We begin our study of genetics with DNA. We start with DNA because genes are made of DNA, or, to be more precise, DNA is the **genetic material**. By studying the structure of DNA we can immediately understand how genes are able to fulfill their two related functions in living organisms, as units of biological information and as units of inheritance. This is the **molecular approach** to genetics, which we will be following in this book. The molecular approach enables us to make a logical, step-by-step progression from the structure of DNA to the process by which the biological information contained in DNA is released to the cell when it is needed. An understanding of DNA structure is also the best starting point for investigating how DNA molecules are copied and passed to offspring during reproduction, and for understanding how DNA molecules can change over time, enabling evolution to take place. The most accessible route into the more complex intricacies of genetics therefore starts with the structure of DNA.

In this chapter, then, we will study the structure of DNA and ask how this structure enables DNA to carry out the two functions of genes. We will learn how biological information is encoded in DNA and how copies can be made of DNA molecules. We will also examine how scientists can read the biological information contained in an individual DNA molecule by DNA sequencing, one of the most important techniques in modern biological research.
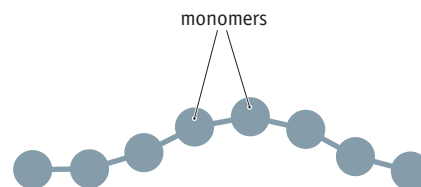
FIGURE 2.1 DNA is a linear polymer. In this depiction, each bead in the chain is an individual monomer.

2.1 THE STRUCTURE OF DNA

DNA (**deoxyribonucleic acid**) is a **polymer**, a long, chainlike molecule made up of subunits called **monomers** (**Figure 2.1**). Many important biological molecules, including not only DNA but also proteins, polysaccharides, and lipids, are polymers of one type or another. In DNA, the monomers are called **nucleotides**, and these are linked together to form a **polynucleotide** chain that can be hundreds, thousands, or even millions of nucleotides in length. First we will study the structure of a nucleotide, and then we will examine how nucleotides are joined together to form a polynucleotide.

Nucleotides are the basic units of a DNA molecule

The basic unit of the DNA molecule is the nucleotide. Nucleotides are found in the cell either as components of nucleic acids or as individual molecules. Nucleotides have several different roles and are not just used to make DNA. For example, some nucleotides are important in the cell as carriers of energy used to power enzymatic reactions.

The nucleotide is itself quite a complex molecule, being made up of three distinct components. These are a sugar, a nitrogenous base, and a phosphate group (**Figure 2.2**). We will look at each of these in turn.
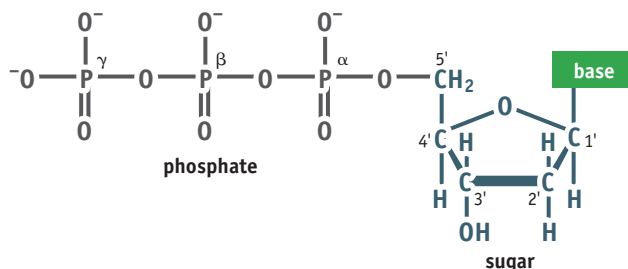
FIGURE 2.2 The components of a deoxyribonucleotide.

Fischer structure

Haworth structure
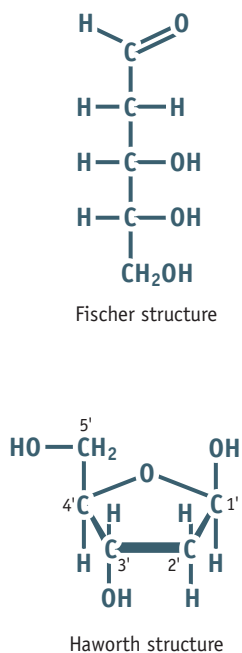
FIGURE 2.3 The two structural forms of 2'-deoxyribose.

The sugar component of the nucleotide is a **pentose**. A pentose is a sugar that contains five carbon atoms. The particular type of pentose present in the nucleotides found in DNA is called 2'-deoxyribose. Pentose sugars can exist in two forms: the straight chain, or Fischer structure; and the ring, or Haworth structure (Figure 2.3). It is the ring form of 2'-deoxyribose that occurs in the nucleotide.

The name 2'-deoxyribose indicates that the standard ribose structure has been altered by replacement of the hydroxyl group (–OH) attached to carbon atom number 2' with a hydrogen group (–H). The carbon atoms are always numbered in the same way, with the carbon of the carbonyl group (–C=O), which occurs at one end of the chain structure, numbered 1'. It is important to remember the numbering of the carbons because it is used to indicate the positions at which other components of the nucleotide are attached to the sugar. It is also important to realize that the numbers are not just 1, 2, 3, 4, and 5, but 1', 2', 3', 4', and 5'. The upper-right stroke is called a "prime," and the numbers are called "one-prime," "two-prime," and so on. The prime is used to distinguish the carbon atoms in the sugar from the carbon and nitrogen atoms in the nitrogenous base, which are numbered 1, 2, 3, and so on.

The nitrogenous bases are single- or double-ring structures that are attached to the 1' carbon of the sugar. In DNA any one of four different nitrogenous bases can be attached at this position. These are called **adenine** and **guanine**, which are double-ring **purines**, and **cytosine** and **thymine**, which are single-ring **pyrimidines**. Their structures are shown in Figure 2.4. The base is attached to the sugar by a **β-*N*-glycosidic bond** attached to nitrogen number 1 of the pyrimidine or number 9 of the purine.

A molecule comprising the sugar joined to a base is called a **nucleoside**. This is converted into a nucleotide by attachment of a phosphate group to the 5' carbon of the sugar. Up to three individual phosphates can be attached in series. The individual phosphate groups are designated α, β, and γ, with the α phosphate being the one attached directly to the sugar.

The full names of the four different nucleotides that polymerize to form DNA are:
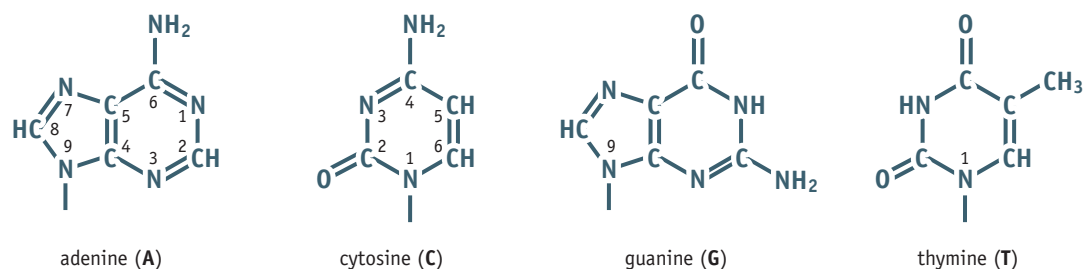
     2'-deoxyadenosine 5'-triphosphate

     2'-deoxyguanosine 5'-triphosphate

     2'-deoxycytidine 5'-triphosphate

     2'-deoxythymidine 5'-triphosphate

Normally, however, we abbreviate these to dATP, dGTP, dCTP, and dTTP, or even just to A, G, C, and T, especially when writing out the sequence of nucleotides found in a particular DNA molecule.

FIGURE 2.4 The structures of the four bases that occur in deoxyribonucleotides.

adenine (**A**)

cytosine (**C**)

guanine (**G**)

thymine (**T**)

Nucleotides join together to make a polynucleotide

The next stage in building up the structure of a DNA molecule is to link the individual nucleotides together to form a polymer. This polymer is called a polynucleotide and is formed by attaching one nucleotide to another through the phosphate groups.

The structure of a trinucleotide, a short DNA molecule comprising three individual nucleotides, is shown in Figure 2.5. The nucleotide monomers are linked together by joining the α phosphate group, attached to the 5′ carbon of one nucleotide, to the 3′ carbon of the next nucleotide in the chain. Normally a polynucleotide is built up from nucleoside triphosphate subunits, so during polymerization the β and γ phosphates are cleaved off. The hydroxyl group attached to the 3′ carbon of the second nucleotide is also lost. The linkage between the nucleotides in a polynucleotide is called a **phosphodiester bond**, "phospho-" indicating the presence of a phosphorus atom and "diester" referring to the two ester (C–O–P) bonds in each linkage. To be precise we should call this a 3′-5′ phosphodiester bond so that there is no confusion about which carbon atoms in the sugar participate in the bond.

