translation

Translation is the process whereby the structural RNAs, and their associated proteins, decode the linear sequence of information contained within the mRNA to produce linear chains of amino acids, called **polypeptides**, that make up proteins. The nucleotide sequence of the mRNA is read as a series of triplets (a **codon**). Each codon specifies the insertion of a single amino acid into the growing peptide chain. Translation begins at an initiation codon, AUG, and ends at one of three different termination codons, UAA, UGA or UAG (see Appendix 1). Each triplet is read in turn from the initiation codon onwards (Figure 1.31). There are 20 different amino acids found within proteins. The triplets of the genetic code, however, provide up to 64 possible different codons (Appendix 1). Thus, there is degeneracy in the genetic code, with most amino acids being encoded by more than one codon. For example, methionine is




**Figure 1.31.** The transcription and translation of a DNA sequence results in the formation of the encoded protein. Mutations in the DNA sequence can have dramatic effects on the protein produced. Shown are the effects of a point mutation on, and the insertion of a single nucleotide into, the coding region of a gene. The resulting changes in the sequence of the resulting protein are shown

encoded by a single codon (AUG), while leucine and arginine are encoded by six codons each. The genetic code is almost universal. The vast majority of genes from all organisms obey the same code; however, some mitochondrial encoded genes deviate from the code. For example, human and yeast mitochondria use the UGA termination codon to insert a tryptophan amino acid into an extending polypeptide chain.

Alterations in the DNA sequence can drastically affect the resulting protein sequence. In the example shown in Figure 1.31, mutation of the DNA sequence altering a sense strand T residue to a G results in the change of a serine codon (UCU) to an alanine codon (GCU). The insertion, or deletion, of nucleotides has an even more dramatic effect on the encoded protein. In the example shown, the insertion of a GC base pair alters the reading frame of the transcript so that the resulting protein is completely different from the original sequence after the point of insertion. Mutations like this are referred to as **frame-shift** mutations. Not all DNA mutations will result in changes to the protein sequence. For example, if the serine codon discussed above had been changed from UCU to UCG then this new codon would still encode serine. Mutations such as these are termed **silent**, as they do not alter the protein sequence.

The tRNAs act as intermediaries in the translation process. They transport a specific amino acid to the mRNA triplet it encodes. Enzymes

called aminoacyl-tRNA synthetases couple the amino acid encoded by a particular codon to the tRNA that contains the appropriate **anticodon**. The codon–anticodon pairing therefore directs the addition of the correct amino acid to the growing polypeptide chain. The mRNA is translated upon **ribosomes**. These bipartite structures consist of a large and a small subunit. Each subunit contains structural RNAs (the rRNAs) complexed with the proteins that perform the various chemical reactions involved in translation. The mechanism of translation can, like DNA replication and transcription, be split into three steps – initiation, elongation and termination. In bacteria, initiation begins when the ribosome assembles on the purine-rich **Shine–Dalgarno sequence** (5′-AGGAGGU-3′) 4–8 bases upstream of the initiator codon (Shine and Dalgarno, 1975). This sequence is complementary to the 3′-end of the 16S rRNA (5′-ACCUCCU-3′) of the 30S ribosome subunit, and positions the ribosome to initiate translation. In eukaryotes, ribosome assembly occurs at the **Kozak sequence** (optimal consensus 5′-ACCAUGG-3′) that surrounds the initiator codon (underlined) (Kozak, 1986). The ribosome then moves along the mRNA in a 5′ to 3′ direction (Figure 1.32). Translation starts at the AUG initiation codon and continues along the mRNA until the ribosome dissociates when it reaches the termination codon. No tRNAs exist for the terminator codons, and instead RF (release factor) protein binds to dissociate the ribosome. Under optimal conditions about 15–20 amino acids can be polymerized per second (Crowlesmith and Gamon, 1982). However, the actual rate of translation may be considerably higher than this since more than one ribosome may be bound to single mRNA at one time, forming a **polysome**. The number of ribosomes bound to a particular mRNA will influence the rate of its translation: the higher the number of ribosomes bound, then the higher will be the number of protein chains produced. The **codon usage** of an individual mRNA can also significantly influence the rate of elongation. The tRNA molecules for certain codons are more abundant than others, leading to a discrepancy in codon usage for highly expressed genes.

