1 Nucleic Acid Structure and Function

In this book it is assumed that you will already have a working knowledge of the essentials of molecular biology, especially the structure and synthesis of nucleic acids and proteins. The purpose of this chapter therefore is to serve as a reminder of some of the most relevant points, and to highlight those features that are particularly essential for an understanding of later chapters.

1.1 Structure of nucleic acids

1.1.1 DNA

In bacteria, the genetic material is double-stranded DNA, although bacteriophages (viruses that infect bacteria; see Chapter 4) may have double-stranded or single-stranded DNA, or RNA. The components of DNA (Figure 1.1) are 2′-deoxyribose (forming a backbone in which they are linked by phosphate residues) and four heterocyclic bases: two purines (adenine, A, and guanine, G) and two pyrimidines (thymine, T, and cytosine, C). The sugar residues are linked by phosphodiester bonds between the 5′ position of one deoxyribose and the 3′ position of the next (Figure 1.2), while one of the four bases is attached to the 1′ position of each deoxyribose. It is the sequence of these four bases that carries the genetic information.

The two strands are twisted around each other in the now familiar double helix, with the bases in the centre and the sugar-phosphate backbone on the outside. The two strands are linked by hydrogen bonds between the bases. The only arrangement of these bases that is consistent with maintaining the helix in its correct conformation is when adenine is paired with thymine and guanine with cytosine. One strand therefore consists of an image of the other; the two strands are said to be *complementary*. Note that the purines are larger than the pyrimidines, and that this arrangement involves one purine opposite a pyrimidine at each position, so the distance separating the strands remains constant.

Molecular Genetics of Bacteria, Fifth Edition, Jeremy W. Dale and Simon F. Park
© 2010 John Wiley & Sons, Ltd

COPYRIGHTED MATERIAL

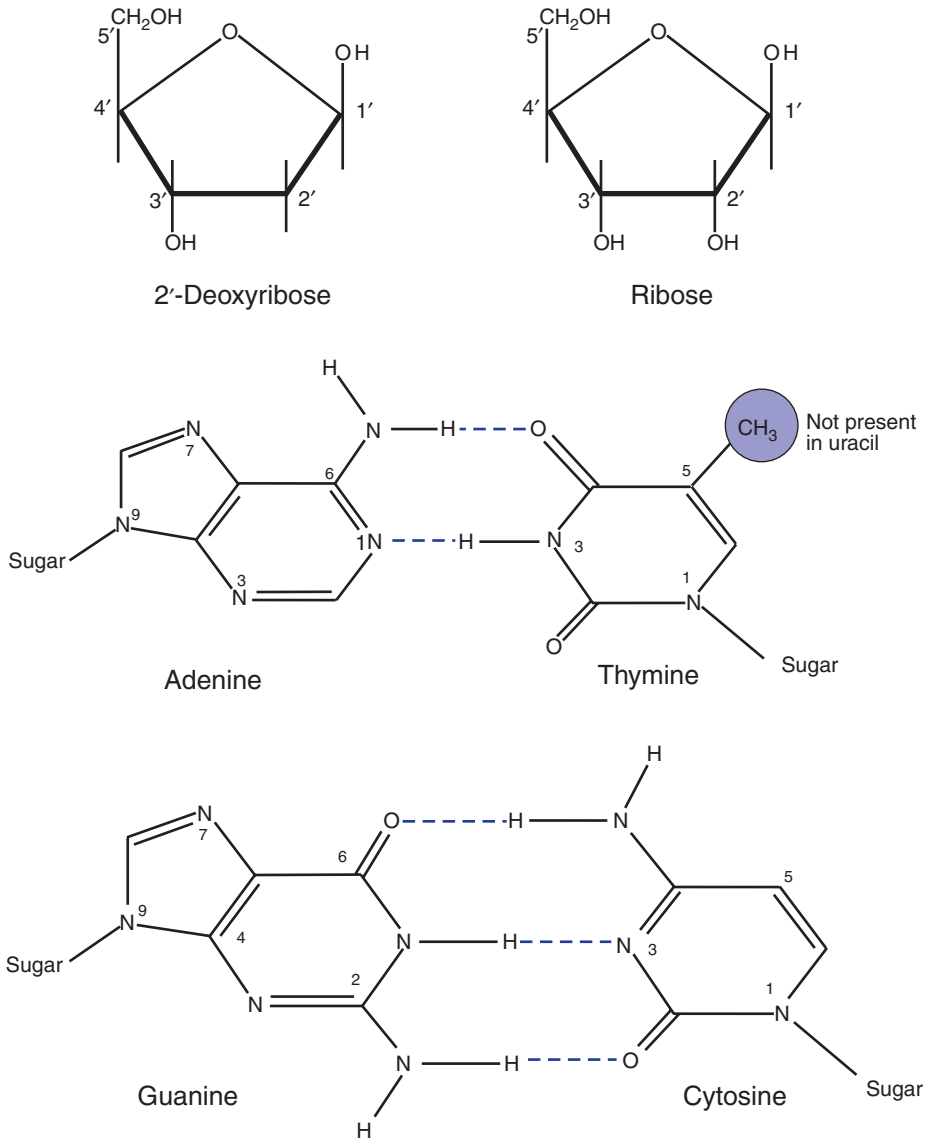


Figure 1.1 Structure of the basic elements of DNA and RNA. RNA contains ribose rather than deoxyribose, and uracil instead of thymine.

1.1.2 RNA

The structure of RNA differs from that of DNA in that it contains the sugar ribose instead of deoxyribose, and uracil instead of thymine (Figure 1.1). It is usually described as single-stranded, but only because the complementary

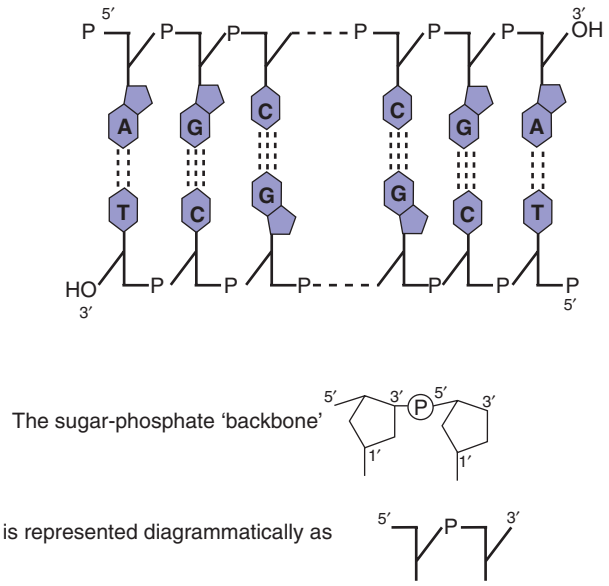


Figure 1.2 Diagrammatic structure of DNA.

strand is not normally made. There is nothing inherent in the structure of RNA that prevents it forming a double-stranded structure: an RNA strand will pair with (hybridize to) a complementary RNA strand, or with a complementary strand of DNA. Even a single strand of RNA will fold back on itself to form double-stranded regions. In particular, transfer RNA (tRNA), and ribosomal RNA (rRNA) both form complex patterns of base-paired regions. The formation of secondary and tertiary structures in RNA via base-pairing can also influence gene expression and this is considered in further detail in Chapter 3.

1.1.3 Hydrophobic interactions

Although geneticists emphasize the importance of the hydrogen bonding between the two DNA strands, these are not the only forces influencing the structure of the DNA. The bases themselves are hydrophobic, and will tend to form structures in which they are removed from the aqueous environment. This is partially achieved by stacking the bases on top of one another (Figure 1.3). The double-stranded structure is stabilized by additional hydrophobic interactions between the bases on the two strands. The hydrogen bonding not only holds the two strands together but also allows the corresponding bases to approach sufficiently closely for the hydrophobic forces to operate. The hydrogen bonding of the bases is, however, of special

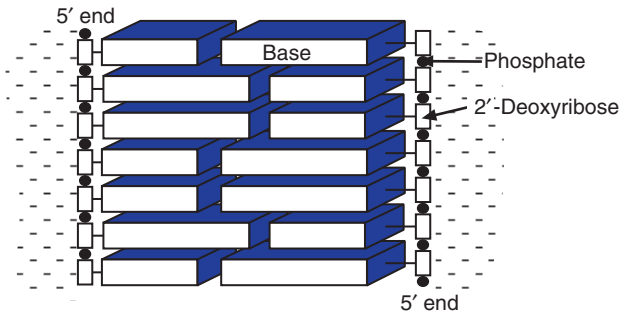


Figure 1.3 Hydrophobic interactions of bases in DNA. The hydrophobic bases stack in the centre of the helix, reducing their contact with water.

importance because it gives rise to the specificity of the base-pairing between the two chains.

Although the bases are hydrophobic, and therefore very poorly soluble in water, nucleic acids are quite soluble, due largely to the hydrophilic nature of the backbone, and especially the high concentration of negatively charged phosphate groups. This will also tend to favour a double-helical structure, in which the hydrophobic bases are in the centre, shielded from the water, and the hydrophilic phosphate groups are exposed.

1.1.4 Different forms of the double helix

A full consideration of DNA structure would be extremely complex, and would have to take into account interactions with the surrounding water itself, as well as the influence of other solutes or solvents. The structure of DNA can therefore vary to some extent according to the conditions. *In vitro*, two main forms are found. The Watson and Crick structure refers to the B form, which is a right-handed helix with 10 base-pairs per turn (Figure 1.4). Under certain conditions, isolated DNA can adopt an alternative form known as the A form, which is also a right-handed helix, but more compact, with about 11 base pairs (bp) per turn. Within the cell, DNA resembles the B form more closely, but has about 10.4 bp per turn (it is *underwound*; see below).

Certain DNA sequences, notably those containing alternating G and C residues, tend to form a left-handed helix, known as the Z form (since the sugar-phosphate backbone has a zigzag structure rather than the regular curve shown in the B form). Although Z DNA was originally demonstrated using synthetic oligonucleotides, naturally occurring DNA within the cell can adopt a left-handed structure, at least over a short distance or temporarily. The switch from left- to right-handed can have important influences on the expression of genes in that region.

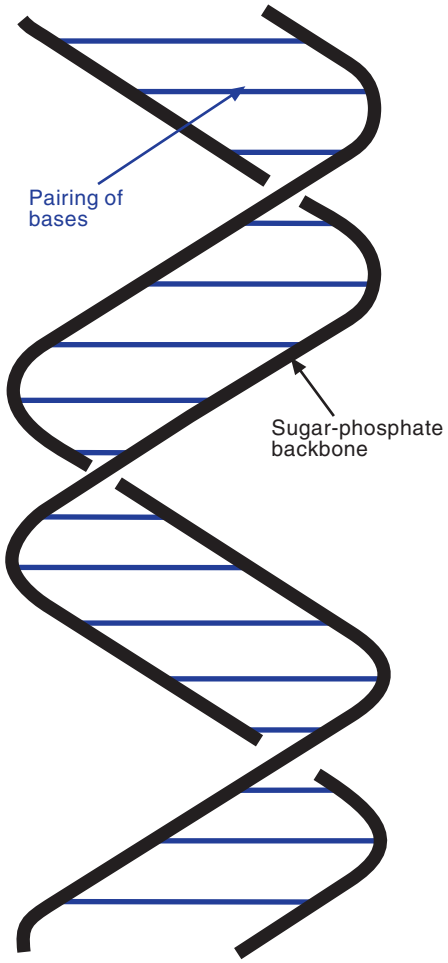


Figure 1.4 Diagrammatic structure of B-form DNA. The two anti-parallel sugar-phosphate chains form a right-handed helix, with the bases in the centre, held together by hydrophobic interactions and hydrogen bonding.