An important feature of the polynucleotide is that the two ends of the molecule are not the same. This is clear from an examination of Figure 2.5. The top of this polynucleotide ends with a nucleotide in which the triphosphate group attached to the 5′ carbon has not participated in a phosphodiester bond and the β and γ phosphates are still in place. This end is called the **5′-P terminus** or simply the **5′ terminus**. At the other end of the molecule the unreacted group is not the phosphate but the 3′ hydroxyl. This end is called the **3′-OH** or **3′ terminus**. The chemical distinction between the two ends means that polynucleotides have a direction, which can be looked on as 5′ → 3′ (down in Figure 2.5) or 3′ → 5′ (up in Figure 2.5). An important consequence of the polarity of the phosphodiester bond is that the chemical reaction needed to extend a DNA polymer in the 5′ → 3′ direction is different from that needed to make a 3′ → 5′ extension. All of the enzymes that make new DNA polynucleotides in living cells carry out 5′ → 3′ synthesis. No enzymes that are capable of catalyzing the chemical reaction needed to make DNA in the opposite direction, 3′ → 5′, have ever been discovered.
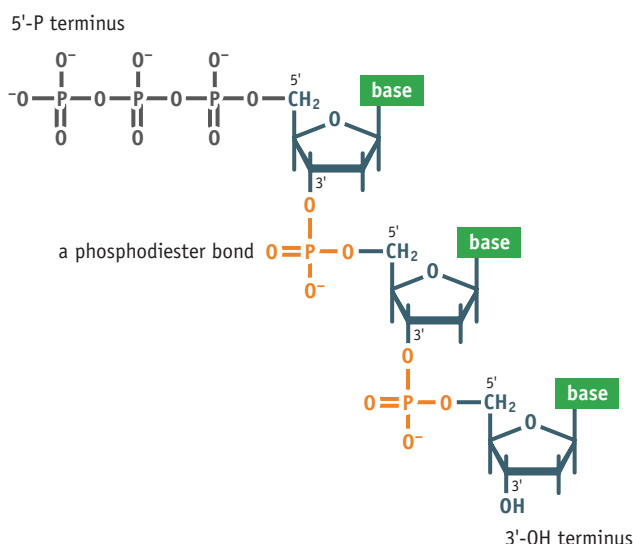

FIGURE 2.5 The structure of a trinucleotide, a DNA molecule comprising three individual nucleotides.

There is apparently no limitation to the number of nucleotides that can be joined together to form an individual DNA polynucleotide. Molecules containing several thousand nucleotides are frequently handled in the laboratory, and the DNA molecules in chromosomes are much longer, sometimes several million nucleotides in length. In addition, there are no chemical restrictions on the order in which the nucleotides can join together. At any point in the chain the nucleotide could be A, G, C, or T.

In living cells, DNA is a double helix

In living cells, DNA molecules almost always contain two polynucleotides, wrapped around one another to form the famous double-helix structure discovered by James Watson and Francis Crick in 1953 (Research Briefing 2.1). The double helix is a complicated structure (**Figure 2.6**), but the key facts about it are not too difficult to understand.

First we will look at the basic features of the double helix. The two polynucleotides in the helix are arranged in such a way that their sugar–phosphate "backbones" are on the outside of the helix and their bases are on the inside. The bases are stacked on top of each other rather like a pile of plates. The two polynucleotides are **antiparallel**, meaning that they run in different directions, one oriented in the 5′ → 3′ direction and the other in the 3′ → 5′ direction. The polynucleotides must be antiparallel in order to form a stable helix. Molecules in which the two polynucleotides run in the same direction are unknown in nature. The double helix is right-handed, so if it were a spiral staircase then the banister (that is, the sugar–phosphate backbone) would be on your right-hand side as you were climbing upward. The final point is that the helix is not absolutely regular. Instead, on the outside of the molecule we can distinguish a **major** and a **minor groove**. These two grooves are clearly visible in Figure 2.6.
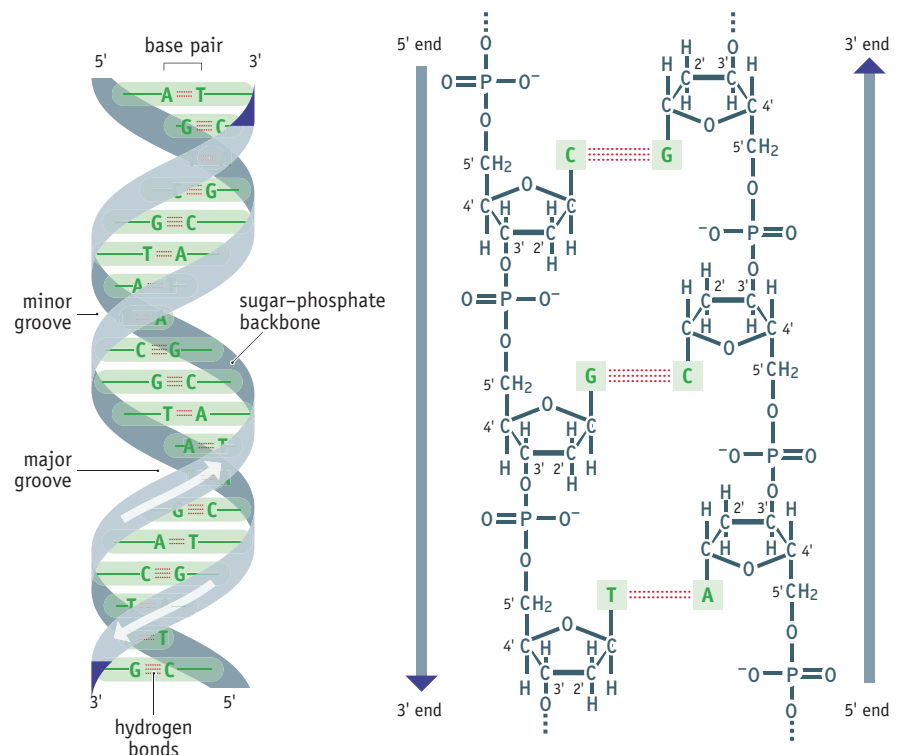
FIGURE 2.6 The double-helix structure of DNA. On the left the double helix is drawn with the sugar–phosphate backbone of each polynucleotide shown as a gray ribbon with the base pairs in green. On the right the chemical structure for three base pairs is given.
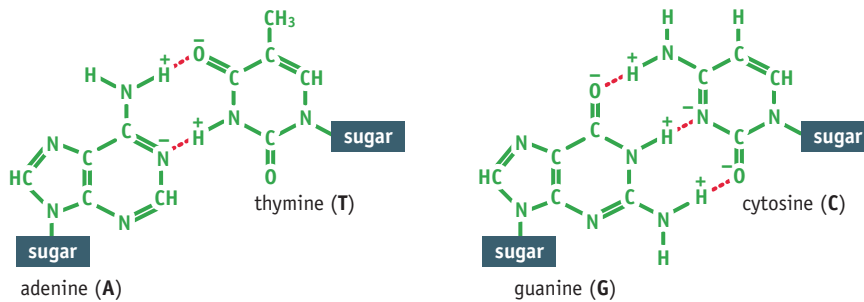
FIGURE 2.7 Base pairing. A base-pairs with T, and G base-pairs with C. The bases are drawn in outline, with the hydrogen bonds indicated by dotted lines. Note that a G–C base pair has three hydrogen bonds whereas an A–T base pair has just two.

Those are the key facts about the double helix. There remains just one additional feature to consider, but this one is the most important of all. Within the helix, an adenine in one polynucleotide is always adjacent to a thymine in the other strand, and similarly guanine is always adjacent to cytosine. This is called **base pairing** and involves the formation of **hydrogen bonds** between an adenine and a thymine, or between a cytosine and a guanine. A hydrogen bond is a weak electrostatic attraction between an electronegative atom (such as oxygen or nitrogen) and a hydrogen atom attached to a second electronegative atom. The base pairing between adenine and thymine involves two hydrogen bonds, and that between guanine and cytosine involves three hydrogen bonds (**Figure 2.7**).

The two base-pair combinations—A base-paired with T, and G base-paired with C—are the only ones that are permissible. This is partly because of the geometries of the nucleotide bases and the relative positions of the atoms that are able to participate in hydrogen bonds, and partly because the pair must be between a purine and a pyrimidine. A purine–purine pair would be too big to fit within the helix, and a pyrimidine–pyrimidine pair would be too small. Because of the base pairing, the sequences of the two polynucleotides in the helix are **complementary**—the sequence of one polynucleotide determines the sequence of the other.

The double helix exists in several different forms

The double helix shown in Figure 2.6 is referred to as the **B-form** of DNA. This is the structure that was studied by Watson and Crick, and it represents by far the most common type of DNA in living cells, but it is not the only version that is known. Other versions are possible because the three-dimensional structure of a nucleotide is not entirely rigid. For example, the orientation of the base relative to the sugar can be changed by rotation around the β-N-glycosidic bond (**Figure 2.8**). This has a significant effect on the double helix, as it alters the relative positioning of the two polynucleotides. The relative positions of the carbons in the sugar can also be changed slightly, affecting the conformation of the sugar–phosphate backbone.

anti-adenosine   syn-adenosine
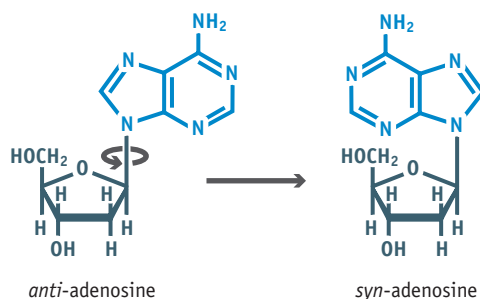
FIGURE 2.8 The structures of *anti-* and *syn-*adenosine, which differ in the orientation of the base relative to the sugar component of the nucleotide. Rotation around the β-N-glycosidic bond can convert one form into the other. The three other nucleotides also have *anti* and *syn* conformations. In A- and B-DNA, all the bases are in the *anti* conformation, but in Z-DNA there is a mixture of the two types.

