1

Deoxyribonucleic acid (DNA)

1.1 The genetic material

The classic experiments of Avery in 1944 demonstrated that DNA (Deoxyribonucleic acid) passes genetic information from one bacterium to another. Strain-specific properties of related bacteria could be transferred by DNA that was free of proteins and other substances. DNA is a polymeric molecule built up from only four similar but distinct monomers — nucleotides that are the 5′-phosphates of deoxyguanosine (dGMP), deoxyadenosine (dAMP), deoxycytidine (dCMP), and thymidine (TMP) (Fig. 1.1), joined by phosphodiester linkages between the 3′- and 5′-positions of successive deoxyribose moieties. The initial letters of the bases in the nucleotides are used as abbreviations when writing out their sequence in DNA. The symbols N, R and Y denote any nucleotide, a purine nucleotide and a pyrimidine nucleotide respectively.

1.2 DNA is a polar helical molecule

One end of a DNA molecule has a phosphoryl radical on the C-5′ of its terminal nucleotide, while the other end possesses a free -OH on the C-3′ of its nucleotide. Thus a polynucleotide exhibits polarity in an analogous way to that of proteins with free -NH₂ and -COOH groups at their ends. The tetranucleotides TCGA and AGCT are different chemical entities with distinct properties, even though they behave very similarly in many respects (Fig. 1.2). By convention, sequences of DNA are written with the
2 Deoxyribonucleic acid (DNA)

![Chemical structures of deoxyribonucleotides](image)

**Fig. 1.1.** The four deoxyribonucleotides that make up DNA, showing how the bases pair by forming hydrogen bonds.

A nucleotide containing the free phosphoryl radical at the left. Sequences to the left of a given nucleotide are said to be on the 5' side (often called upstream), and those to the right are said to be on the 3' side (often called downstream).
1.2 DNA is a polar helical molecule

![DNA structure](image)

**Fig. 1.2.** Polarity in DNA. The two tetranucleotides TCGA (left) and AGCT (right) are different although they have the nucleotides in the same order.

DNA generally exists as a *duplex* in double strands because of the propensity of the bases to hydrogen bond to each other in a specific way (Fig. 1.1). Such pairs of nucleotides are known as *base-pairs* (bp; with kbp used as an abbreviation for 1000 bp). A bonds with T, and G with C, though very occasional mismatches or alternative bonding can occur. Thus a double-stranded DNA always contains equal molar proportions of A and T and of G and C though the content of A (or T) and G (or C) varies widely in DNA from different sources.

This double-stranded molecule takes up a helical conformation in which the continuous deoxyribose–phosphate strands twine round the outside of
4 Deoxyribonucleic acid (DNA)

Fig. 1.3. A 20 bp length of duplex DNA showing both B and Z forms. The lines running round the outside represent the backbone of polydeoxyribose phosphate, while the horizontal lines represent the edges of the base pairs in the interior of the molecule. (Reprinted, with permission, from S.B. Zimmerman, Annu. Rev. Biochem., 51, © 1982 by Annual Reviews Inc.)

the helix with the base-pairs in the interior (Fig. 1.3). The association of A and T with two hydrogen bonds is less stable than that of G and C which has three hydrogen bonds, so that in regions rich in A and T residues, the helix can be more easily destabilised and unwound than in G–C rich regions.

The polarity of the two DNA strands is anti-parallel since one runs in the 5' to 3' direction while the complementary strand runs the opposite way. The helix can adopt several conformations. The commonest form (B-DNA) has a pitch of just over ten residues per turn, and is right-handed when viewed end on. Z-DNA can arise under certain conditions when there are alternating purine and pyrimidine residues in the sequence. This is a left-handed helix with a pitch of 11.5 residues per turn (Fig. 1.3). This form may have some biological significance, since short stretches are sometimes found in regions that are involved in the control of the genome. Z-DNA can be detected by the binding of antibodies that specifically recognise it. A number of proteins, other than antibodies, that also react with Z-DNA are present in a wide range of cells.

In B-DNA there are two grooves with different dimensions running helically along it. The major groove (to the top of Fig. 1.1) is about 12 Å wide and 8.5 Å deep, while the minor groove (at the bottom of Fig. 1.1) is only 6 Å wide and 7.5 Å deep. The N and O atoms of the bases and the H atoms of the amino groups lining each groove can serve as hydrogen bond acceptors and donors for making specific contacts with appropriate atoms in DNA-binding proteins that control the replication and transcription of DNA.
1.3 Compact forms of DNA

Table 1.1. Chromosome numbers and DNA content of cells of a representative set of species