1.1.5 Supercoiling

Within the cell, the DNA helix is wound up into coils; this is known as supercoiling. Figure 1.5 shows a simple demonstration of supercoiling, which you can easily try out for yourself. Take a strip of paper, and twist one end to introduce one complete turn (i.e. the same side of the paper is facing you at each end). It will now look as in Figure 1.5a. Then bring the two ends towards each other; the conformation will change to that shown in Figure 1.5b, which is a simple form of supercoiling. Notice that not only has the strip of paper become supercoiled, but also the degree of twisting seems to have changed

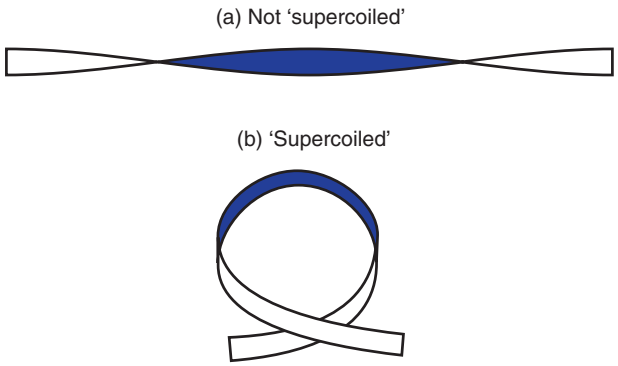


Figure 1.5 Interaction between twisting and supercoiling. (a) A ribbon with a single complete twist, without supercoiling. (b) The same ribbon, allowed to form a supercoil; the ribbon is now not twisted.

(in this example it now appears not to be twisted at all). If you have kept hold of both ends, the twist of the strip cannot have disappeared completely; it has merely changed to a different form. If you pull the ends apart again, it will change back to the form shown in Figure 1.5a.

There are three parameters involved: twist (T), linking number (L) and writhe (W). The twist is the number of turns of the strip whereas writhe (essentially a measure of the degree of supercoiling) can be considered as the number of times the strip crosses over itself in a defined direction. These two parameters vary according to the conformation: in Figure 1.5a there is one twist ($T = 1$) but no supercoiling ($W = 0$), whereas in Figure 1.5b there is no twist ($T = 0$) and the strip crosses itself once ($W = 1$). The linking number, which is a measure of the overall twisting of the strip, is equal to the sum of the other two parameters, i.e. $L = T + W$.

If the ends of the strip are not free to rotate, then the linking number will remain constant. Most of the DNA molecules we will be considering are circular, and therefore do not contain ends that can rotate. Unless there is a break in the DNA, any change in the twist will be balanced by a change in supercoiling, and vice versa. This is illustrated by Figure 1.6.

In Figure 1.6a the strip (or DNA molecule) is not supercoiled ($W = 0$) but contains one complete twist ($T = +1$); the linking number (L) is $+1$. In Figure 1.6b the overall shape has been changed by rotating one end of the structure (i.e. introducing a degree of supercoiling). The strip crosses itself once, and by convention a crossover in this direction is assigned a negative value, so $W = -1$. At the same time the twist has changed; there are now two complete twists, so $T = +2$. Since $L = T + W$, we see that L remains the same ($+1$).

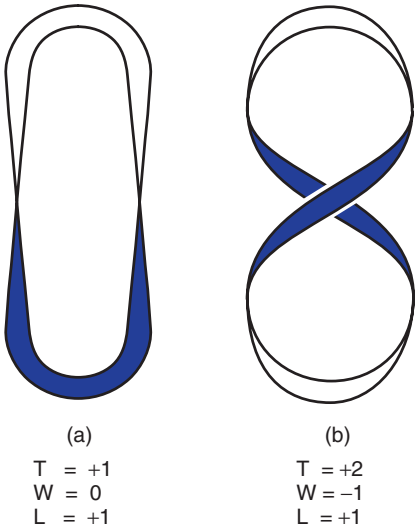


Figure 1.6 Supercoiling of a circular molecule. In (a), the 'molecule' has one twist and no supercoils. Rotating one end of the molecule (b) introduces negative supercoils and increases the amount of twisting. Since the linking number (L) remains the same, the two forms are interchangeable without breaking the circle. See the text for further explanation.

The two structures shown in Figure 1.6 are interchangeable by rotating one end, without opening the circle. With an intact circle, you can change the twist and the writhe jointly but not separately. Any change in supercoiling will involve a compensating change in the twist (and vice versa) so that the linking number remains constant. *It is only possible to alter the linking number in circular DNA by breaking and rejoining DNA strands*, for example through the action of topoisomerases (see below).

Bacterial DNA is normally negatively supercoiled. Another way of putting it is to say that the DNA is underwound so that if the DNA was not supercoiled, the degree of twisting of the helix would be less than that seen in relaxed linear DNA. If the DNA is nicked (i.e. one strand is broken, leaving it free to rotate) it relaxes into an open circular, non-supercoiled form. Chromosomal DNA is usually broken into linear fragments during lysis of the cell, but bacterial plasmids (see Chapter 5) are usually small enough to be isolated intact in a supercoiled form.

The compact supercoiled structure of the DNA is also significant in that the chromosome, in its expanded state, would be a thousand times longer (about 1 mm) than the bacterial cell itself. To put it another way, a bacterial operon of four genes, in its non-supercoiled B form, would stretch from one end of the cell to the other. Supercoiling is only the start of story, as the bacterial chromosome consists of a large number of supercoiled loops arranged on

a core to produce a highly compact and organized structure known as the *nucleoid*. Supercoiling (and other structural features) of the DNA are also important in the regulation of gene expression (see Chapter 3).

Action of topoisomerases

Supercoiling of bacterial DNA is not achieved by physically twisting the circular molecule in the way we illustrated in Figure 1.6. Instead the cell uses enzymes known as *DNA topoisomerases* to introduce (or remove) supercoils from DNA by controlled breaking and rejoining of DNA strands.

DNA topoisomerases can be considered in two classes. Type I topoisomerases act on a segment of DNA by breaking one of the strands and passing the other strand through the gap, followed by resealing the nick. Since this increases the number of times the two strands cross one another, the linking number is increased by 1, which results in an increase in either T or W. The *Escherichia coli* topoisomerase I acts only on negatively supercoiled DNA; the increase in the value of W means that the degree of negative supercoiling is reduced (the DNA becomes relaxed).

Type II topoisomerases break both strands and pass another duplex region through the gap. In Figure 1.7, we start with a non-supercoiled circle (structure A) and move the centres of the right- and left-hand loops across one another to give structure B. Although this may look at first glance to be supercoiled, it is not. The two crossovers are in opposite directions and therefore cancel one another out. (Mathematically $W = -1$ at the upper point, and $W = +1$ at the lower one, so overall $W = 0$.) If both strands of the helix are broken between points L and M, and the lower strands (X – Y) are moved through the gap, followed by resealing the strands between L and M, structure C is formed. Now, both crossovers are in the same direction, so the structure *is* supercoiled ($W = -2$). There is a corresponding reduction in the linking number. The enzyme has introduced a negative supercoil, and at the same time has reduced the winding of the helix. An important example of this type of enzyme is DNA gyrase, which is able to introduce negative supercoils into newly replicated DNA.

1.1.6 Denaturation and hybridization

Since the two strands of DNA are only linked by non-covalent forces, they can easily be separated in the laboratory, for example by increased temperature or high pH. Separation of the two DNA strands, *denaturation*, is readily reversible. Reducing the temperature, or the pH, will allow hydrogen bonds between complementary DNA sequences to reform; this is referred to as *re-annealing* (Figure 1.8). If DNA molecules from different sources are denatured, mixed and allowed to re-anneal, it is possible to form hydrogen bonds between similar DNA sequences (*hybridization*). This forms the basis of the use of DNA probes to detect specific

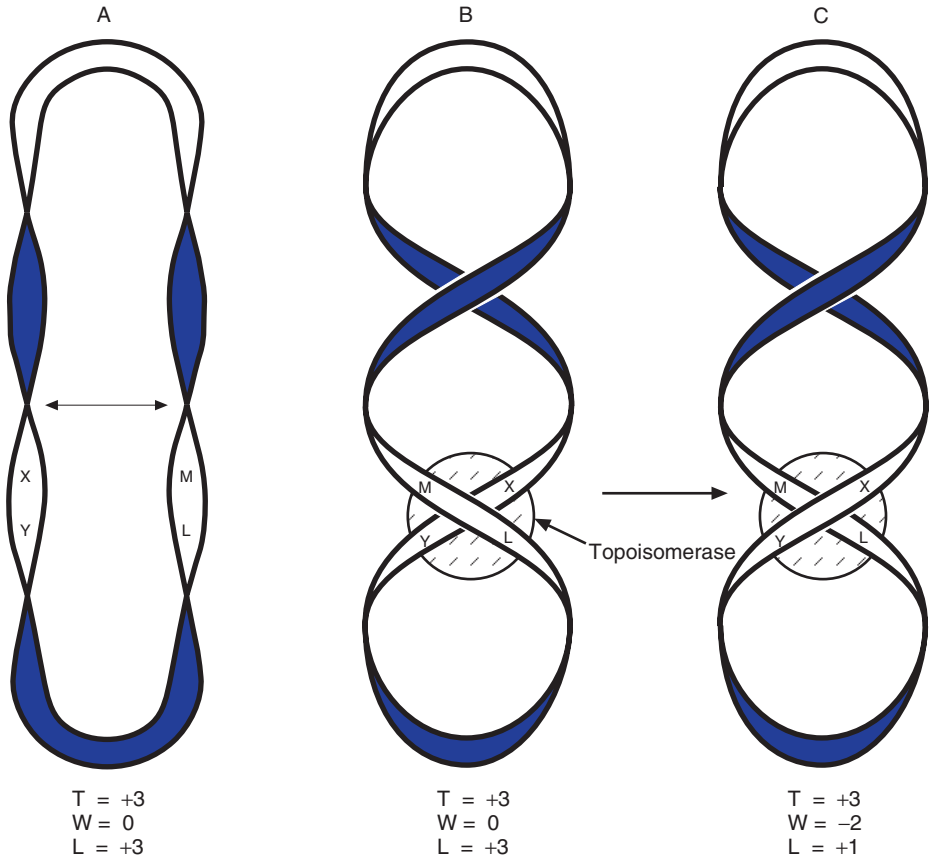


Figure 1.7 Action of Type II topoisomerase. Structure A is not supercoiled, and is converted to B by bending the two sides as shown by the arrow. B is not supercoiled: the two crossing points are of opposite sign and cancel one another. The topoisomerase makes a double-strand break between L and M, passes the X-Y region through the gap, and re-seals the break between L and M. This changes the sign of W at that point, so structure C is now negatively supercoiled.

DNA sequences. The specificity can be adjusted by altering the conditions used for re-annealing (or subsequent washing). Higher temperature, or lower ionic strength, gives greater *stringency* of hybridization. High-stringency hybridization is used to detect closely related sequences, or to distinguish between sequences with only small differences, whereas low-stringency conditions are used to detect sequences that are more remotely related to your probe. This technique forms an important part of modern molecular biology, and we will encounter many applications in subsequent chapters.

Temporary separation of localized regions of the two DNA strands also occurs as an essential part of the processes of replication and transcription.