RESEARCH BRIEFING 2.1

The discovery of the double helix

The discovery of the double helix, by James Watson and Francis Crick of Cambridge University, UK, in 1953, was the most important breakthrough in twentieth-century biology. The nature of the double helix revealed how genes can replicate, a puzzle that had seemed almost impossible to solve just a few years earlier. According to Watson in his book *The Double Helix*, the work was a desperate race against the famous American biochemist Linus Pauling, who initially thought that DNA was a triple helix. This mistake gave Watson and Crick the time they needed to complete their description of the double-helix structure.

When Watson and Crick began their work, the structures of the nucleotides, and the way these are linked together to form a polynucleotide, were already known. What was not known was the actual structure of DNA in a living cell. Was it a single polynucleotide, perhaps folded up in some way? Or were there two or more polynucleotides in a DNA molecule?

To solve the structure of DNA, Watson and Crick used **model building**—they built a scale model of what they thought a DNA molecule must look like. The model had to obey the laws of chemistry, which meant that if a polynucleotide was coiled in any way then its atoms must not be placed too close together. It was equally vital that the model take account of the results of other investigations into DNA structure. One of these studies was carried out by Erwin Chargaff at Columbia University in New York, the other, by Rosalind Franklin at King's College, London.

Chargaff's base ratios paved the way for the correct structure

Erwin Chargaff became interested in DNA in the 1940s, when scientists first realized that DNA *might* be the genetic material. He decided to use a new technique, called **paper chromatography**, to measure the amounts of each of the four nucleotides in DNA from different tissues and organisms. In paper chromatography, a mixture of compounds is placed at one end of a paper strip, and an organic solvent, such as *n*-butanol, is then allowed to soak along the strip. As the solvent moves it carries the compounds with it, but at different rates depending on how strongly each one absorbs into the paper matrix (**Figure 1**).

Chargaff purified DNA from different sources and treated each sample with acid to break the molecules into their component nucleotides (**Figure 2**). He then used paper chromatography to separate the nucleotides in each mixture so their concentrations could
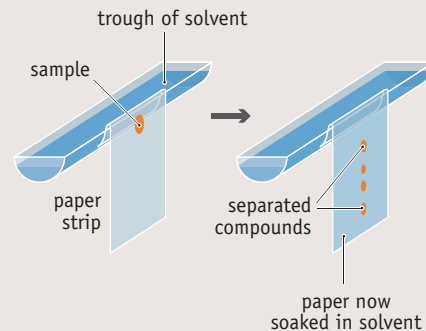

FIGURE 1 Paper chromatography.

be measured. The results were quite startling. They revealed a simple relationship between the proportions of the nucleotides in any one sample of DNA. The relationship is that the number of adenines equals the number of thymines, and the number of guanines equals the number of cytosines. In other words, A = T and G = C.

Chargaff did not speculate to any great extent, at least not in his publications, about the relevance of these base ratios to the structure of DNA. In fact, Watson and Crick appear to have been unaware of his results until they met Chargaff when he visited Cambridge University. But once they became aware of the A = T and G = C relationship, they knew this had to be accounted for in their model of DNA.
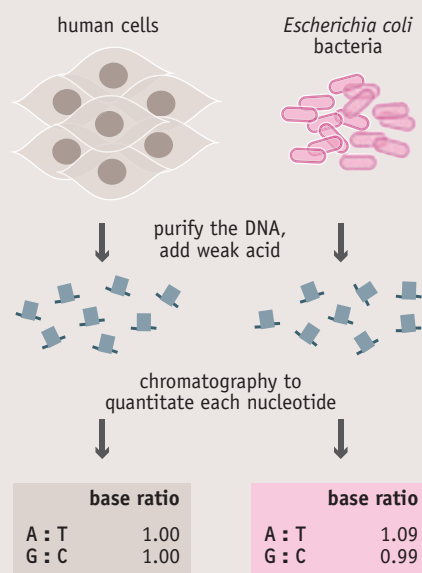

human cells		*Escherichia coli* bacteria	
base ratio		**base ratio**	
A : T	1.00	A : T	1.09
G : C	1.00	G : C	0.99

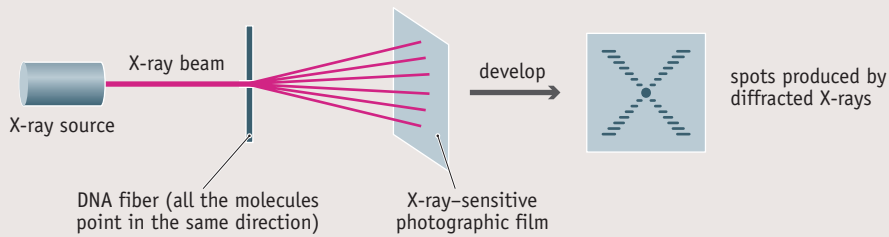**FIGURE 2** Chargaff's experiments.

X-ray diffraction analysis indicates that DNA is a helical molecule

The second piece of evidence available to Watson and Crick was the **X-ray diffraction pattern** obtained when a DNA fiber is bombarded with X-rays. X-rays have very short wavelengths—between 0.01 and 10 nm—comparable with the spacings between atoms in chemical structures. When a beam of X-rays is directed onto a DNA fiber some of the X-rays pass straight through, but others are diffracted and emerge at a different angle (**Figure 3**). As the fiber is made up of many DNA molecules, all positioned in a regular array, the individual X-rays are diffracted in similar ways, resulting in overlapping circles of diffracted waves which interfere with one another. An X-ray–sensitive photographic film placed across the beam reveals a series of spots and smears, called the X-ray diffraction pattern.

X-ray diffraction pictures of DNA fibers were made by Rosalind Franklin, during 1952, using techniques previously developed by Maurice Wilkins (**Figure 4**). The pictures immediately showed that DNA is a helix, and mathematical calculations based on them revealed that the helix has two regular periodicities of 0.34 nm and 3.4 nm. But how do these deductions relate to Chargaff's base ratios and to the actual structure of DNA?

Pulling together the evidence

Watson and Crick put together all the experimental data concerning DNA and decided that the only structure that fitted all the facts was the double helix shown in Figure 2.6. The main difficulty was deciding how many polynucleotides were present in a single molecule. This could be estimated from the density of the DNA in a fiber. Several measurements of DNA fiber density had been reported, and they did not agree. Some suggested that there were three polynucleotides in a single molecule, others suggested two. Pauling thought the first set of measurements were correct, and devised a triple-helix structure that was completely wrong. Watson and Crick decided it was more likely to be two.

Once two polynucleotides had been decided on, it became clear that the sugar–phosphate backbone had to be on the outside of the molecule. This was the only way the various atoms could be spaced out appropriately within the models that Watson and Crick built. The models also indicated that the two strands had to be antiparallel and the helix right-handed. The X-ray diffraction data enabled the dimensions of the helix to be set. The periodicity of 0.34 nm indicated the spacing between individual base pairs, and that of 3.4 nm gave the distance needed for a complete turn of the helix.

What about Chargaff's base ratios? These were the key to solving the structure. Watson realized, on the morning of Saturday, March 7, 1953, that the pairs formed by adenine–thymine and guanine–cytosine have almost identical shapes (see Figure 2.7). These pairs would fit neatly inside the double helix, giving a regular spiral with no bulges. And if these were the only pairs that were allowed, then the amount of A would equal the amount of T, and G would number the same as C. Everything fell into place, and the greatest mystery of biology—how genes can replicate—had been solved.
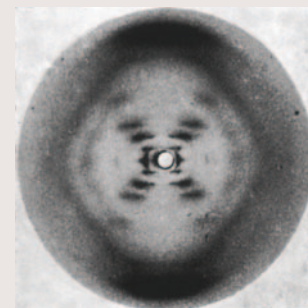

FIGURE 4 Franklin's "photo 51" showing the diffraction pattern obtained with a fiber of DNA. The cross shape indicates that DNA has a helical structure, and the relative positions of the various dots and smears enable the periodicities within the molecule to be calculated. (From R. Franklin and R. G. Gosling, *Nature* 171: 740–741, 1953. With permission from Macmillan Publishers Ltd.)

A-DNA **B-DNA**
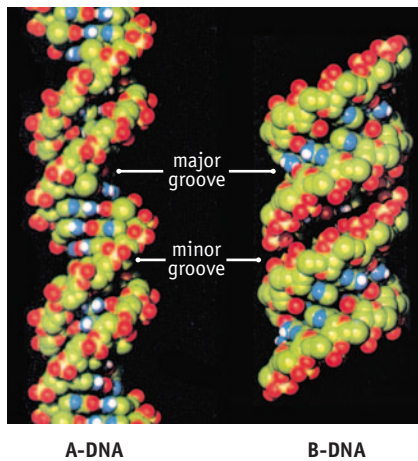
FIGURE 2.9 Comparison between the A- and B-DNA versions of the double helix. These are space-filling models in which the individual atoms are represented by spheres. The atoms of the sugar–phosphate backbone are shown in green and red, and those in the bases in blue and white. (From J. Kendrew (ed.) Encyclopedia of Molecular Biology. Oxford: Blackwell, 1994. With permission from Wiley–Blackwell.)