Translation has proven to be particularly important to the genetic engineer since many antibiotics target translation as their mechanism of inhibiting bacterial growth. The following antibiotics function by the inhibition of translation:

- *chloramphenicol* – inhibits peptidyl transferase on the 50S ribosomal subunit.
- *erythromycin* – inhibits translocation by 50S ribosomal subunit.
- *fusidic acid* – inhibits translocation by preventing the dissociation of an elongation factor from the ribosome.

- *puromycin* – an aminoacyl-tRNA analogue that causes premature chain termination.
- *streptomycin* – causes mRNA misreading and inhibits chain initiation
- *tetracycline* – inhibits binding of aminoacyl-tRNA to ribosomal A-site.

The protein chain contains information that directs post-translational processes and cellular compartmentalization in eukaryotes. All protein translation begins with a methionine residue at the amino-terminal end of the newly synthesized polypeptide (coded by the AUG codon). Many proteins, in both prokaryotes and eukaryotes, have this residue removed by an enzyme called **methionine aminopeptidase** such that the final protein sequence begins with the amino acid encoded by the second codon (Bradshaw, Briday and Walter, 1998). Soluble cytosolic proteins are simply released from the ribosome after polypeptide synthesis is complete, and are already in the correct location to undertake

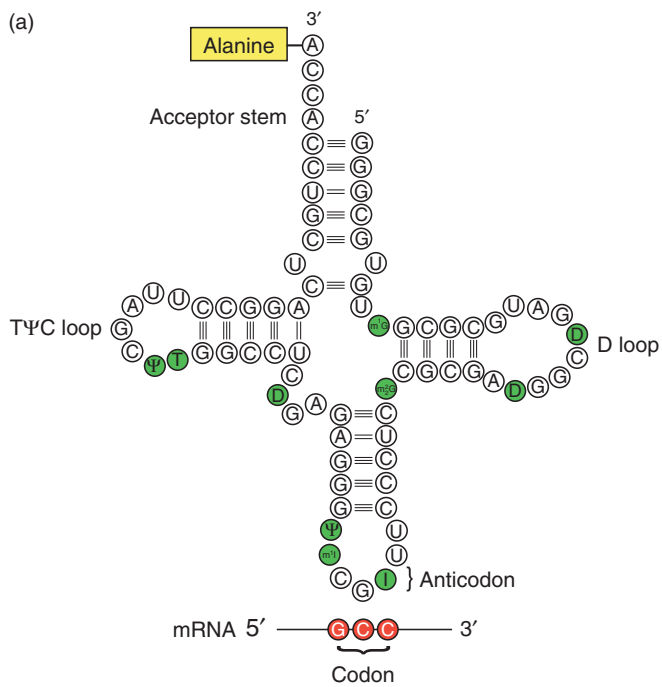



**Figure 1.32.** (a) Cloverleaf structure of yeast alanine tRNA showing the location of the unusual bases (shaded green). tRNA molecules can contain the following unusual bases: pseudouridine (Ψ), inosine (I), dihydrouridine (D), ribothymidine (T), methylguanosine ($m^1G$), dimethylguanosine ($m^2_2G$) and methylinosine ($m^1I$). The tRNA anticodon base pairs with the appropriate codons in mRNA