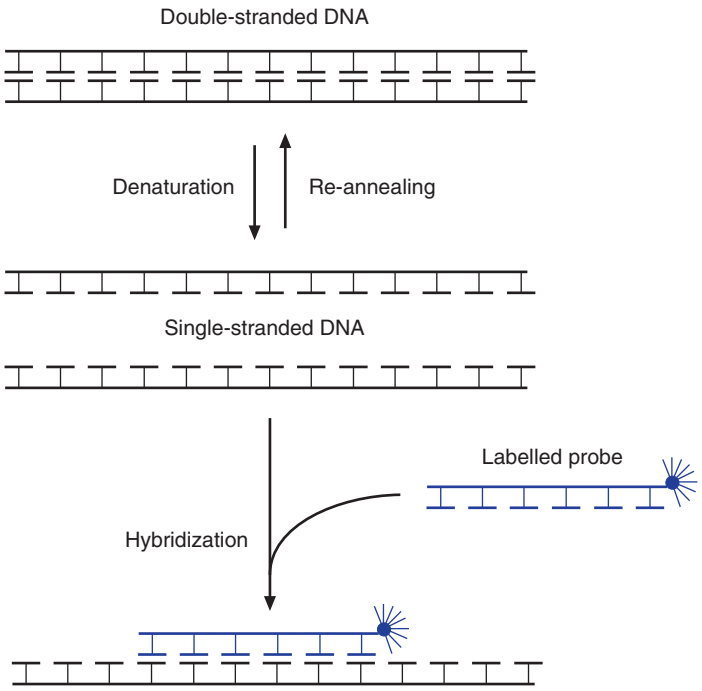


Figure 1.8 Denaturation and hybridization of DNA.

Note that there are three hydrogen bonds linking guanine and cytosine while the adenine–thymine pairing has only two hydrogen bonds. The two DNA strands are therefore more strongly attached in those regions with a high G + C content. Because of this, such regions are more resistant to denaturation and conversely re-anneal more readily. The influence of base composition on the ease of separation of two nucleic acid strands may play an important role in the control of processes such as the initiation of RNA synthesis where an A–T-rich region may facilitate the initial separation of the DNA strands (Chapter 3).

1.1.7 Orientation of nucleic acid strands

A further noteworthy feature of the helix is that each strand can be said to have a direction, based on the orientation of the linkages in the sugar-phosphate backbone. Each phosphate group joins the 5′ position of one sugar residue to the 3′ position of the next deoxyribose. In Figure 1.2, the upper strand has a free 5′ group at the left-hand end and a 3′ OH group at the right-hand end. It is therefore said to run (from left to right) in the 5′ to 3′ direction. Conversely, for the lower strand, the 5′ to 3′ direction runs from right to left.

By convention, if a single DNA (or RNA) strand is shown, it reads in the 5′ to 3′ direction from left to right (unless otherwise stated). If both strands are shown, the upper strand reads (left to right) from the 5′ to 3′ end.

All nucleic acids are synthesized in the 5′ to 3′ direction. That is, the new strand is elongated by the successive addition of nucleotides to the free 3′ OH group of the preceding nucleotide. The phosphate to make the link is provided by the substrate which is the nucleoside 5′-triphosphate (ATP, GTP, CTP and UTP for RNA; dATP, dGTP, dCTP and dTTP for DNA).

1.2 Replication of DNA

A DNA strand can act as a template for synthesis of a new nucleic acid strand in which each base forms a hydrogen-bonded pair with one on the template strand (G with C, A with T, or A with U for RNA molecules). The new sequence is thus complementary to the template strand. The copying of DNA molecules to produce more DNA is known as *replication*; the synthesis of RNA using a DNA template is called *transcription*.

Replication is a much more complicated process than implied by the above statement. Some of the main features are summarized in Figure 1.9. The opposite polarity of the DNA strands is a complicating factor. One of the new strands (the 'leading' strand) can be synthesized continuously in the 5′ to 3′ direction. The enzyme responsible for this synthesis is DNA polymerase III. With the other new strand, however, the overall effect is of growth in the 3′ to 5′ direction. Since nucleic acids can only be synthesized in the 5′ to 3′ direction, the new 3′ to 5′ strand (the 'lagging' strand) has to be made

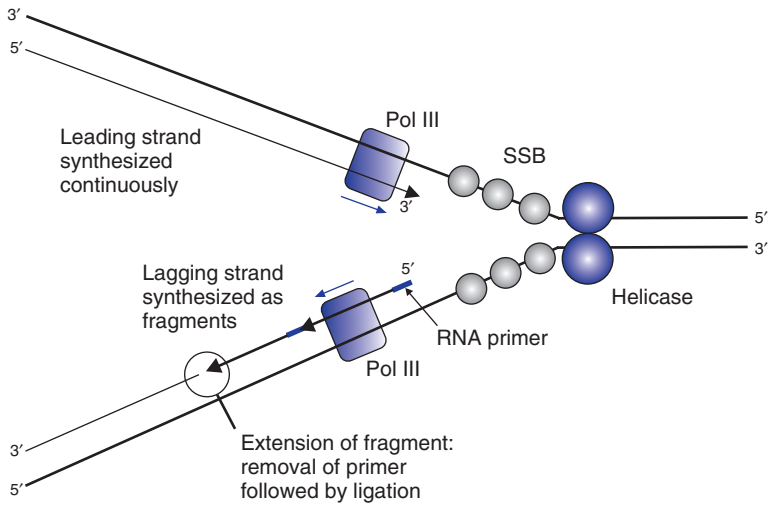


Figure 1.9 Simplified view of the main features of DNA replication. Note that the diagram does not show the helical structure of the DNA.

'backwards', i.e. in the opposite direction to overall replication. This can be done by making the new strand in short fragments (known as Okazaki fragments) which are subsequently joined together by the action of another enzyme, DNA ligase.

Furthermore, DNA polymerases are incapable of starting a new DNA strand, but can only extend a previously existing molecule. This restriction does not apply to RNA polymerases, which are able to initiate synthesis of new nucleic acids. Each fragment is therefore started with a short piece of RNA, produced by the action of a special RNA polymerase (*primase*). This RNA primer can then be extended by DNA polymerase III. The primer is subsequently removed, and the gap filled in, by a different DNA polymerase (DNA polymerase I); this enzyme can carry out both of these actions since it has exonuclease as well as polymerase activity. After the gap has been filled, the fragments that have been produced are joined together by DNA ligase.

1.2.1 Unwinding and rewinding

Before any of these events can take place, it is necessary for the two strands to be separated, for a short region at least. This is achieved by enzymes known as *helicases* which bind to the template strand and move along it, separating the two strands. The separated strands are prevented from re-associating by the binding of another protein, the single-stranded DNA-binding protein or SSB. A number of copies of the SSB will bind to the DNA strands, maintaining a region of DNA in an extended single-stranded form.

A further complication arises from the twisting of the two DNA strands around each other. DNA molecules within the cell cannot normally rotate freely, and not only because bacterial DNA is usually circular. Therefore it is not possible to produce a pair of daughter molecules by just separating the two strands and synthesizing the complementary strands, as is implied by the simplified representation in Figure 1.9. The strands have to be unwound to be separated. If they are not free to rotate, separating the strands at one point will cause overwinding further along. Unless this problem is overcome, the molecules would quickly get in a hopeless tangle. (If you don't understand this, try it for yourself with some bits of string!) The resolution of the problem requires the action of topoisomerases, as described earlier. By allowing the double helix to unwind ahead of the replication fork, they permit the strands to separate for replication. One topoisomerase, DNA gyrase, has the important role of introducing negative supercoils into the newly replicated DNA.

1.2.2 Fidelity of replication; proofreading

It is essential that the newly synthesized DNA is a precise (complementary) copy of the template strand. This does not arise simply by the nucleotides

aligning themselves in the right position, but involves the specificity of the DNA polymerase in selecting nucleotides that are correctly aligned.

Most DNA polymerases are more complex enzymes than the name suggests, as they also possess exonuclease activity. We have already encountered one such activity: the removal of the RNA primer from the Okazaki fragments is achieved by means of the 5′ to 3′ exonuclease activity of the DNA polymerase (i.e. it can remove bases from the 5′ end of a chain) as it extends the next fragment. The fidelity of replication is enhanced by a second exonuclease function of DNA polymerases: the 3′ to 5′ exonuclease activity, which is able to remove the nucleotide at the growing end (3′ end) of the DNA chain. This is not as perverse as it sounds, since the 3′ to 5′ exonuclease only operates if there is an incorrectly paired base at the 3′ end. The DNA polymerase will only extend the DNA chain, by adding nucleotides to the 3′ end, if the last base at the 3′ end is correctly paired with the template strand. If it is not, polymerization will stop, and the 3′ to 5′ exonuclease function will remove the incorrect nucleotide, allowing a further attempt to be made (Figure 1.10). The reasons for the occurrence of errors in adding bases to the growing DNA chain are dealt with in Chapter 2.

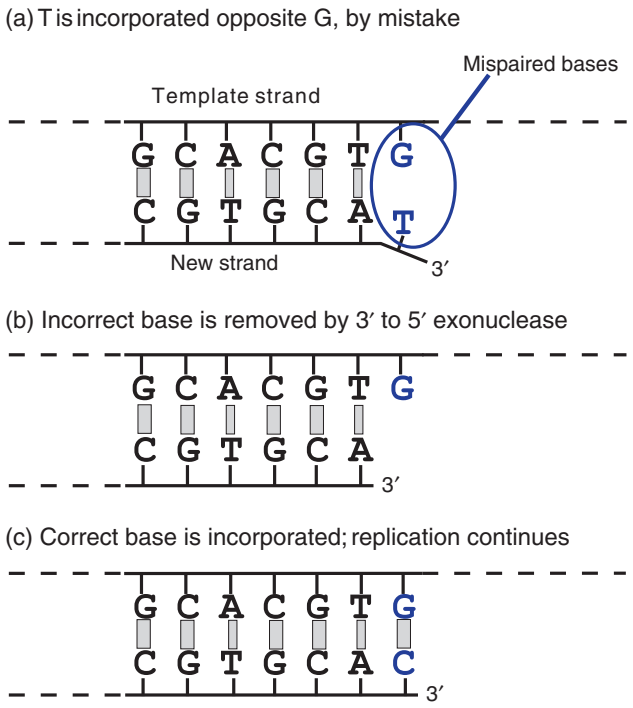


Figure 1.10 Elimination of mispaired bases by proofreading. (a) An incorrect base has been added to the growing DNA strand; this will prevent further extension. (b) The mispaired base is removed by the 3′ to 5′ exonuclease action of DNA polymerase. (c) The correct base is added to the 3′ end; DNA synthesis continues.

This mechanism of correcting errors, known as proofreading or error-checking, adds considerably to the fidelity of replication, thus reducing the rate of spontaneous mutation. There is a price to be paid however, as extensive error-checking will slow down the rate of replication. The balance between the rate of replication and the extent of error-checking will be determined by the nature of the DNA polymerase itself. Some DNA polymerases do not show efficient proofreading and therefore result in a much higher degree of spontaneous errors. The rate of spontaneous mutation shown by an organism is therefore (at least in part) a genetic characteristic that is subject to evolutionary pressure.

The fidelity of replication is further enhanced by DNA repair mechanisms, which are described later in this chapter.

1.3 Chromosome replication and cell division

Bacterial cells are generally regarded as having a single, circular chromosome. This is a simplification in several ways. Firstly, many bacteria often contain additional DNA molecules known as *plasmids*. In most cases these are additional, dispensable, elements, but in some bacterial species all strains carry two or more different DNA molecules, both (or all) of which appear to be essential for normal growth. These can equally well be regarded as essential plasmids or as additional chromosomes. Secondly, not all bacterial DNA is circular. Some bacteria (notably *Streptomyces*) have a linear chromosome and/or linear plasmids. These topics are discussed further in Chapter 5.

More fundamentally, immediately before cell division there must be at least two complete copies of the chromosome, in order to ensure that both daughter cells acquire a copy. Therefore, the chromosome must replicate in tune with the cell-division cycle, which means that at an intermediate time in the cycle part of the chromosome will have been copied, with the consequence that there are at least two copies of this part of the chromosome.