TABLE 2.1 FEATURES OF THE DIFFERENT CONFORMATIONS OF THE DNA DOUBLE HELIX

Feature	B-DNA	A-DNA	Z-DNA
Type of helix	Right-handed	Right-handed	Left-handed
Number of base pairs per turn	10	11	12
Distance between base pairs (nm)	0.34	0.29	0.37
Distance per complete turn (nm)	3.4	3.2	4.5
Diameter (nm)	2.37	2.55	1.84
Major groove	Wide, deep	Narrow, deep	Flat
Minor groove	Narrow, shallow	Wide, shallow	Narrow, deep

The common, B-form of DNA has 10 base pairs (abbreviated as 10 bp) per turn of the helix, with 0.34 nm between adjacent base pairs and hence a pitch (the distance needed for one complete turn) of 3.40 nm. The diameter of the helix is 2.37 nm. The second type of DNA double helix to be discovered, called the **A-form**, is more compact, with 11 bp per turn, 0.29 nm between each two base pairs, and a diameter of 2.55 nm (**Table 2.1**). But comparing these dimensions does not reveal what is probably the most significant difference between these two conformations of the double helix. This is the extent to which the internal regions of the DNA molecule are accessible from the surface of the structure. A-DNA, like the B-form, has two grooves, but with A-DNA the major groove is even deeper, and the minor groove shallower and broader (**Figure 2.9**).

In Chapter 9 we will learn that expression of the biological information contained within a DNA molecule is mediated by DNA-binding proteins that attach to the double helix and regulate the activity of the genes contained within it. To carry out their function, each DNA-binding protein must attach at a specific position, close to the gene whose activity it must influence. The protein can achieve this, with at least some degree of accuracy, by reaching down into a groove, within which it can "read" the DNA sequence without opening up the helix by breaking the base pairs (**Figure 2.10**). This is possible with B-DNA, but easier with A-DNA because the bases are more exposed in the minor groove. It has been suggested that some DNA-binding proteins induce formation of a short stretch of A-DNA when they attach to a DNA molecule.
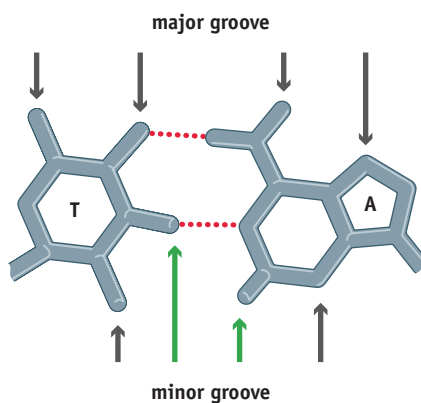
FIGURE 2.10 A DNA-binding protein can recognize individual base pairs without opening up the double helix. This drawing shows an A–T base pair in outline, with arrows indicating the chemical features that can be recognized by accessing the base pair via the major groove (above) and minor groove (below). Using these features, a binding protein can recognize that the base pair is an A–T rather than a G–C. In the major groove the distribution of these chemical features also enables the orientation of the base pair to be identified, so the strand containing the A can be distinguished from the one containing the T. For some time it was believed that this was not possible in the minor groove, because only the two features indicated by the black arrows were known to be present. The two additional features indicated by the green arrows have recently been discovered, and these are thought to enable the orientation of the pair to be discernible via the minor groove.

FIGURE 2.11 The structure of the Z-DNA version of the double helix. The atoms of the sugar–phosphate backbone are shown in green and red, and those in the bases in blue and white. (From J. Kendrew (ed.) Encyclopedia of Molecular Biology. Oxford: Blackwell, 1994. With permission from Wiley–Blackwell.)
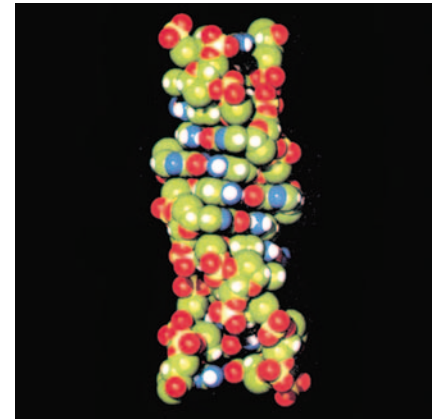
A third type, **Z-DNA**, is more strikingly different. In this structure the helix is left-handed, not right-handed as it is with A- and B-DNA, and the sugar–phosphate backbone adopts an irregular zigzag conformation (**Figure 2.11**). Z-DNA is more tightly wound, with 12 bp per turn and a diameter of only 1.84 nm (Table 2.1). It is thought to form around regions of B-DNA that have become slightly unwound, as occurs when a gene is being transcribed into RNA. Unwinding results in torsional stress, which might be relieved to some extent by forming the more compact Z version of the helix (**Figure 2.12**).

2.2 THE MOLECULAR EXPLANATION OF THE BIOLOGICAL ROLE OF DNA

To fulfill its role as the genetic material, a DNA molecule must possess properties that enable the genes it contains to act as units of biological information and units of inheritance. We will now ask ourselves how these requirements are met by the structure of the double helix.

Biological information is contained in the nucleotide sequence of a DNA molecule

DNA is able to act as a store of biological information because of its polymeric structure and because there are four different nucleotides. The order of nucleotides in a DNA molecule—the **DNA sequence**—is, in essence, a *language* made up of the four letters A, C, G, and T. The biological information contained in genes is written in this language, which we call the **genetic code**. The language is read through the process called **gene expression**.

There are no chemical restrictions on the order in which the nucleotides can join together in a DNA molecule. At any point the nucleotide could be A, G, C, or T. This means that a polynucleotide just 10 nucleotides in length could have any one of 4^{10} = 1,048,576 different sequences (**Figure 2.13**). The average length of a gene is about 1000 nucleotides. This length of DNA can exist as 4^{1000} different sequences, which we will look on as simply a very big number, greater than the supposed number of atoms in the observable universe, which is a paltry 10^{80}. Bear in mind that we can also have genes 999 and 1001 nucleotides in length, as well as many other lengths, each of these lengths providing its own immense number of possible DNA sequence variations. The early geneticists at the start of the twentieth century were mystified by the ability of genes to exist in many different forms,

unwinding creates torsional stress

Z-DNA region   Z-DNA region
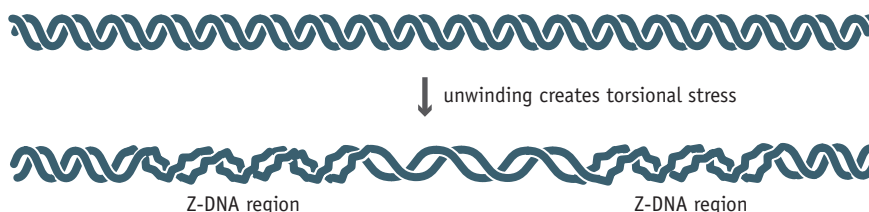
FIGURE 2.12 Regions of Z-DNA might form on either side of an underwound segment of B-DNA, in order to relieve the torsional stress that is created.

A G C T A A G G G T
A C C T A A G G G T
A A C T A A G G G T
A T C T A A G G G T
A G G T A A G G G T
A G T T A A G G G T
A C G T A A G G G T
A C G A A A G G G T
A G C T T C G G G T
A G C T G G G G G T
A G C T A A G G G A
T C A A T T T T A A

FIGURE 2.13 Twelve of the 1,048,576 different possible sequences for a DNA molecule 10 nucleotides in length. The orange nucleotides are ones that are different from the equivalent nucleotide in the topmost sequence.

in order to account for all those present in the myriad of species alive today and which lived in the past. An answer was so difficult to imagine that some biologists wondered whether genes really were physical structures inside cells. Perhaps they were just abstract entities whose invention by geneticists made it possible to explain how biological characteristics are passed from parents to offspring. Now that we understand the structure of DNA the immense variability required by the genetic material is no puzzle at all.

Complementary base pairing enables DNA molecules to replicate

In order to act as units of inheritance, genes must be able to replicate. Copies of the parent's genes must be placed in the fertilized egg cell so that this cell receives the information it needs to develop into a new living organism that displays a mixture of the biological characteristics of its parents. Genes must also be replicated every time a cell divides, so that each of the two daughter cells can be given a complete copy of the biological information possessed by the parent cell.

Before the structure of the double helix was known, the ability of genes to make copies of themselves was an even greater mystery than their ability to exist in an almost infinite number of different forms. No chemical capable of replication had ever been found in the natural world. The discovery of the double helix provided a clear and obvious solution to the problem. The key lies with the complementary base pairing that links nucleotides that are adjacent to one another in the two strands of the double helix. The rules are that A can base-pair only with T, and G can base-pair only with C. This means that if the two polynucleotides in a DNA molecule are separated, then two perfect copies of the parent double helix can be made simply by using the sequences of these preexisting strands to dictate the sequences of the new strands (**Figure 2.14**).