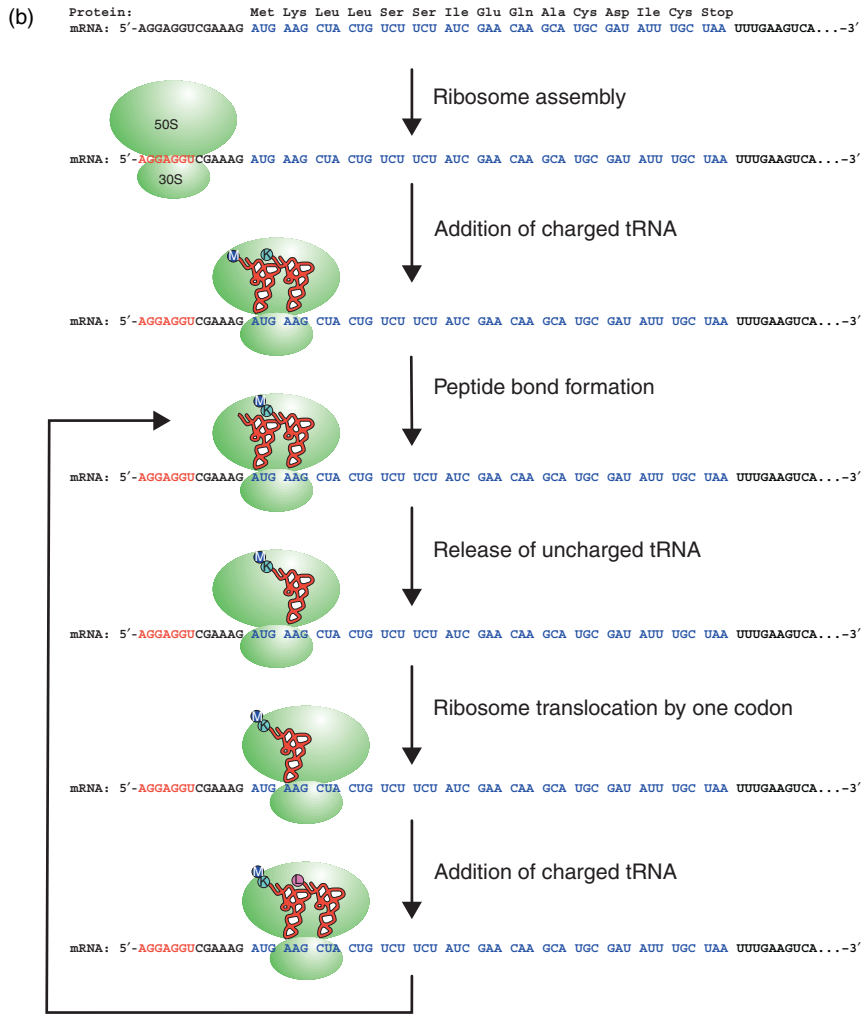



**Figure 1.32.** (b) Translation of an mRNA molecule on a ribosome in bacteria. The ribosome assembles on the mRNA just upstream of the initiation codon (AUG) at the Shine–Dalgarno sequence (shown in red). The 30S ribosomal subunit binds directly to this sequence and then recruits the 50S subunit. The assembled 70S ribosome is a large complex of proteins and rRNA that is of a physical size to cover approximately 35 bases of mRNA. A special initiator tRNA (fMet-tRNA$_f$) binds to the initiator codon (in the P site) through base pairing of the codon with the anticodon present in the tRNA. The next aminoacyl-tRNA (charged with its appropriate amino acid) binds to the next codon (in the A site) prior to peptide bond formation. The formation of a peptide bond involves the transfer of the amino acid(s) attached to the tRNA in the P site to the aminoacyl-tRNA in the A site. The uncharged tRNA then leaves the complex and the ribosome moves to the next codon. Subsequent amino acids are added in a similar fashion

their specific function. Many eukaryotic proteins are, however, destined for a particular compartment of the cell. The polypeptide itself encodes the information required for its final destination. Eukaryotic proteins that are destined to accumulate within the nucleus contain a nuclear localization signal (NLS), consisting of a short stretch of predominately basic amino acids (e.g. PKKKRLV), which direct the protein into the nucleus (Goldfarb *et al*., 1986). Proteins that are destined to be exported from the cell, or to be incorporated in cellular membranes, are transported into the endoplasmic reticulum (ER) during translation. Such proteins containing a stretch of predominantly hydrophobic amino acids at their amino terminus, called the **signal sequence,** which directs the polypeptide–ribosome complex to the ER, where the newly synthesized amino acid chain is inserted through the ER membrane. Inside the ER, the signal is cleaved off as the growing peptide chain is secreted into the ER. The protein is then transported by the Golgi apparatus to its final destination or, if it contains an ER retention signal (the amino acids KDEL) at its extreme carboxy-terminal end, it is retained within the ER itself (Munro and Pelham, 1987).