Replication of a bacterial chromosome normally starts at a fixed point (the origin of replication, *oriV*) and proceeds in both directions to a termination point (*ter*) that is approximately opposite to the origin (Figure 1.11). In *E. coli*, this takes about 40 minutes. There is then a period of at least 20 minutes before cell division, making a minimum of 60 minutes between the initiation of replication and cell division. However under favourable conditions *E. coli* will grow much faster than that, dividing perhaps every 20 minutes. How can the cells be dividing faster than the chromosome replicates, and still allow every daughter cell to acquire a complete copy of the chromosome? The answer lies in the timing of the initiation of replication. Initiation is stimulated, not by cell division, but as a function of the size of the cell.

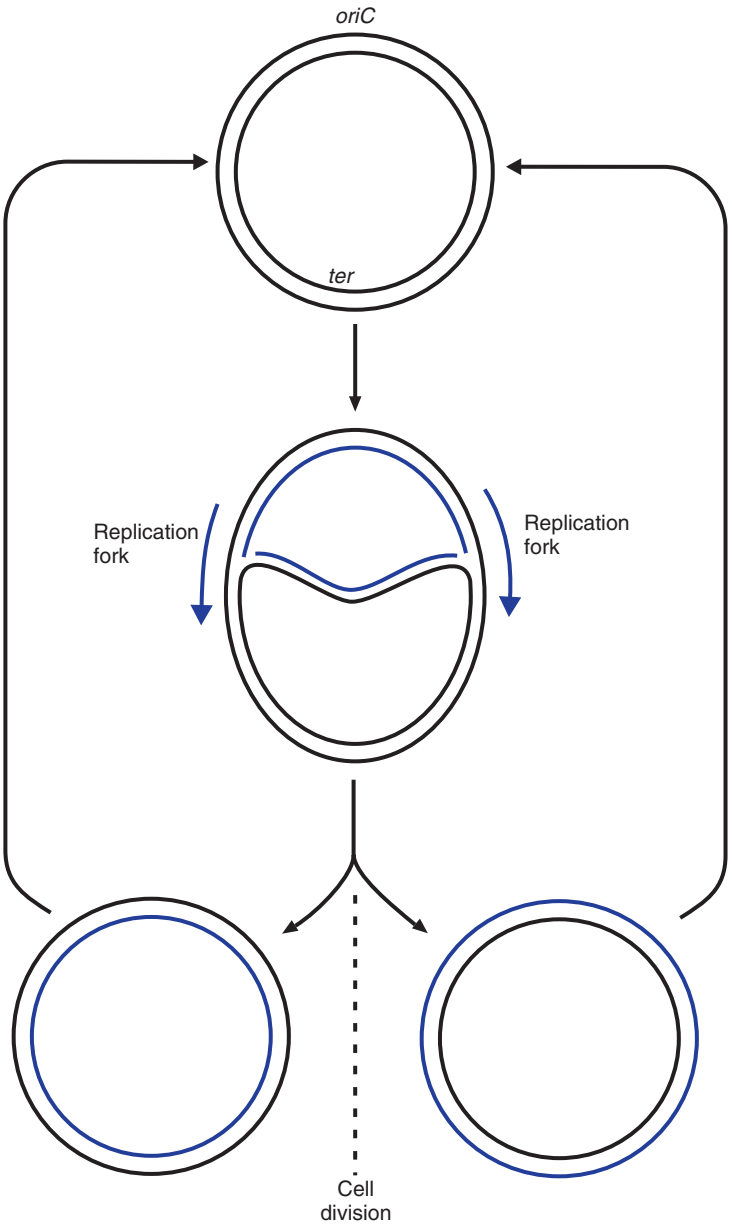


Figure 1.11 Chromosome replication. Bidirectional replication starts at *oriC* and continues to the termination site *ter*, producing two double-stranded molecules.

Consequently, when the cells are growing rapidly, there are several sets of replication forks copying the chromosome, so that when the cell is ready to divide there are not just a pair of completely replicated chromosomes, but each of these has in turn already been partly replicated by a second pair of replication forks (Figure 1.12).

There are two key regulatory points to be considered: the link between the completion of chromosome replication and subsequent cell division, and the control of the initiation of replication. Both are too complex (and still incompletely understood) to be considered fully here, but we can consider some aspects. On the first point, one (simplified) model is that the replicating chromosome occupies a region of the membrane at the midpoint of the cell, which prevents the initiation of cell division at that point. When replication has finished, the two separate molecules can be pulled apart, towards the poles of the cell, thus freeing the site for cell division to start. We will look further at this in Chapter 9.

The second point, the control of initiation, is more difficult. Initiation of new rounds of replication is triggered when the cell reaches a critical mass. It is tempting to think that this means that an inhibitor of replication is diluted out as the cell grows, but it is far from that simple. We know that initiation requires a protein called DnaA which binds to specific DNA sequences known as *DnaA boxes*; the origin of replication contains a number of DnaA boxes. Wrapping the DNA around the aggregated DnaA proteins facilitates the separation of the strands that is necessary for the initiation of replication (Figure 1.13). The availability of active DnaA is a significant component of initiation. However, the full story of the control of initiation is more complex, and still incompletely understood.

1.4 DNA repair

In addition to the measures, described earlier, that enhance the fidelity of replication, the cell also possesses mechanisms to correct damaged DNA, including replication errors that have escaped the proofreading process and damage that may have occurred in non-replicating DNA.

1.4.1 Mismatch repair

The simplest of replication errors is one that leads to the wrong base being incorporated into the new strand. If this occurs, and is not dealt with by the proofreading mechanism, it would lead to mutation. However, the cell has an effective mechanism for removing such mismatches, and replacing them by the correct nucleotide. In order to do this, it has to know which of the two strands contains the correct information. Mature DNA is methylated, i.e. it contains additional methyl groups, especially on the adenine in the sequence GATC, due to the action of deoxyadenosine methylase (Dam methylase). The new

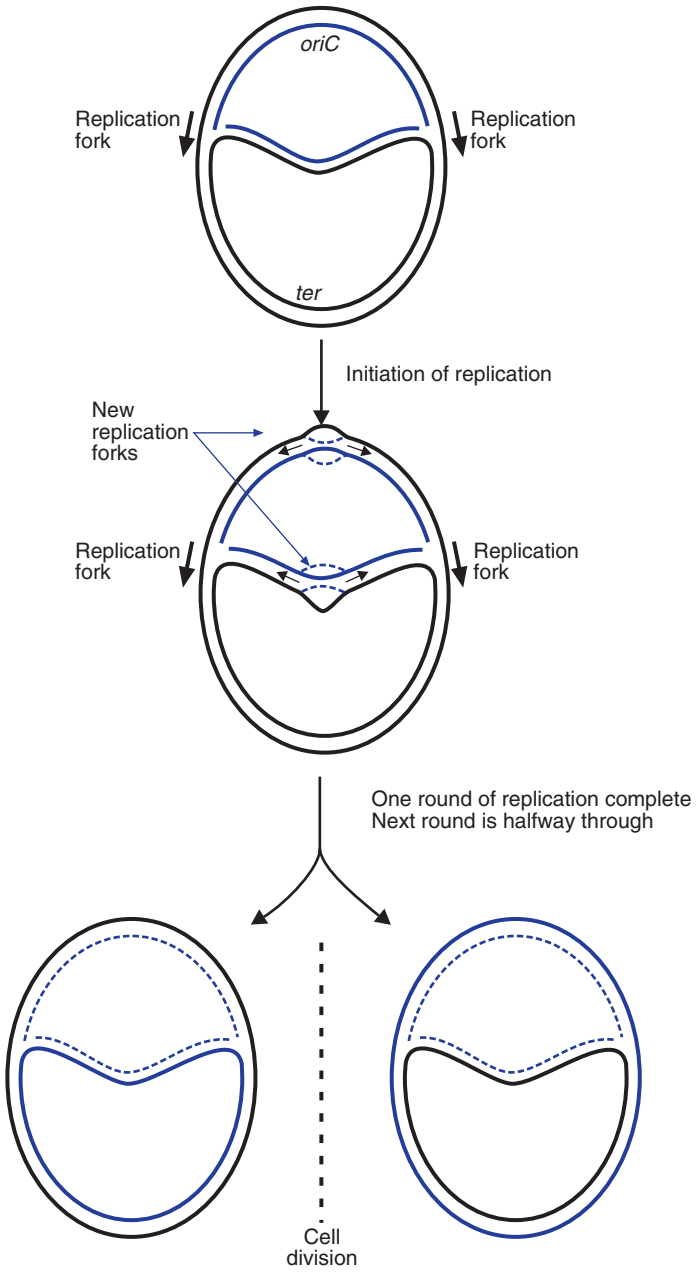


Figure 1.12 Chromosome replication at higher growth rates. When the interval between cell divisions is less than the time need for replication of the chromosome, a new round of replication starts before the previous one has finished.

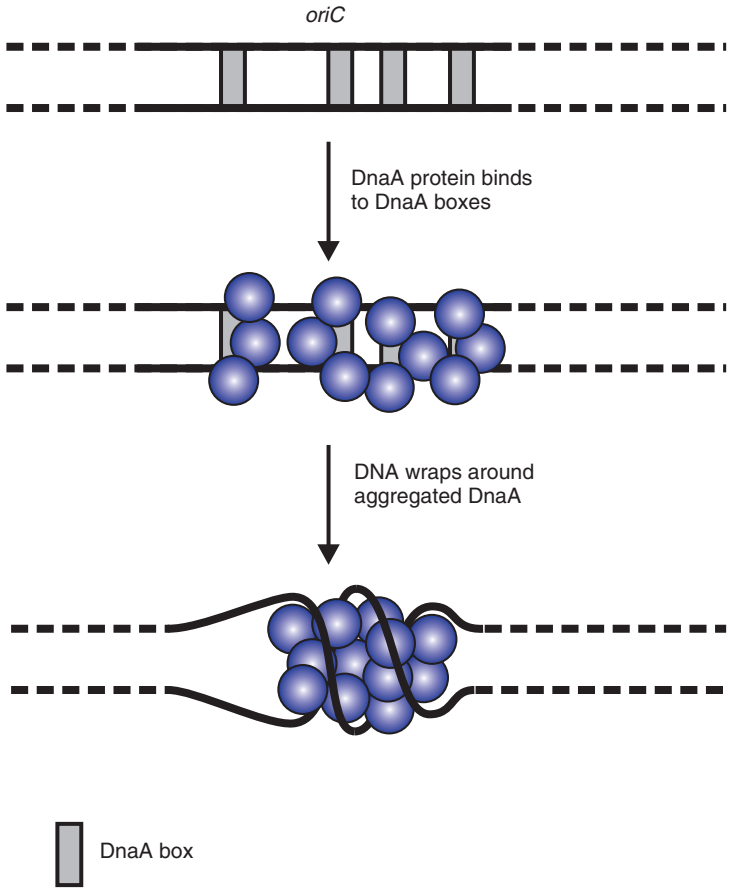


Figure 1.13 Binding of DnaA protein to DnaA boxes at *oriC*. The DnaA protein binds to several sites (DnaA boxes) at the origin of replication. Wrapping the DNA around the bound DnaA helps to separate the strands, enabling the initiation of replication.

strand does not, at first, contain these methyl groups, so we refer to the DNA as *hemimethylated*. The new and old strands can therefore be distinguished, at least until the new strand becomes methylated; hence the mechanism is referred to as *methyl-directed mismatch repair*. The system recognizes the mismatched bases, removes a short region of the non-methylated strand, and fills in the gap.

1.4.2 Excision repair

Other types of DNA damage, in particular the formation of pyrimidine dimers by ultraviolet irradiation (see Chapter 2), give rise to distortion of the double helix, which can activate a repair mechanism known as excision repair.