In living cells, DNA molecules are copied by enzymes called **DNA polymerases**. A DNA polymerase builds up a new polynucleotide by adding nucleotides one by one to the 3' end of the growing strand, using the base-pairing rules to identify which of the four nucleotides should be added at each position (**Figure 2.15**). This is called **template-dependent DNA synthesis**. It ensures that the double helix that is made is a precise copy of
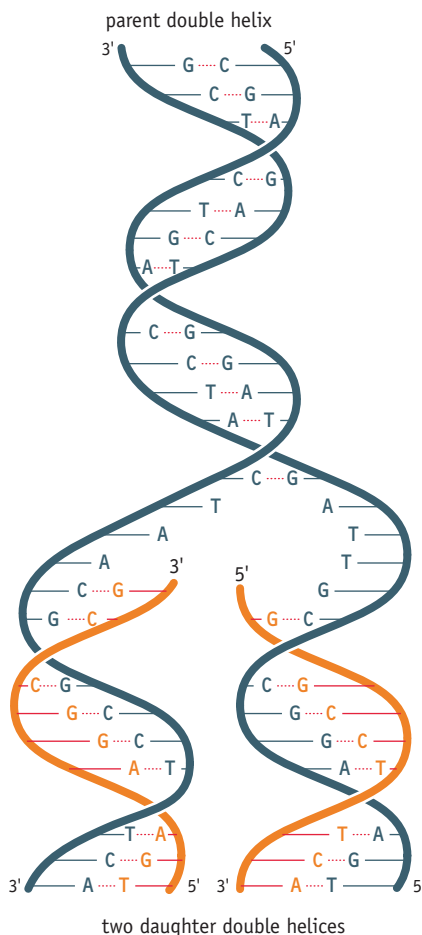
FIGURE 2.14 Complementary base pairing provides a means for making two copies of a double helix. The polynucleotides of the parent double helix are shown in blue, and the newly synthesized strands are in orange.

FIGURE 2.15 Template-dependent DNA synthesis. The new DNA, shown in orange, is extended by adding nucleotides one by one to its 3' end.

the double helix from which the original polynucleotide was obtained. The structure of DNA therefore explains how genes are able to replicate and hence to act as units of inheritance.

2.3 HOW DNA IS SEQUENCED

Now that we understand the structure of DNA we can begin to appreciate why **DNA sequencing** is so important in modern biology. By working out the sequence of nucleotides we can gain access to the biological information contained in a DNA molecule. From the sequence it might be possible to identify the genes present in the DNA molecule, and possibly to deduce the functions of those genes. We should therefore complete our study of DNA by examining the methods used to obtain nucleotide sequences.

The DNA molecules present in living cells are extremely long, possibly many millions of base pairs in length. It is not yet possible to work out the sequence of a molecule of this length in a single experiment. It might be possible one day, as DNA sequencing technology is advancing by leaps and bounds, but at present this is far beyond our capabilities. In fact, the maximum amount of sequence that can be read in a single experiment is just 750 bp. The strategy in a sequencing project is therefore to break these long DNA molecules down into many short fragments, each of which is individually sequenced. The sequence of the original molecule is then assembled by searching for overlaps between the sequences of the short fragments (**Figure 2.16**). Immediately we can understand why big sequencing projects make use of automated techniques, with the sequencing experiments carried out by robots, and computer programs used to find the overlaps and build up the contiguous blocks of sequence, or **contigs**.

### Clone libraries contain DNA fragments ready to be sequenced

The first step in a DNA sequencing project is to prepare a collection of short DNA fragments ready for sequencing. This is done by **DNA cloning**, and the resulting collection is called a **clone library**.

DNA cloning is a sophisticated technology, but we do not need to delve too deeply into it in order to understand how it is used to prepare a clone library. We will learn much more about it when we study the industrial applications of genetics in Chapter 22. First, we must break down the DNA molecules that we wish to sequence into the shorter fragments from which the library will be prepared. Often the DNA is fragmented by **sonication**, a technique that uses high-frequency sound waves to make random cuts in the molecules (**Figure 2.17**).

The next task is to separate the fragments resulting from sonication so that each one can be sequenced individually. This is the actual cloning part of the procedure, and it requires a **vector**, a DNA molecule that is able to replicate inside a cell of the bacterium *Escherichia coli*. The vector is usually a modified version of a naturally occurring, self-replicating DNA molecule called a **plasmid**. In its replicative state the plasmid is a circular molecule, but at the start of the cloning experiment it is obtained in a linear form. The linear vector molecules are mixed with the DNA fragments, and a **DNA ligase** enzyme is added, which joins DNA molecules together end-to-end. Various combinations are produced, including some in which a vector molecule has become linked to one of the DNA fragments and the two ends have then joined together to produce a circular **recombinant** vector (**Figure 2.18A**).
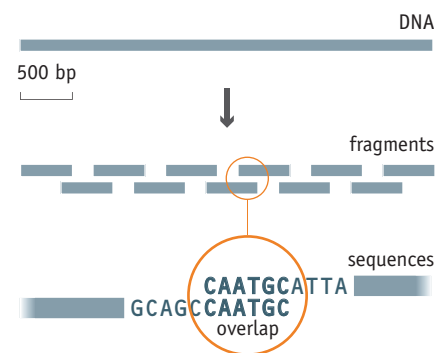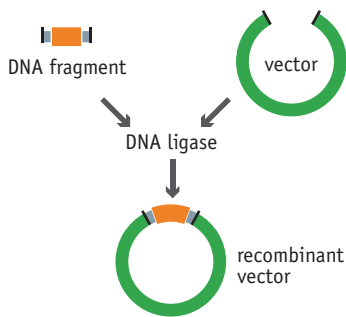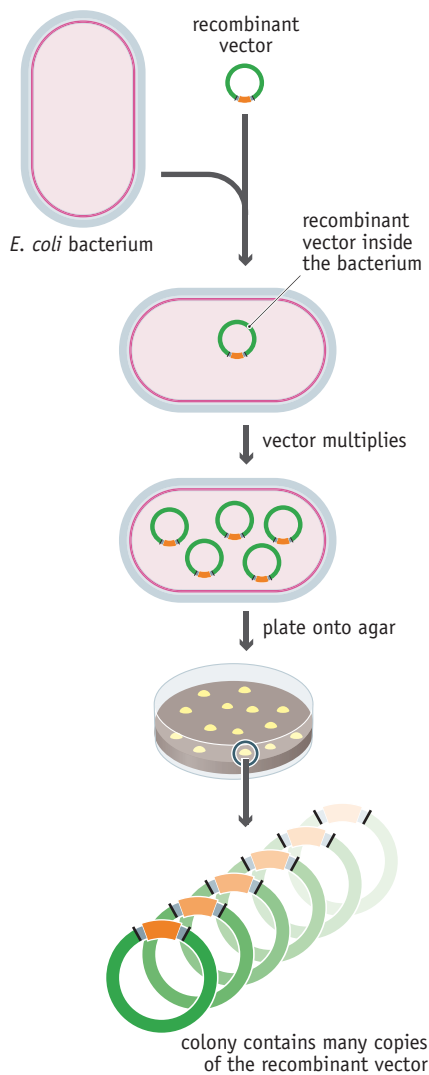

FIGURE 2.16 Assembling long DNA sequences from the sequences of short fragments. The original DNA molecule is broken into smaller fragments. These are sequenced individually, and overlaps between these sequences identified in order to assemble the complete sequence of the starting molecule.
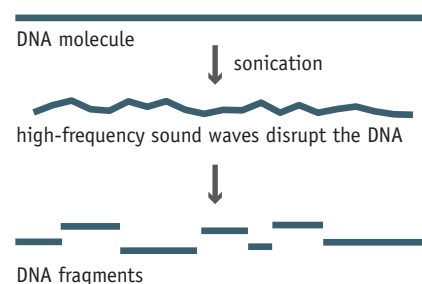
FIGURE 2.17 Using sonication to break a DNA molecule into fragments.

(A) construction of a recombinant DNA molecule

DNA fragment

vector

DNA ligase

recombinant vector

FIGURE 2.18 The steps in DNA cloning. (A) A recombinant DNA molecule is constructed by ligation of a DNA fragment into a cloning vector. (B) In the actual cloning stage, the recombinant vector is taken up by a bacterium. The recombinant vector is then replicated until there are several hundred copies per cell. On a solid agar medium, the bacterium will divide many times to produce a colony. The colony will contain many copies of the recombinant vector, these molecules being the direct descendents of the molecule taken up by the original bacterium.