<table>
<thead>
<tr>
<th>Species</th>
<th>Number of chromosomes</th>
<th>DNA content kilobase-pairs</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Bacillus subtilis</em></td>
<td>1</td>
<td>$2 \times 10^3$</td>
</tr>
<tr>
<td><em>Escherichia coli</em></td>
<td>1</td>
<td>$3.8 \times 10^3$</td>
</tr>
<tr>
<td><em>Saccharomyces cerevisiae</em></td>
<td>34</td>
<td>$14 \times 10^3$</td>
</tr>
<tr>
<td><em>Drosophila melanogaster</em></td>
<td>8</td>
<td>$2 \times 10^5$</td>
</tr>
<tr>
<td>Sea urchin</td>
<td>52</td>
<td>$1.6 \times 10^6$</td>
</tr>
<tr>
<td>Frog</td>
<td>26</td>
<td>$45 \times 10^6$</td>
</tr>
<tr>
<td>Chicken</td>
<td>78</td>
<td>$2.1 \times 10^6$</td>
</tr>
<tr>
<td>Mouse</td>
<td>40</td>
<td>$4.7 \times 10^6$</td>
</tr>
<tr>
<td>Human</td>
<td>46</td>
<td>$5.6 \times 10^6$</td>
</tr>
<tr>
<td>Maize</td>
<td>20</td>
<td>$30 \times 10^6$</td>
</tr>
</tbody>
</table>

All figures, except those for bacteria, are for diploid cells.

(Chapters 4, 6 and 7). Since the overall diameter of the α-helix of a protein is about 12 Å, this can fit snugly into the major groove when there are suitable complementary atoms to make the hydrogen bonds between the two molecules.

In addition to hydrogen bonding between the complementary bases, individually weak stacking interactions between adjacent rings of the purine and pyrimidine bases may contribute significantly to the stability of the structure.

1.3 DNA molecules are very long but can be twisted into compact forms

A double helix of DNA with $10^9$ bp is 340 mm long and only 2 nm in diameter, and can be visualised using an electron microscope. In prokaryotes the DNA is circular so there are no free 3′- and 5′-ends, and all the chromosomal DNA is in a single molecule. In eukaryotes each chromosome contains a single linear molecule of DNA, and different species possess different numbers of chromosomes. The amount of DNA in their cells varies widely, generally with an increase in the DNA content as species become more complex (Table 1.1). Individual chromosomes are morphologically distinguishable when they are suitably stained. For purposes of identification they have been given numbers in order, starting with the largest one.
6 Deoxyribonucleic acid (DNA)

With the exception of the sex chromosome all eukaryotic chromosomes are paired, one partner coming from each parent. In mammals, females carry two X chromosomes, while males carry an X chromosome inherited from the mother and a Y chromosome from the father, but other methods of sex determination also occur. The two chromosomes of a pair are said to be homologous since they are nearly always identical in their organisation and frequently in the genes they carry. However, since there are many mutant genes in a population, a pair of homologous chromosomes may carry different genes at particular loci, known as alleles.

A mutant gene encoding a defective product can generally be complemented by a ‘good’ copy of the gene on the homologous chromosome, but if there is a defect on the single copy of the X chromosome in a male this is not possible. Thus there are many sex-linked diseases that are carried by females but expressed only in males. These disorders can actually occur in females but the chances of a female inheriting two defective genes are very low.

Since all somatic cells contain a homologous pair of each of the chromosomes, they are known as diploid. The gametes – sperm and ova – which only contain one member of each pair of chromosomes are referred to as haploid cells. The contribution of one parent to the genetic make-up of the offspring is known as the haplotype.

In cells, both linear and circular molecules of DNA occur in compacted forms. The helix is coiled on itself several times (like the element in an electric light bulb) so that the overall length is greatly reduced at the expense of an increase in diameter. This conformation is stabilised by proteins in eukaryotic cells (Chapter 6.2) resulting in the packaging of the DNA into a minimum of space. When portions of the DNA become functional there is some uncoiling of this structure accompanied by temporary separation of the two helical strands.

In a circular DNA molecule containing 4000 bp (such as might occur in a bacterial plasmid – Chapter 3.2) the double helix is in the B form. As this has a pitch of ten residues per turn there should be 400 turns, but the DNA has only about 380 turns because the helix is unwound to a certain extent. This is known as negative supercoiling, and is an important structural feature of DNA. It gives rise to a puckered form of the molecule so that it has a more compact structure, and also places considerable torsional strain on it. An analogy can be made by twisting a rubber band held firmly at two diametrically opposite positions. A molecule in which there is no supercoiling is said to be relaxed, and there is a dynamic balance between relaxed and supercoiled forms of DNA as a result of the action of two classes of enzymes called topoisomerases I and II that catalyse the production of one form from the other.
1.3 Compact forms of DNA

Fig. 1.4. The effect of supercoiling on the electrophoretic mobility of Simian virus 40 DNA, treated with a class II topoisomerase. The thick band at the left is the fully relaxed DNA, while molecules with increasing degrees of supercoiling appear as bands of increasing mobility. The arrow shows the direction of electrophoresis. (Reproduced from W. Keller, Proc. Natl. Acad. Sci., USA (1975), 72, 2550.)

Fig. 1.5. The action of DNA gyrase in forming a negative supercoil in a circular molecule of DNA. By convention, when the upper strand crosses over the lower strand from right to left the supercoiling is said to be positive. Negative supercoiling is the converse of this.