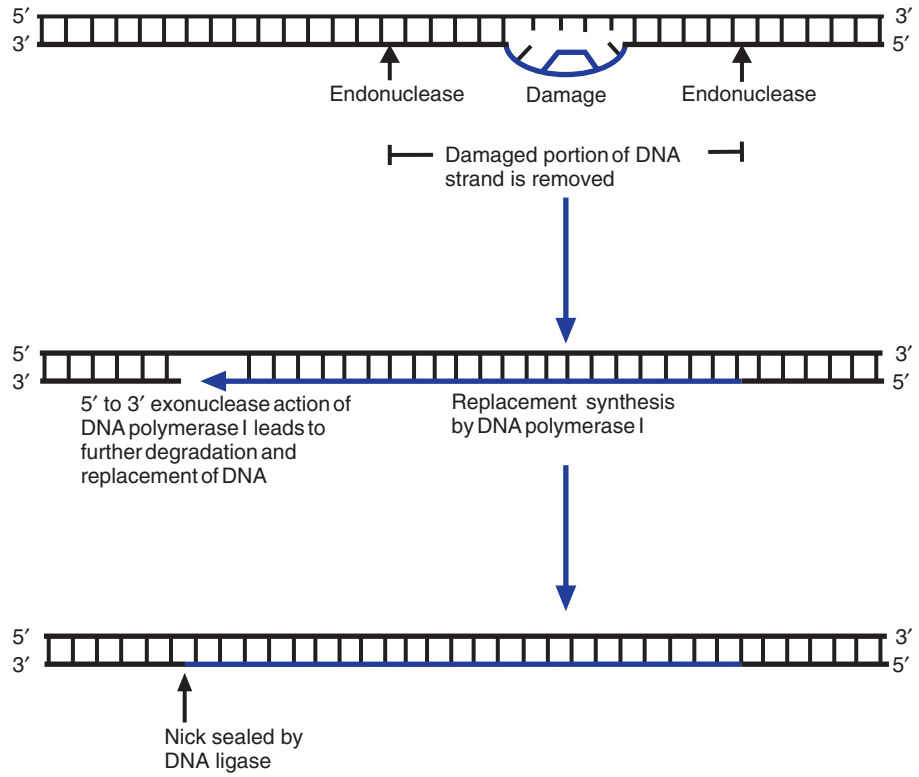


Figure 1.14 Mechanism of excision repair. Endonuclease cleavage removes a portion of the damaged strand. The gap is filled in by DNA polymerase I; the 5′ to 3′ exonuclease action of DNA polymerase I allows it to remove more DNA and replace it. Finally, the sugar-phosphate backbone is resealed by DNA ligase.

The essence of this mechanism is summarized in Figure 1.14. The process is initiated by an endonuclease (a complex enzyme, coded for by genes known as *uvrA, uvrB* and *uvrC*, since mutations in these genes cause reduced ultraviolet resistance). This enzyme cuts the DNA strand on either side of the damage, which exposes a 3′ OH group; this can be used as a primer by DNA polymerase I to replace the short region of DNA between the nicked sites (15–20 bases long). The final step is the joining of the newly repaired strand to the existing DNA, by DNA ligase.

1.4.3 Recombination (post-replication) repair

There is another type of mutant that is abnormally sensitive to ultraviolet radiation, although possessing a functional excision repair system. These bacteria are defective in a gene (*recA*) that is among other things responsible for general recombination (see Chapter 2). This indicates that excision repair

is not the only mechanism for dealing with ultraviolet damage, but that there is a further repair mechanism involving general recombination.

Forms of DNA damage that interfere with the base-pairing between the strands will normally prevent replication, due in part to the requirement of the DNA polymerase for an accurately paired 3′ end. The replication fork will therefore pause. It is possible however for replication to restart beyond the lesion, thus leaving a gap in the newly synthesized strand (Figure 1.15). This

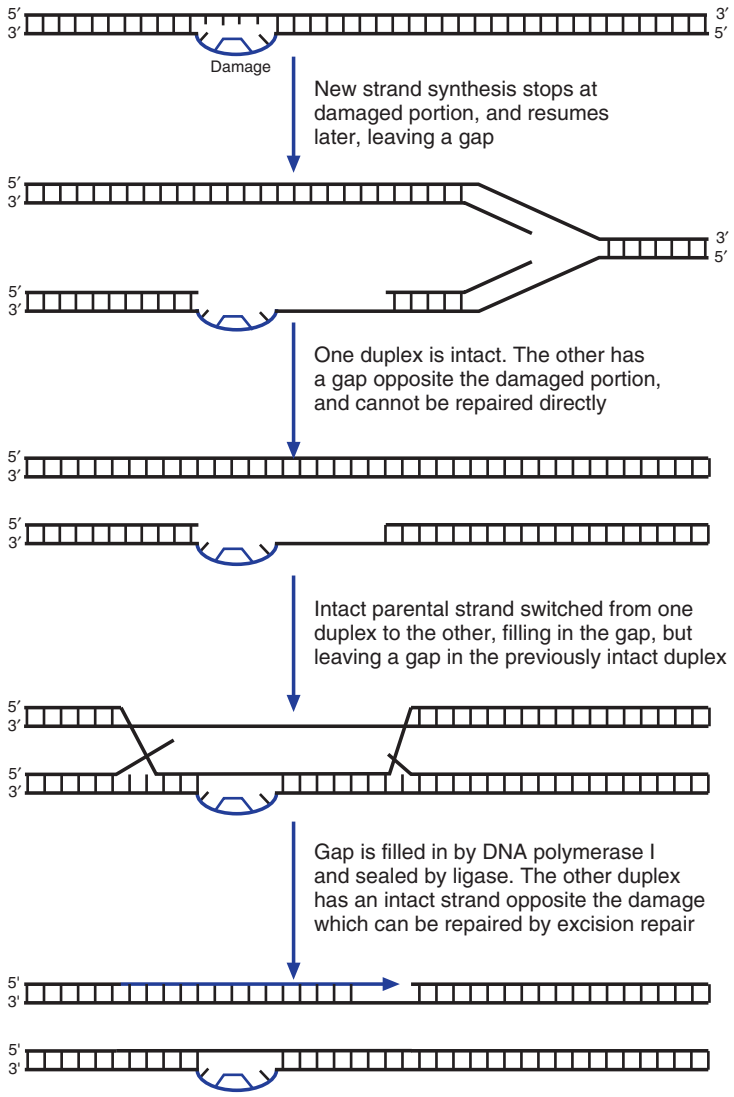


Figure 1.15 Post-replication repair. After replication stops at a damaged site, subsequent re-initiation leaves a gap in the new strand. This can be repaired by exchange of DNA, allowing the original lesion to be repaired by excision repair.

portion of DNA cannot be repaired by excision repair, which requires one intact strand. The gap can however be filled using a portion of DNA from the other pair of strands, by a recombination process (i.e. cutting and rejoining the DNA). Although this merely re-assorts the damage rather than directly repairing it, it does achieve a situation where the damage is repairable. The original damage can now be repaired by excision repair, while the gap in the other DNA molecule can be filled in by DNA polymerase I and DNA ligase.

1.4.4 SOS repair

An alternative strategy, when faced with overwhelming levels of DNA damage preventing normal replication, is to temporarily modify or abolish the specificity of the DNA polymerase. This enables it to continue making a new DNA strand, despite the absence of an accurately paired 3′ end. Although the new strand is produced, it obviously is likely to contain many mistakes, and the process is therefore described as 'error-prone' repair. SOS repair is the cause of mutations arising from ultraviolet irradiation, and is considered further in Chapter 2.

1.5 Gene expression

The expression of the genetic material occurs for the most part through the production of proteins, involving two consecutive steps in which the information is converted from one form to another: transcription and translation. With those proteins that consist of several different subunits, each one is the product of a distinct region of DNA. The complete protein is thus the product of several different genes, mutation in any of which may lead to the absence of a functional product. The term *cistron*, meaning that region of the DNA that codes for a single polypeptide chain, is used where it is desirable to emphasize the distinction between a single polypeptide and a multimeric protein.

1.5.1 Transcription

The first step is the conversion of the information into messenger RNA (mRNA). This process (transcription) is carried out by RNA polymerase. As with DNA synthesis, the RNA strand is made in the 5′ to 3′ direction. However, there are major differences between transcription and replication. Firstly, only a comparatively short molecule is produced, and secondly, only one of the DNA strands is transcribed. (Some genes use one strand, and some use the other, but in general any specific region of DNA is only transcribed from one strand.) Since only a single strand is made, it can be produced continuously, using a single enzyme; there is no need for lagging-strand synthesis. In addition, the production of relatively short

single-stranded RNA causes fewer topological problems: the enzyme and the RNA product can essentially rotate around the helix, so there is no need for the helicases and topoisomerases that are essential for replication. Furthermore, RNA polymerase can start synthesis from scratch: no primer is needed. Transcription is therefore considerably simpler than replication.

Promoters

Since transcription results in the synthesis of comparatively short mRNA molecules (often just a few kilobases long, corresponding to a defined block of several genes) there must be a large number of signals around the chromosome that direct the RNA polymerase to start transcription at the required place, and to stop when the block of genes has been transcribed. The start signals (*promoters*) also convey the information as to the direction in which transcription should proceed, or which strand to work from, which is another way of saying the same thing.

In *E. coli*, depending on growth conditions, 2000–5000 copies of RNA polymerase may be engaged on mRNA synthesis at any time. The basic structure of RNA polymerase consists of four polypeptides: two identical α chains plus two other chains (β and β') that are related to one another but are not identical. This structure ($\alpha_2\beta\beta'$) is referred to as the *core* enzyme. The specificity of promoter binding is due to a fifth subunit, the σ (sigma) factor; the complete structure including the σ factor is called the *holoenzyme*. As we will see in Chapter 3, there are different classes of promoters, recognized by different sigma factors, which allows selective expression of certain groups of genes.

After the RNA polymerase holoenzyme binds to the promoter region, the initial structure (the closed complex) is converted to an open complex (Figure 1.16) in which there is localized separation of the two DNA strands. This exposes the bases of the coding strand, allowing base-pairing of the ribonucleoside triphosphates for synthesis of the RNA. The first phosphodiester bond is formed and the σ factor dissociates from the complex. From now on, the core enzyme alone is required for extension of the RNA strand. A short region of the newly formed RNA remains base-paired to the DNA template, which keeps the DNA strands from re-associating, and therefore permits continued RNA synthesis, until a termination signal is reached, when the mRNA and the RNA polymerase are released.

Transcriptional terminators

A characteristic feature of a transcriptional terminator is the presence of a short sequence that is complementary to the sequence just preceding it. When such a sequence is transcribed, the RNA formed can establish a

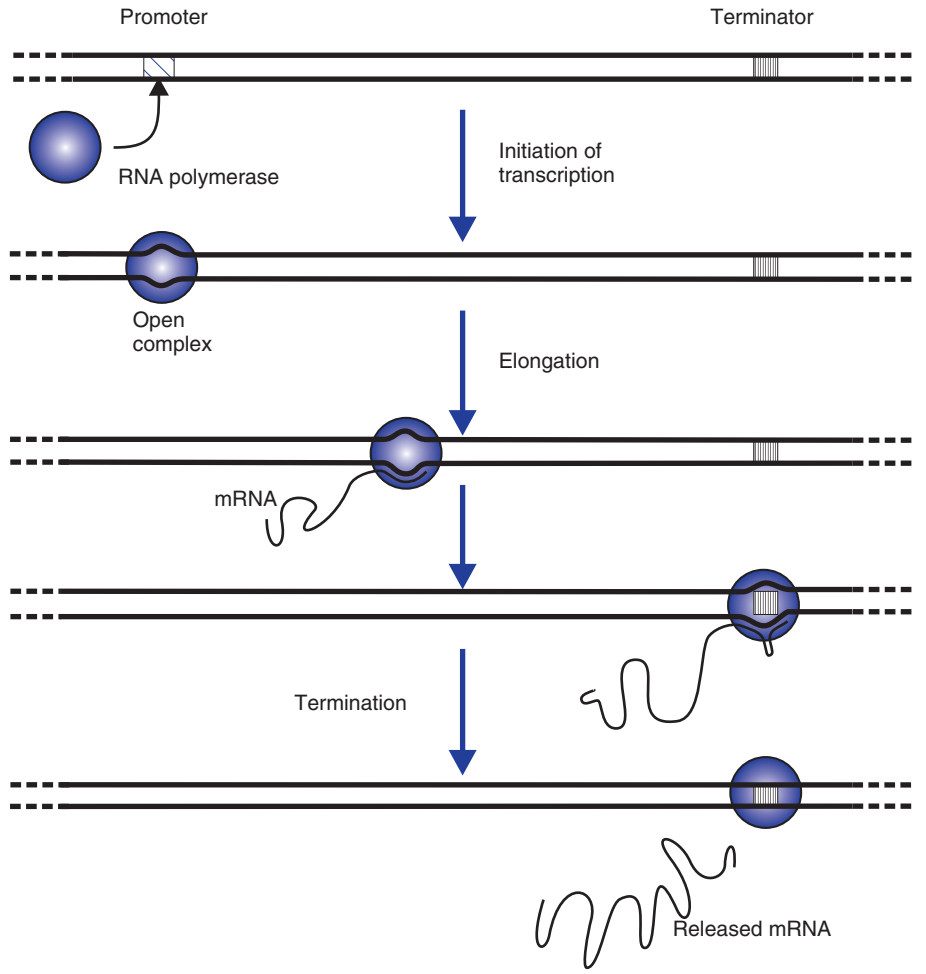


Figure 1.16 Main features of transcription.

stem-loop structure as shown in Figure 1.17. In most terminator sequences, the stem-loop structure is followed by a run of U residues.