(B) cloning in *Escherichia coli*

recombinant vector

E. coli bacterium

recombinant vector inside the bacterium

vector multiplies

plate onto agar

colony contains many copies of the recombinant vector

In the next stage of the cloning procedure, the molecules resulting from ligation are mixed with *E. coli* cells, which have been chemically treated so that they are **competent**, a term used to describe cells that are able to take up DNA molecules from their environment (**Figure 2.18B**). Some of the molecules will be taken up by the bacteria—but usually no more than one molecule by any individual cell. Linear molecules, and circular molecules that lack the vector sequences, cannot be propagated inside a bacterium. Circular vectors, on the other hand, including the recombinant ones, are able to replicate, so that multiple copies are produced, exactly how many depending on the identity of the cloning vector, but usually several hundred per cell.
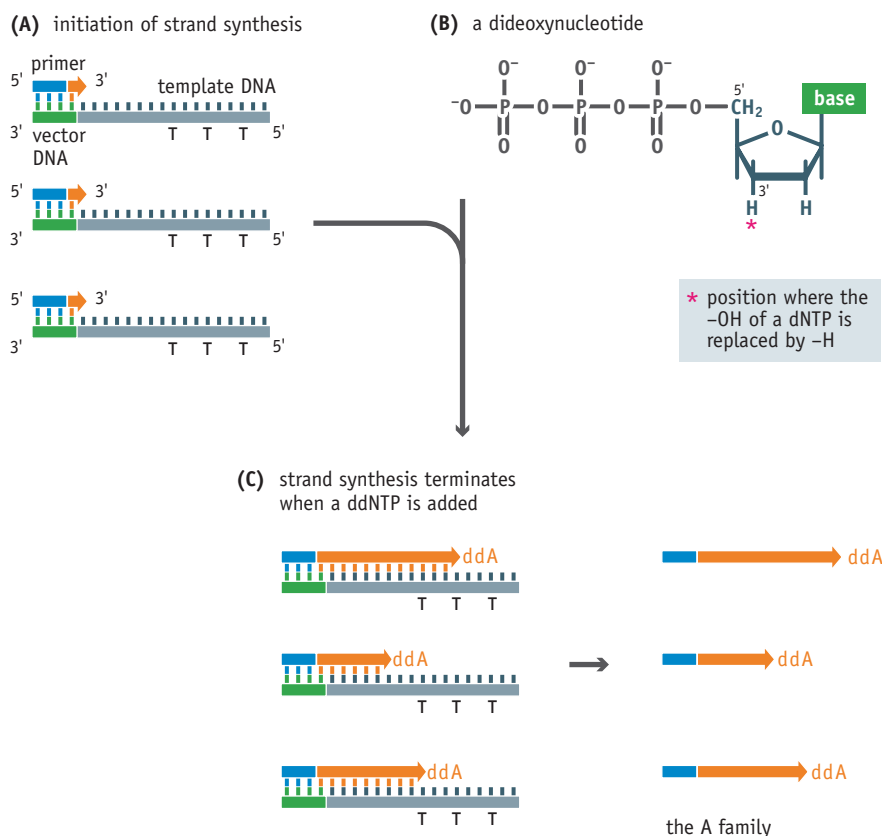
The bacteria are now spread onto a solid agar medium so that individual cells can divide and give rise to colonies. Each colony will contain many copies of the vector, these molecules being the direct descendents of the original molecule taken up by the bacterium that gave rise to the colony. If that original molecule was a recombinant vector, then the colony will contain multiple copies of a single DNA fragment, separate from all the other fragments that were present in the starting mixture. Other colonies will contain other cloned fragments. This collection of colonies is the clone library.

Each fragment in a clone library is individually sequenced

There are several methods for DNA sequencing, but most experiments are carried out using the **chain termination** method, which was first developed in the 1970s and has been used, with modifications and improvements, ever since. In the chain termination method, the sequence is obtained by making copies of a cloned DNA fragment in a way that enables the position of each of the four nucleotides to be identified. Chain termination sequencing therefore makes use of a purified DNA polymerase, the type of enzyme that makes copies of DNA molecules.

First the recombinant vector molecules are purified from one of the clones in the clone library. A short piece of DNA called an **oligonucleotide**, usually about 20 nucleotides in length, is then attached to the vector part of the recombinant DNA molecule, just adjacent to the site into which the cloned fragment has been inserted (**Figure 2.19A**). This oligonucleotide acts as a **primer** for synthesis of a new DNA strand. The new strand is made by the DNA polymerase and extends over the cloned fragment, so that a copy is made of this fragment.

DNA polymerases can make polynucleotides that are several thousand nucleotides in length, but in a sequencing experiment synthesis of the new strand is terminated prematurely. This occurs because as well as the four standard deoxynucleotides, a small amount of each of four **dideoxynucleotides** (ddNTPs—ddATP, ddCTP, ddGTP, and ddTTP) is added to the reaction. A dideoxynucleotide differs from a deoxynucleotide in that the –OH group

(A) initiation of strand synthesis

(B) a dideoxynucleotide

* position where the −OH of a dNTP is replaced by −H

(C) strand synthesis terminates when a ddNTP is added

the A family

FIGURE 2.19 The steps in chain termination DNA sequencing. (A) Chain termination sequencing involves the synthesis of new strands of DNA that are complementary to a single-stranded template. Strand synthesis begins at the primer, and proceeds in the direction indicated by the orange arrow. (B) Strand synthesis does not proceed indefinitely because the reaction mixture contains small amounts of each of the four dideoxynucleotides. These block further elongation because they have a hydrogen atom rather than a hydroxyl group attached to the 3′ carbon. (C) Incorporation of ddATP results in chains that are terminated opposite Ts in the template. This generates the A family of terminated molecules. Incorporation of the other dideoxynucleotides generates the C, G, and T families.

at the 3′ position is replaced by an H. The polymerase enzyme does not discriminate between deoxy- and dideoxynucleotides, but once incorporated a dideoxynucleotide blocks further elongation because it lacks the 3′-OH group needed to form a connection with the next nucleotide (**Figure 2.19B**).

Because the normal deoxynucleotides are also present, in larger amounts than the dideoxynucleotides, the strand synthesis does not always terminate close to the primer. In fact, several hundred nucleotides might be polymerized before a dideoxynucleotide is eventually incorporated. The result is a set of "chain-terminated" molecules, all of different lengths, and each ending in a dideoxynucleotide whose identity indicates the nucleotide—A, C, G, or T—that is present at the equivalent position in the DNA that has been copied (**Figure 2.19C**).

Capillary gel electrophoresis is used to read the sequence

To read the DNA sequence, all that we have to do is identify the dideoxynucleotide at the end of each chain-terminated molecule. This might sound difficult, but in fact it is easy. The technique involved is called **capillary gel electrophoresis**.

A capillary gel is a long, thin tube of **polyacrylamide**, typically 80 cm in length with a diameter of 0.1 mm, contained in a plastic sleeve (**Figure 2.20A**). The mixture of chain-terminated molecules is placed at one end of the capillary and an electric current is applied. The migration rate of each molecule through the capillary gel depends on its electric charge. Each phosphodiester bond in a DNA molecule carries a single negative electric charge (see Figure 2.5), so the overall negative charge of a polynucleotide is directly proportional to its length. The chain-terminated molecules therefore migrate through the capillary gel at different rates, the shortest ones

RESEARCH BRIEFING 2.2

Genes are made of DNA

Today we are so familiar with the fact that DNA is the genetic material that it comes as quite a surprise to learn that this idea was considered ridiculous by most biologists until the 1940s, and that experimental proof that human genes are made of DNA was not obtained until the 1970s. Why did it take so long to establish this fundamental fact of genetics?

At first it was thought that genes might be made of protein

The first speculations about the chemical nature of genes were prompted by the discovery in the very early years of the twentieth century that genes are contained in chromosomes. **Cytochemistry**, in which cells are examined under the microscope after staining with dyes that bind specifically to just one type of biochemical, showed that chromosomes are made of DNA and protein, in roughly equal amounts. One or the other must therefore be the genetic material.

In deciding whether it was protein or DNA, biologists considered the properties of genes and how these properties might be provided for by the two types of compound. The most fundamental requirement of the genetic material is that it be able to exist in an almost infinite variety of forms. Each cell contains a large number of genes, several thousand in the simplest bacteria, and tens of thousands in higher organisms. Each gene specifies a different biological characteristic, and each presumably has a different structure. The genetic material must therefore have a great deal of chemical variability.

This requirement appeared not to be satisfied by DNA, because in the early part of the twentieth century it was thought that all DNA molecules were the same. On the other hand, it was known, correctly, that proteins are highly variable polymeric molecules, each one made up of a different combination of 20 chemically distinct amino acids. There are many different proteins, distinct from one another by virtue of their different amino acid sequences. Proteins therefore possess the variability that would be required by the genetic material. Not surprisingly, biologists during the first half of the twentieth century concluded that genes were made of protein and looked on the DNA component of chromosomes as much less important—perhaps a structural material, needed to hold the protein "genes" together.

## Two experiments suggested that genes might be made of DNA

The errors regarding DNA structure lingered on, but by the late 1930s it had become accepted that DNA, like protein, has immense variability. The notion that protein was the genetic material initially remained strong, but was eventually overturned by the results of two experiments.

The first of these was carried out by Oswald Avery, Colin MacLeod, and Maclyn McCarty, of Columbia University, New York. They studied what researchers of the time were calling the **transforming principle**. They prepared an extract from dead cells of *Streptococcus pneumoniae*, a bacterium that causes pneumonia. Something in the extract was known to transform a harmless strain of *S. pneumoniae* into one capable of causing the disease. This transforming principle must therefore contain genes that provide the bacteria with the biological characteristics they need to cause pneumonia. Avery and his colleagues showed in 1944 that the active component of the extract, the transforming principle, is DNA (**Figure 1**).