A supercoiled DNA molecule migrates more rapidly upon electrophoresis than does the relaxed form. A family of otherwise identical DNA molecules with different degrees of supercoiling can be made visible as a ladder of bands by this technique (Fig. 1.4).

DNA gyrase (a class II topoisomerase) brings about the formation of the supercoiled form in the cell. Since it is a strained structure, this requires an input of energy that is met by the hydrolysis of ATP. The reaction proceeds by breaking both strands of the DNA so that another part of the molecule can be passed through the break. The broken strands are then resealed (Fig. 1.5). There are several different type II topoisomerases in all organisms, each needed for specific reactions in which single-stranded DNA is required.

Class I topoisomerases relax supercoiled DNA. This requires no input of energy since a less strained molecule is produced, and involves nicking one strand of the DNA, when the 5'-phosphate at the break point becomes bound to a tyrosyl residue on the enzyme. This strand now has a free 3'-OH group and this end is rotated round the other intact strand. The nicked strand is then resealed (Fig. 1.6).
8 Deoxyribonucleic acid (DNA)

![Diagram of DNA structure and enzymatic reactions]

Fig. 1.6. Relaxation of supercoiled DNA by topoisomerase I.

Both these topoisomerases play important roles in replication and transcription (Chapters 4, 6 and 7), since each of these processes requires single-stranded DNA as a template. The DNA ahead of the replication or transcription complexes is positively supercoiled, and the type II topoisomerases that open it up are generally known as helicases. The DNA behind these complexes may be partially relaxed by the action of a type I topoisomerase.

1.4 Replication of DNA is semi-conservative

When a cell divides, each of the daughter cells contains a full complement of DNA, identical to that of each parent (except during the production of gametes in eukaryotes). The DNA is precisely replicated by the separation of the two strands, followed by pairing of deoxyribonucleoside triphosphates through specific hydrogen bonding with the bases in each strand. These nucleotides are joined together (ligated) by the enzyme DNA polymerase with the release of inorganic pyrophosphate (Fig. 1.7). Synthesis always proceeds from the 5'-end of the growing chain. In practice the two parental strands do not separate completely, but are opened up at the replication fork that is moved along as the process proceeds.

Because of the opposite polarity of the parental strands, only one strand (the leading strand) is in the correct orientation for continuous synthesis of the new strand. On the other (lagging) strand, that is incorrectly oriented, comparatively short lengths of new DNA are formed (named Okazaki fragments after their discoverer), which are then joined by a ligase (Fig. 1.8). This process of replication is semi-conservative since one strand is newly synthesised, while the other is from the parental cell. This outline of replication
1.5 The functional unit of DNA

Replication fork, showing strand separation and incorporation of nucleotides. Continuous synthesis takes place on the lower strand. Discontinuous synthesis on the upper strand does not start so near the replication fork.

may sound simple, but the actual process is extremely complex and not yet completely understood. In *E. coli*, where it has been most intensively studied, at least 14 different polypeptides, some of which are enzymes, are known to be involved. There are further details in Chapters 4 and 6.

1.5 The gene or cistron is the functional unit of DNA

Replication ultimately involves the whole DNA molecule, but the functional unit of DNA is much smaller than this – perhaps only a few thousand or even hundred base-pairs. These units correspond to the original concept of genes postulated by earlier geneticists who associated them with various variable (phenotypic) characters that are inherited predictably. A more recent term with a very similar meaning is *cistron*, but this is not widely used, and we still talk of genes. The complete collection of genes in an organism is called the *genome*.

Genes contain information that directs the synthesis of other molecules in a highly specific way. Although only four different nucleotides are used in building a DNA molecule, the number of possible arrangements is extremely large. In a length of DNA containing 100 nucleotides there are $4^{100} (1.7 \times 10^{60})$
10 Deoxyribonucleic acid (DNA)

Fig. 1.8. Continuous and discontinuous synthesis of duplex DNA during replication. Note that the scale is much smaller than in Fig. 1.7.

Possible sequences. As the genetic material of any organism contains millions of nucleotides there is obviously room to account for all the amazingly wide diversity of living organisms.

Most of the information encoded in the genome is used to specify the sequence of amino acids that are made into proteins to serve the wide variety of essential functions within the cell. The expression of this information is mediated by molecules of ribonucleic acid (RNA), formed on a template of DNA by the process called transcription (see Chapter 2), but some RNA is used directly for various specific purposes.

DNA sequences flanking both ends of those that encode information for making RNAs also have important functions in regulating the activity of genes. In prokaryotes, factors such as the availability of metabolites determine which genes are active at any particular time. In metazoans there is generally specialisation of cells of various types so that some genes in certain tissues are permanently inactive (‘switched off’), while others may be activated by particular signals in the local environment. There are also a number of pseudogenes scattered throughout most genomes, with very considerable sequence homology to functional genes from which they have probably been derived.