A model that accounts for the termination of transcription is presented in Figure 1.18. RNA polymerase requires a short length of unwound DNA (about 17 bp) in which the two DNA chains are separated. However, the two DNA strands will tend to snap back together unless something prevents them.

It is the mRNA itself that is responsible for keeping the DNA 'bubble' open, by remaining base-paired with the complementary DNA strand for a short while. Under physiological conditions, the RNA–DNA hybrid is

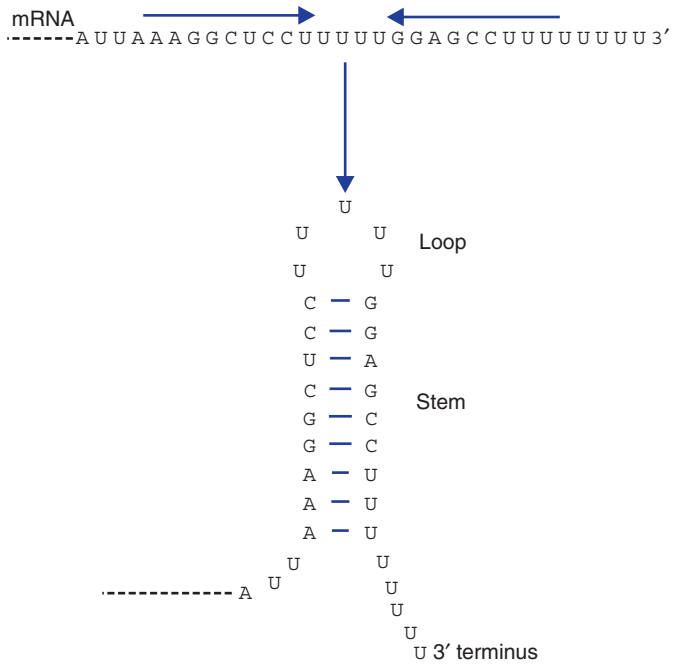


Figure 1.17 Structure of a typical terminator. The regions marked with arrows are complementary, and so can anneal together resulting in the formation of a stem-loop structure.

more stable than the DNA–DNA pairing. The size of the bubble is limited, however, by topological constraints. The helix has to be unwound to some extent to allow the strands to separate, which causes stress in the molecule, since the two strands can only be separated by increasing the winding on either side. The larger the unwound region, the greater the stress. Beyond this point therefore the remainder of the mRNA molecule is dissociated from the template DNA (Figure 1.18a).

When the RNA polymerase encounters a transcriptional terminator sequence it will transcribe it into RNA, with the consequent formation of a stem-loop structure which includes a portion of RNA that would otherwise be engaged on keeping the DNA bubble open (Figure 1.18b). The bubble will close up, which will hinder the activity of the RNA polymerase. The enzyme will therefore pause at this point, a few bases beyond the stem-loop structure (Figure 1.18c). Since the stem-loop structure in a typical termination sequence is followed by a string of U residues (Figure 1.17), all that keeps the mRNA attached at this stage is the relatively weak hydrogen bonding of the A–U base pairs. As a result, the RNA tends to dissociate from the DNA template, thereby terminating mRNA synthesis.

With some terminators, the stem-loop structure is not followed by a run of U residues. Although the RNA polymerase may pause at these sites,

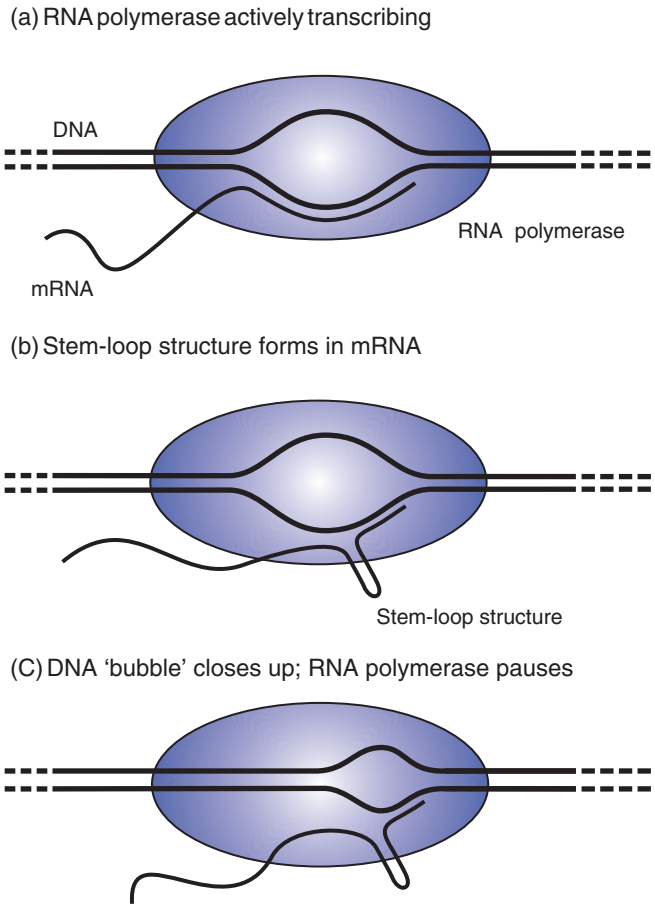


Figure 1.18 Model for transcription termination. (a) During transcription, the presence of the mRNA keeps the DNA strands separated over short region; (b) When the termination site is reached, the mRNA forms a stem-loop structure; (c) This allows the DNA strands to re-associate, leading to a pause in transcription.

termination is dependent on the activity of another protein known as the rho factor. These are therefore known as rho-dependent terminators.

1.5.2 Translation

The genetic code

The mRNA carries the information for the sequence of amino acids in a protein in the form of the genetic code (see Appendix F) in which each occurrence of one of the 64 groups of three nucleotides (triplets or *codons*) codes for a specific amino acid (or for a stop signal).

The code is almost universal, in all organisms, although there are occasional minor differences, such as the use of UGA, which is normally a stop codon, to code for tryptophan or cysteine. This indicates that the code as we know it must have originated early in the evolutionary process, and then became fixed because of the effect that any change would have on virtually every gene in the cell.

Ribosomes

Bacterial ribosomes typically consist of two subunits, with sedimentation coefficients of 50 S and 30 S, and the whole structure being referred to as a 70 S ribosome (sedimentation coefficients are not additive). The larger (50 S) subunit has two RNA molecules (23 S and 5 S) plus 31 different polypeptides, whereas the smaller one (30 S) contains a single RNA molecule (16 S) and 21 polypeptides. Note that the structure of eukaryotic ribosomes is different in several respects.

The ribosomal RNA molecules form a very stable three-dimensional structure by extensive base-pairing, which allows them to perform a scaffolding role in the attachment of the ribosomal proteins. The role of the rRNA extends further than this, as it is involved in recognition of the mRNA (see below) and in the catalytic events leading to peptide-bond formation.

In bacteria, the ribosomes attach to a specific sequence on the mRNA (the ribosome-binding site, or RBS, also known as the Shine–Dalgarno sequence after the workers who first recognized its significance). This sequence is partly complementary to the 3′ end of the 16 S rRNA (Figure 1.19), so that binding of the ribosomes can be mediated by hydrogen bonding between the complementary base sequences. This will normally occur as soon as the binding site is available, so the mRNA will start to be translated while it is still being formed (Figure 1.20).

However, it is not the complete ribosome that initiates these events. Ribosomes that are not involved in translation dissociate into their constituent 50 S and 30 S subunits. For translation to start, a 30 S subunit binds to an RBS, and an initiator tRNA (see below) associates with an adjacent initiation codon (usually AUG, but sometimes GUG or even less commonly CUG). The 50 S subunit can then attach to this initiation complex, and the process of

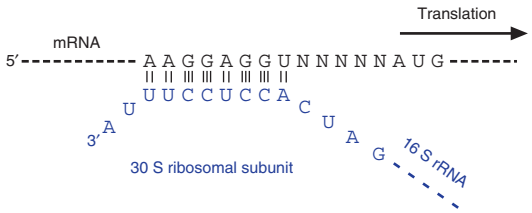


Figure 1.19 Ribosome-binding site (Shine–Dalgarno sequence).

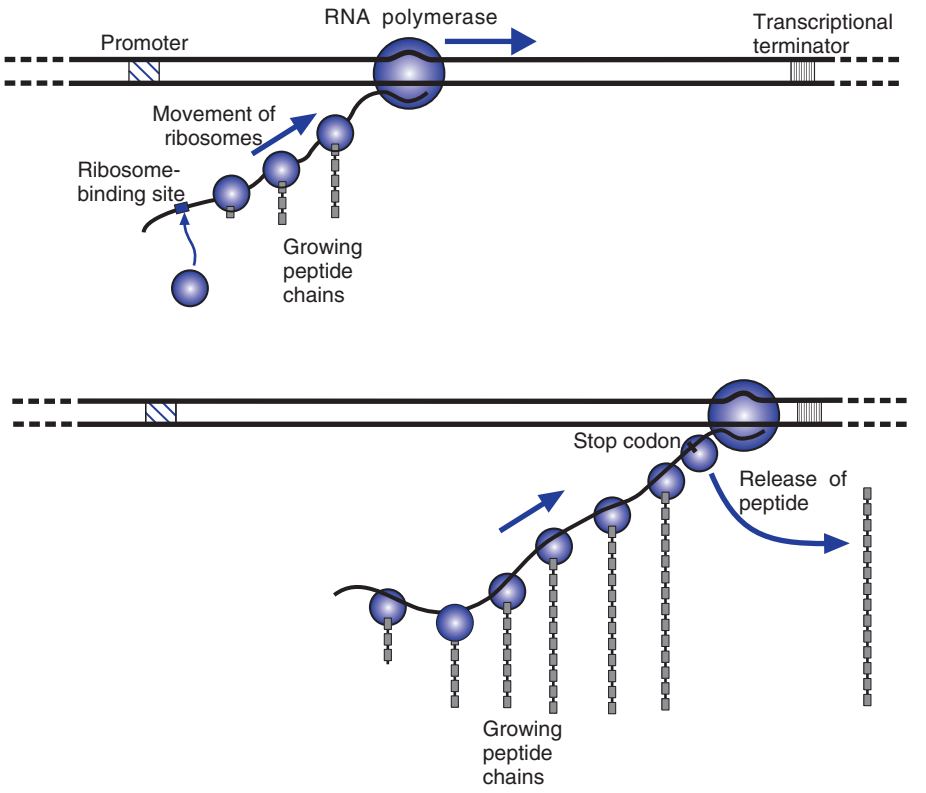


Figure 1.20 Translation of mRNA.

translation can get underway. These steps also require the action of several non-ribosomal proteins known as initiation factors.