The second experiment was performed in 1952 by Alfred Hershey and Martha Chase, at Cold Spring Harbor, New York. They showed that when a bacterium is infected with a virus, the DNA of the virus enters the cell, but most of the virus protein stays outside (**Figure 2**). This was a vital observation because, during the infection cycle, the genes of the infecting viruses are used to direct synthesis of new viruses, and this synthesis occurs within the bacterial cells. If it is the DNA of the infecting viruses that enters the
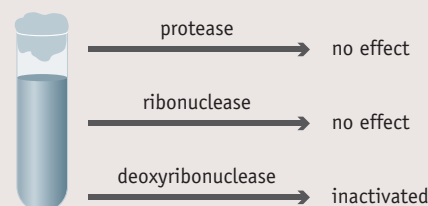
FIGURE 1 Avery and his colleagues treated extracts containing the transforming principle with a protease enzyme, which specifically degrades protein, with a ribonuclease, which breaks down RNA, and with a deoxyribonuclease, which degrades DNA. The protease and the ribonuclease had no effect on the ability of the extract to transform harmless *S. pneumoniae* bacteria. The deoxyribonuclease, on the other hand, inactivated the transforming principle. The active component of the transforming principle must therefore be DNA.
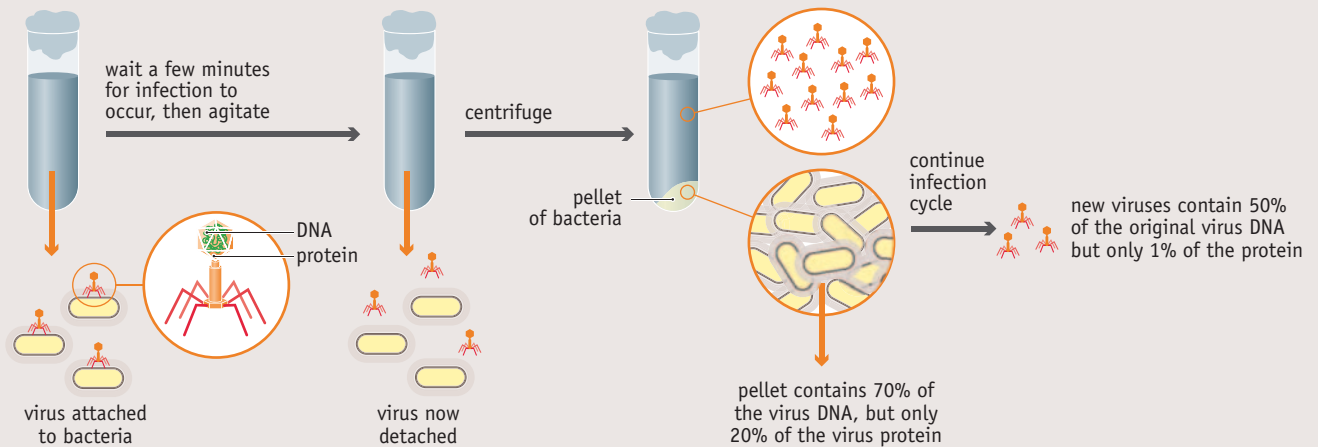
FIGURE 2 The Hershey–Chase experiment showed that the genes of T2 bacteriophage, a type of virus that infects *E. coli*, are made of DNA. Hershey and Chase knew that T2 bacteriophage was made of DNA and protein. We now know that the DNA contained a protein "head," which is attached to a "body" and "legs," also made of protein. Hershey and Chase added some T2 bacteriophages to a culture of *E. coli* bacteria, and then waited a few minutes to allow the viruses to inject their genes into the cells. They then agitated the culture in a blender in order to detach the empty virus particles from the surfaces of the bacteria. The culture was then centrifuged, which collects the bacteria plus virus genes as a pellet at the bottom of the tube, but leaves the empty virus particles in suspension. Hershey and Chase found that the bacterial pellet contained most of the viral DNA, but only 20% of the viral protein. Hershey and Chase also allowed some of the bacteria to complete their infection cycles. The new viruses that were produced inherited 50% of the DNA from the original T2 bacteriophages, but only 1% of the protein.

cells, then it follows that the virus genes must be made of DNA. Hershey and Chase also showed that the new viruses that are produced contain DNA, but only small amounts of protein, from the original infecting viruses. In other words, the new viruses inherit DNA but not protein from their parents.

The double helix convinced biologists that genes are made of DNA

Although from our perspective the Avery and the Hershey–Chase experiments provided the key results to tell us that genes are made of DNA, biologists at the time were not so easily convinced. Both experiments had limitations that enabled skeptics to argue that protein could still be the genetic material. For example, there were worries about the specificity of the **deoxyribonuclease** enzyme that Avery used to inactivate the transforming principle. This result, a central part of the evidence that the transforming principle is DNA, would be invalid if the enzyme contained trace amounts of a contaminating **protease** and so was also able to degrade protein.

The Hershey–Chase experiment is less open to criticism, but Hershey and Chase did not believe that it proved that genes are made of DNA. In the paper describing their results, they state, "The chemical identification of the genetic part [of the virus] must wait, however, until some questions … have been answered."

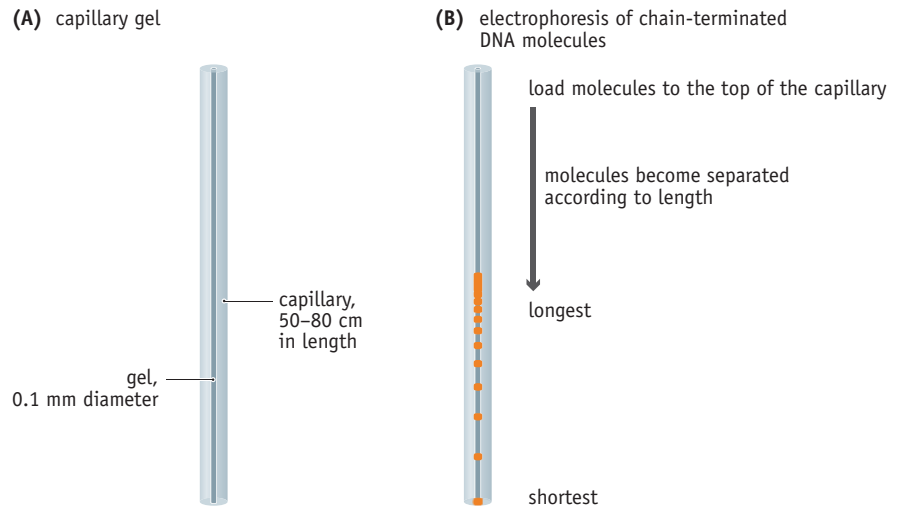Whether these two experiments actually proved that genes are made of DNA is not really important. They made biologists much more receptive to the notion that DNA *might* be the genetic material. This meant that when the double-helix structure of DNA was discovered by Watson and Crick in 1953, revealing how genes can replicate, the scientific world immediately accepted that genes really are made of DNA.

It was not proved that human genes are made of DNA until cloning was invented

If we wish to apply the most rigorous scientific principles, then we might argue that the Avery experiment shows that bacterial genes are made of DNA, and the Hershey–Chase experiment does the same for virus genes, but neither tells us anything about the genes of higher organisms such as humans.
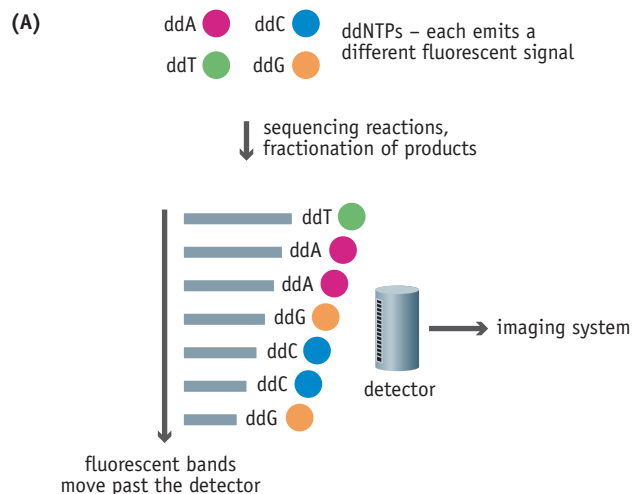
It was not until 1979 that an experiment was carried out that showed that human genes are made of DNA. David Goeddel, working for the biotechnology company Genentech, used DNA cloning to transfer the gene for somatotrophin, one of the human growth hormones, into *Escherichia coli*. The bacteria acquired the ability to make somatotrophin, showing that the DNA that had been transferred did indeed contain the human gene. The purpose of this experiment was to develop a means of using bacteria to produce the large amounts of somatotrophin needed for its clinical use in the treatment of growth defects, and it is rightly looked on as one of the major breakthroughs in the development of the biotechnology industry. The fact that it also finally laid to rest the question about the chemical nature of the genetic material, first asked some 75 years earlier, went unnoticed.