Since the mRNA is read in consecutive groups of three (with no punctuation), it could code for three completely different proteins, depending on where it starts; i.e. there are three potential *reading frames*. The position of the RBS and the initiation codon determines the reading frame. As we will see in Chapter 2, the addition or deletion of a single base will change the reading frame, and the coding property of the subsequent message is totally different.

Transfer RNA

Recognition of each triplet codon is mediated by small RNA molecules known as transfer RNA (tRNA). There is at least one tRNA species specific for each amino acid. However, they are all quite similar in their structure, consisting of a single RNA chain of 75–100 nucleotides folded back on itself in a form usually depicted as a cloverleaf structure (Figure 1.21); the actual three-dimensional structure is more complex and compact than this simplified

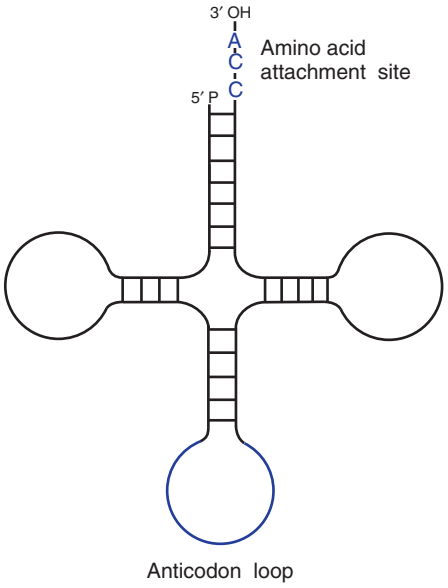


Figure 1.21 Diagrammatic structure of tRNA.

two-dimensional diagram. Two parts of this molecule have clear functions in protein synthesis: the acceptor arm, formed by base-pairing of the 5′ and 3′ terminal regions, provides the site for attachment of an amino acid (by acylation of the 3′ end), and the anticodon arm, which contains the bases (the anticodon) that recognize the triplet codon in the mRNA by base-pairing.

The appropriate amino acid is added to the tRNA by a specific enzyme (one of a number of aminoacyl tRNA synthetases) which has a crucial dual specificity: it is capable of recognizing a single tRNA species and also the correct amino acid with which that tRNA should be charged. Thus for example the codon UGG (which codes for tryptophan) will be recognized by a specific tRNA (designated $tRNA^{Trp}$). This tRNA will be recognized by the tryptophanyl tRNA synthetase. This therefore ensures that the tRNA is charged with the appropriate amino acid.

So there are three separate elements to the specificity of this process: codon–anticodon interaction, recognition of the specific tRNA by the aminoacyl tRNA synthetase, and recognition by the enzyme of the appropriate amino acid. Since tRNA molecules are all basically quite similar, and some amino acids (such as isoleucine and valine) are also similar to one another, it would not be surprising if mistakes were made occasionally. The low frequency of such errors (it has been estimated that one protein molecule in a thousand contains one incorrect amino acid) is due to the existence of an editing mechanism whereby the synthetase is able to cleave the amino acid from an incorrectly charged tRNA molecule.

If all three elements of specificity were absolute, then there would have to be at least 61 different tRNA species: one for each of the 64 codons less the three stop codons, for which there is no corresponding tRNA. For many of the amino acids there are indeed multiple tRNA species with different codon specificities. Some of these tRNA molecules are present at comparatively low levels in the cell, which would indicate that there could be a difficulty in translating that particular codon. This can be correlated to some extent with the frequency of occurrence of particular codons (codon usage): those codons that require a rare tRNA species tend also to occur less commonly, at least in highly expressed genes.

However, this is not the complete story. For many tRNA molecules, the codon–anticodon recognition is not absolutely precise; in particular, there is some latitude allowed in the matching of the third base of the codon. A rather complex set of rules (the wobble hypothesis) has been developed to account for the extent of allowable mismatching. So some tRNA molecules are able to recognize more than one codon. The number of tRNA species required for recognition of the complete set of codons is thus considerably less than 61 (commonly between 30 and 40).

Mechanism of protein synthesis

In bacteria, the initiation codon is recognized by a specific tRNA molecule, $tRNA^{fMet}$. After this tRNA molecule is charged with methionine, the amino acid is modified, to *N*-formylmethionine. Aminoacylated tRNA molecules normally bind to a site on the ribosome known as the A (Acceptor) site, while their anticodon region pairs with the mRNA. Only after peptide-bond formation is the tRNA able to move to a second site on the ribosome, the P (Peptide) site. The fMet-$tRNA^{fMet}$ (i.e. the $tRNA^{fMet}$ charged with formylmethionine) is unique in being able to enter the P site directly.

The $tRNA^{fMet}$ anticodon recognizes (forms base pairs with) the start codon on the mRNA, in association with binding of a 30 S ribosome subunit to the nearby RBS. The 50 S ribosome subunit then joins the complex (Figure 1.22). The charged tRNA corresponding to the second codon then enters the A site on the ribosome and peptide-bond formation occurs by transfer of the fMet residue to the second amino acid. The $tRNA^{fMet}$, now uncharged, is released, and the ribosome moves one codon along the mRNA, which is accompanied by movement of the second tRNA molecule (now charged with a dipeptide) from the A site to the P site; this step is known as translocation. The A site is thus free to accept the charged tRNA corresponding to the third codon. The cycle of peptide-bond formation, translocation and binding of a further aminoacylated tRNA requires several additional non-ribosomal proteins (elongation factors) and is accompanied by the hydrolysis of GTP.

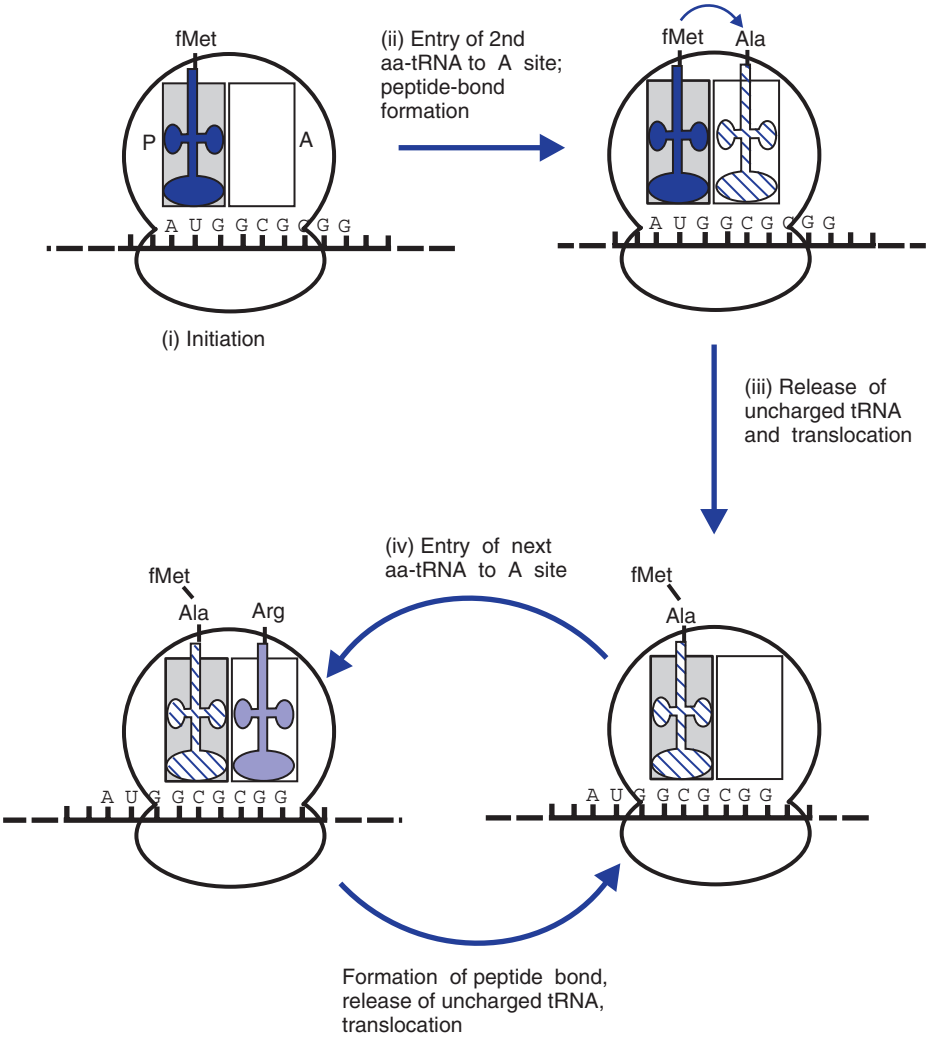


Figure 1.22 Outline of the mechanism of protein synthesis. (i) The initiation complex is formed by binding of a 30 S ribosomal subunit to the mRNA, followed by the initiating mRNA and a 50 S subunit. (ii) The second tRNA enters the A site and the first peptide bond is formed by transfer of the N-terminal fMet to the free amino group of the second amino acid. (iii) The uncharged initiating tRNA is released and the ribosome moves along the mRNA. The second tRNA is now in the P site, and the A site is free for the next tRNA. (iv) Entry of the next tRNA to the A site, ready for peptide bond formation as in step (ii). Steps (iii) and (iv) are then repeated, until a stop codon is reached.

When the ribosome has moved far enough, the RBS on the mRNA is exposed again and another ribosome can attach to it. A single mRNA molecule will therefore carry a number of ribosomes actively translating the sequence (Figure 1.20).

Each ribosome moves along the mRNA until a stop codon is reached. The absence of a corresponding tRNA species capable of recognizing this codon causes translation to stop at this point. The polypeptide chain is then released, with the aid of proteins known as release factors, and the ribosome dissociates from the mRNA.

1.5.3 Post-translational events

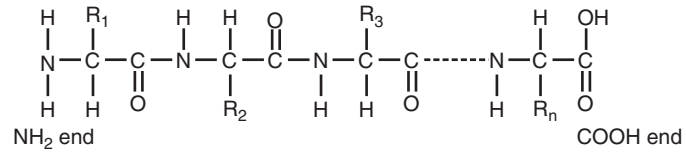
Translation of the information into a polypeptide does not complete the story. The formation of a biologically active product involves several further steps. First of all, the protein has to fold correctly. There are three conformational levels, in addition to the primary structure (which is the amino acid sequence itself). The secondary structure is the spatial arrangement of successive amino acids, which may form regular structures such as α-helices or β-sheets (Figure 1.23), which are stabilized by non-covalent interactions (such as hydrogen bonds). Different regions of the protein will adopt different secondary structures, separated by turns or less defined loops.

The various elements of secondary structure are in turn folded together to form the tertiary structure (Figure 1.24). This conformation is stabilized by non-covalent interactions, and also by covalent disulphide bridges between cysteine residues. The tertiary structure may include two or more semi-autonomous regions known as *domains*. Many proteins consist of several (identical or different) polypeptide chains; the way in which these polypeptides are associated constitutes the quaternary structure (Figure 1.25).

It is not easy to predict the structure that will be adopted by a polypeptide chain, partly because of the large number of possibilities and partly because the polypeptide will tend to form secondary (and higher-order) structures as it is being produced, thereby posing constraints on the structure adopted by the subsequent parts of the chain. The final structure is not necessarily the most thermodynamically stable form that can be adopted by refolding of the entire polypeptide.

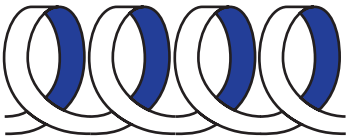
Furthermore, the folding of the polypeptide is not entirely spontaneous. Cells contain proteins known as *molecular chaperones* which assist in obtaining the correct conformation of proteins, for example by interacting with the nascent polypeptide to prevent it from adopting an incorrect conformation until the complete protein is produced. Molecular chaperones can also play an important role in the refolding of denatured proteins, which can provide a degree of protection against heat and other stress conditions. Some of these molecular chaperones are specifically produced under conditions that lead to the accumulation of denatured protein, and are known as heat-shock proteins, or stress proteins. A further role of molecular chaperones is concerned

(a) Primary structure



(b) Secondary structures

(i) α helix



(ii) β sheets

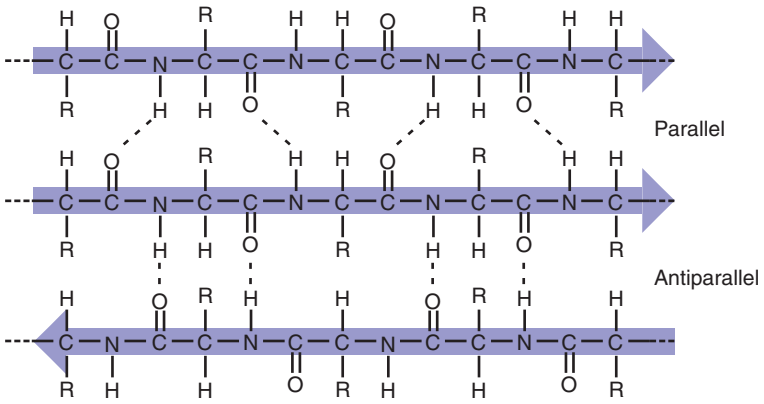


Figure 1.23 Primary and secondary protein structure. (a) Primary structure consists of the sequence of amino acids. (b) Two main forms of secondary structure are the alpha (α) helix (i) and beta (β) sheet (ii).

with the assembly of polypeptide subunits into multimeric proteins or larger structures; for example, the assembly of bacteriophage heads (Chapter 4) may require the action of molecular chaperones.

Secretion

Many bacterial proteins have functions that require them to be present on the surface of the cell or in the extracellular environment. The first barrier

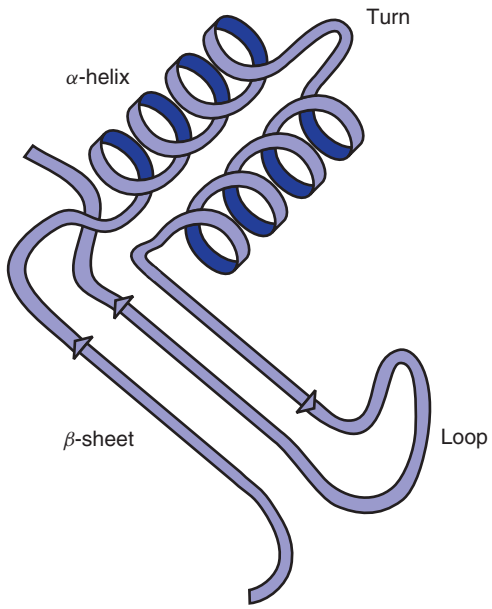


Figure 1.24 Protein secondary and tertiary structure. Tertiary structure refers to the way in which the elements of secondary structure are arranged. This shows two α-helices separated by a turn, a β-sheet, and a relatively loose and unconstrained region known as a loop.

to protein export is the cytoplasmic membrane, and the most common mechanism for transport of proteins across this membrane is known as the general secretory pathway (GSP, sometimes called the Sec-dependent pathway). All proteins that utilize this system have a specific sequence at the N-terminus which targets the protein to this pathway and which is cleaved during transport. In most Gram-positive bacteria, this mechanism is sufficient for export of proteins to the cell surface or to the surrounding medium. However, Gram-negative bacteria also have an outer membrane, and the GSP by itself will deliver proteins, not to the outside of the cell, but to the region known as the periplasm, between the cytoplasmic and outer membranes.

Although Gram-negative bacteria are less prolific than Gram-positive bacteria in secreting proteins, they do have important secretion mechanisms. The most common of these, the Type II mechanism, is dependent on the GSP for transport of proteins to the periplasm and then uses a specific multiprotein complex to transport the protein across the outer membrane.

Most Gram-negative bacteria (and some Gram-positive bacteria) also possess a number of separate secretion systems which are not dependent on the GSP and which are, therefore, termed Sec-independent. The Type IV secretory apparatus is notable as it bears many similarities with the

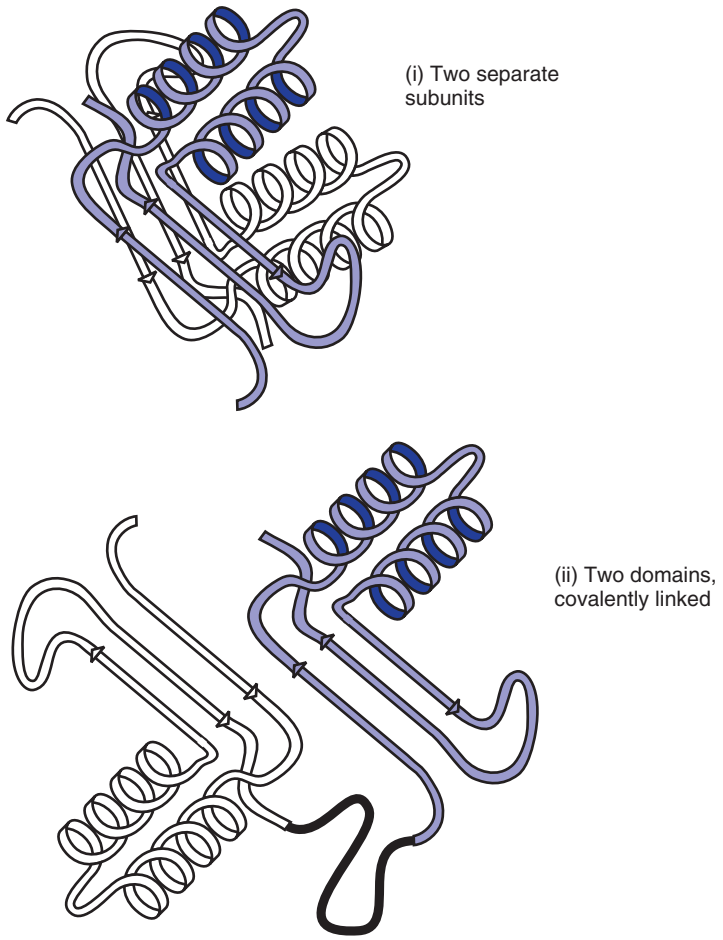


Figure 1.25 Protein structure: subunits and domains. (i) A protein may consist of two or more separate polypeptides (identical or different); the way in which these subunits associate is known as the quaternary structure. (ii) A single polypeptide may fold into separate *domains*, which often have distinct functions. An enzyme from one source may have two separate subunits, while a related enzyme from another source has two domains of a single polypeptide.

conjugal plasmid transfer systems described in Chapter 6, and can be used by a bacterium to introduce proteins into eukaryotic cells. Perhaps the most remarkable secretion system is the Type III pathway, which has been likened to a molecular syringe, and which is used to inject proteins directly into the cytosol of eukaryotic cells (this type of system is described in more detail in Chapter 9). Secretion is only triggered by direct contact between the bacterium and the host cell. Consequently, this pathway is widely used by pathogens to introduce effector molecules into host cells, thus subverting the normal function of the cell to the benefit of the bacterium.

Other post-translational modifications

In addition to the events described above, proteins may undergo a wide range of additional post-translational modifications, such as glycosylation, biotinylation, addition of lipids, and proteolytic cleavage. The full range of these is too complex to be covered here, but the outcome is that the final structure can be influenced very strongly by the nature of the cell itself. Since the post-translational events may be essential in obtaining a product with full biological activity, the difficulty in obtaining accurate post-translational modification can severely affect the outcome of attempts to obtain functional gene expression in heterologous hosts (see Chapter 8).

1.6 Gene organization

In bacteria, genes with related functions are often (but not always) located together in a group known as an *operon* (Figure 1.26). An operon has a single promoter, and is transcribed into a single polycistronic mRNA molecule, which carries the information for several proteins. This group of genes will therefore be coordinately controlled: growth of the cell under the appropriate conditions will induce all the genes in the operon simultaneously. This is discussed more fully in Chapter 3.

After the ribosome has translated the first cistron in an operon, it may dissociate, in which case translation of the next cistron will require attachment of ribosomes to another binding site adjacent to the initiation codon of the second cistron. In some cases the start codon for the second gene is very close to the stop codon of the first (in fact the sequences may actually overlap). If this occurs, then after the first polypeptide has been released, the ribosome may start translating again at the nearby start codon, without dissociating.

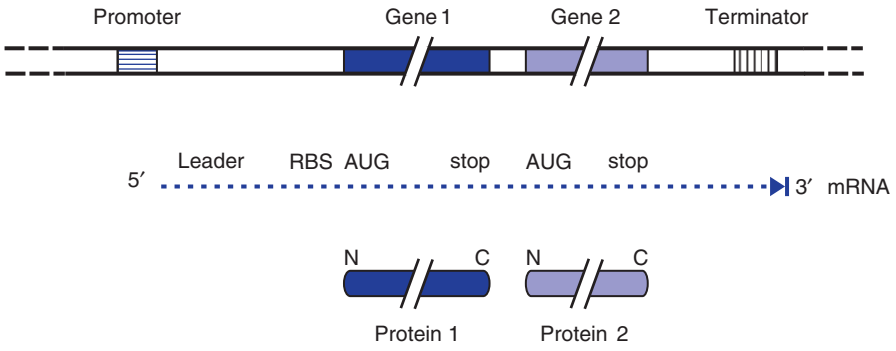


Figure 1.26 Structure and expression of a bacterial operon. A typical operon is transcribed, from a single promoter, into a polycistronic mRNA from which several independent polypeptides can be translated. RBS, ribosome-binding site.

Some major differences between bacteria and eukaryotes are worth noting. The mRNA in eukaryotes acts as a true 'messenger', being produced within the nucleus and migrating to the cytoplasm for translation to occur. In bacteria, transcription and translation occur in the same compartment, and the ribosomes will attach to the mRNA as soon as a RBS is available. So, in bacteria, the mRNA is being actively translated while it is still being made. Most bacterial mRNA is extremely short-lived; it typically has a half-life of a few minutes only, which may be less than the time required for producing or translating it. This can only be achieved by the coupling of transcription and translation. A further difference related to the mechanism of ribosome binding is that eukaryotic mRNA (in general) codes for a single polypeptide only. The ribosome in eukaryotic cells attaches to the 5′ end of the mRNA and migrates until it reaches the start codon. (Internal ribosome entry sites do exist in eukaryotes, but they are the exception rather than the rule.)

A further difference is that in eukaryotes, the initial product of transcription is a precursor of the mRNA. This precursor, which is found only in the nucleus, contains additional sequences (introns) that are removed by a process known as splicing or processing. In some cases the final size of the mRNA is less than 10% of that of the original gene. Generally, bacterial genes do not contain introns, but there are a few examples of prokaryotic genes (mainly from bacteriophages) that do contain introns.

Finally, eukaryotic mRNA is often (but not always) polyadenylated, i.e. it has a run of adenine residues at the 3′ end. The presence of poly-A tails is often used as the basis of procedures for purifying mRNA from eukaryotic cells. Bacterial mRNA, on the other hand, is not consistently polyadenylated, although a small proportion of bacterial RNA molecules may carry a short oligo-A tail.