FIGURE 2.20 Capillary gel electrophoresis. (A) A capillary gel. (B) The principle behind separation of DNA molecules by capillary gel electrophoresis. During electrophoresis, the DNA molecules become separated according to their lengths, the shortest ones moving most quickly through the gel. The molecules therefore form a series of bands in the capillary tube.

(A) capillary gel

capillary, 50–80 cm in length

gel, 0.1 mm diameter

(B) electrophoresis of chain-terminated DNA molecules

load molecules to the top of the capillary

molecules become separated according to length

longest

shortest

most quickly and the longer ones the slowest. By the time they get to the end of the capillary they have formed bands, each band containing the molecules that are one nucleotide longer than those in the preceding band (**Figure 2.20B**).

Each of the four dideoxynucleotides emits a fluorescent signal of a different wavelength. A fluorescence detector is therefore used to identify which of the four dideoxynucleotides is present in each band. By identifying whether the dideoxynucleotide is an A, a C, a G, or a T, the detector is able to build up the DNA sequence of the cloned fragment nucleotide by nucleotide (**Figure 2.21**). The sequence can be printed out for examination by the operator, or entered directly into a storage device for future analysis.

(A)

ddA ● ddC ●
ddT ● ddG ●

ddNTPs – each emits a different fluorescent signal

sequencing reactions, fractionation of products

ddT ●
ddA ●
ddA ●
ddG ●
ddC ●
ddC ●
ddG ●

detector

imaging system

fluorescent bands move past the detector

FIGURE 2.21 Reading the sequence generated by a chain termination experiment. (A) Each dideoxynucleotide emits a different fluorescent signal. During electrophoresis, the chain-terminated molecules move past a fluorescence detector, which identifies which dideoxynucleotide is present in each band. The information is passed to the imaging system. (B) A DNA sequencing printout. The sequence is represented by a series of peaks, one for each nucleotide position. In this example, a red peak is an A, blue is a C, orange is a G, and green is a T.

(B)

C A C C G C A T C G A A A T T A A C T T C C A A A G T T A A G C T T G G

10   20   30
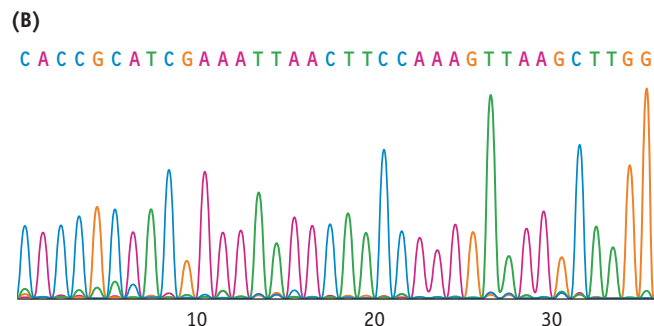
As mentioned above, a maximum of about 750 bp of sequence can be obtained in a single experiment. Automated sequencers with multiple capillaries working in parallel can read up to 96 different sequences in a 2-hour period, which means that 864,000 bp of information can be generated per machine per day. This, of course, requires round-the-clock technical support, with robotic devices used to prepare the sequencing reactions and to load the chain-terminated molecules into the sequencers. If such a factory approach can be established and maintained, then the data needed to sequence DNA molecules that are millions of base pairs in length can be generated in just a few weeks.

KEY CONCEPTS

- The structure of DNA reveals how genes are able to play their dual roles as units of biological information and as units of inheritance.

- DNA is a polymer in which the individual units are called nucleotides. There are four different nucleotides, usually referred to as A, C, G, and T, the abbreviations of their full chemical names. Nucleotides can be linked together in any order.

- The sequence of nucleotides in a DNA molecule is a language made up of the four letters A, C, G, and T. The biological information contained in a gene is written in this language.

- The double-helix structure reveals how DNA molecules are able to replicate. The two strands of the double helix are held together by hydrogen bonds, in such a way that A can base-pair only with T and G can pair only with C. This means that if the two polynucleotides in a DNA molecule are separated, then two perfect copies of the parent double helix can be made as the sequences of these preexisting strands dictate the sequences of new complementary strands.

- There are variations of the double-helix structure, called A-, B-, and Z-DNA. B-DNA is the most common type in the cell.

- Modern techniques of DNA sequencing are able to read the order of nucleotides in DNA molecules. Only a short sequence, of up to 750 bp, can be obtained in a single experiment, but automated systems enable many experiments to be carried out at once.

QUESTIONS AND PROBLEMS (Answers can be found at www.garlandscience.com/introgenetics)

Key Terms

Write short definitions of the following terms:

3'-OH terminus	chain termination	dideoxynucleotide
5'-P terminus	clone library	DNA cloning
β-*N*-glycosidic bond	competent	DNA ligase
adenine	complementary	DNA polymerase
A-form	contigs	DNA sequence
antiparallel	cytochemistry	DNA sequencing
base pairing	cytosine	gene expression
B-form	deoxyribonuclease	genetic code
capillary gel electrophoresis	deoxyribonucleic acid	genetic material

guanine
hydrogen bond
major groove
minor groove
model building
molecular approach
monomer
nucleoside
nucleotide
oligonucleotide

paper chromatography
pentose
phosphodiester bond
plasmid
polyacrylamide
polymer
polynucleotide
primer
protease
purine

pyrimidine
recombinant plasmid
sonication
template-dependent DNA synthesis
thymine
transforming principle
vector
X-ray diffraction pattern
Z-DNA

Self-study Questions

2.1   Draw the structure of a nucleotide.

2.2   What are the complete chemical names of the four nucleotides found in DNA molecules?

2.3   Draw a fully annotated diagram of the structure of a short DNA polynucleotide containing each of the four nucleotides.

2.4   Explain why the two ends of a polynucleotide are chemically distinct.

2.5   Outline the two major types of experimental analysis that laid the foundations for the deduction of the structure of DNA by Watson and Crick.

2.6   What are the important features of the double-helix structure?

2.7   What is meant by complementary base pairing, and why is it important?

2.8   If the sequence of one polynucleotide of a DNA double helix is 5'-ATAGCAATGCAA-3', what is the sequence of the complementary polynucleotide?

2.9   Thirty percent of the nucleotides in the DNA of the locust are As. What are the percentage values for: (a) T, (b) G + C, (c) G, (d) C?

2.10   DNA from the fungus *Neurospora crassa* has an AT content of 46%. What is the GC content?

2.11   What are the main differences between the A- and B-forms of DNA?

2.12   Describe how Z-DNA differs from the A- and B-forms, and outline one possible function of Z-DNA in living cells.

2.13   Explain how DNA provides the variability needed by the genetic material.

2.14   Draw a diagram to illustrate the process called template-dependent DNA synthesis.

2.15   Explain why in the laboratory a long DNA molecule is initially sequenced as a set of shorter contigs.

2.16   Outline how a clone library is prepared.

2.17   Describe how a chain termination DNA sequencing experiment is carried out.

2.18   Explain how a DNA sequence is read by capillary gel electrophoresis.

2.19   Why did biologists originally think that protein is the genetic material?

2.20   Outline the two experiments carried out in the 1940s and 1950s that indicated that genes are made of DNA.

2.21   To what extent is the statement "Genes are made of DNA" consistent with the results of the Avery and Hershey–Chase experiments?

Discussion Topics

2.22   Is the statement "Genes are made of DNA" universally correct?

2.23   An A–T base pair is held together by two hydrogen bonds and a G–C base pair by three hydrogen bonds. In which parts of the genome might you expect to find AT-rich sequences?

2.24   Discuss why the double helix gained immediate universal acceptance as the correct structure for DNA.

2.25   The human genome has a GC content of 40.3%. In other words, 40.3% of the nucleotides in the genome are either G or C. The GC contents for different organisms vary over a wide range. The genome of the malaria parasite, *Plasmodium falciparum*, has a GC content of just 19.0%, whereas that of the bacterium *Streptomyces griseolus* is 72.4%. Speculate on the reasons why the GC contents for different species should be so different.

2.26 Explore the reasons why, in the early twentieth century, some biologists thought that genes were abstract entities invented by geneticists to explain how biological characteristics are passed from parents to offspring.

2.27 In Research Briefing 2.2, we noted that a formal proof that human genes are made of DNA was not obtained until 1979 when a human DNA molecule was first transferred into *E. coli* and shown to direct synthesis of a human protein. What other experiments could be carried out to demonstrate that DNA is the genetic material in humans?

2.28 The scheme for DNA replication shown in Figure 2.14 is the same as that proposed by Watson and Crick immediately after their discovery of the double-helix structure. Many biologists thought that this process would be impossible in a living cell, especially for the circular DNA molecules present in many bacteria. Why was this?

2.29 A DNA polymerase builds up a new polynucleotide by adding nucleotides one by one to the 3′ end of the growing strand. Enzymes that make DNA in the opposite direction, by adding nucleotides to the 5′ end, are unknown. This fact complicates the process by which a double-stranded DNA molecule is replicated. Explain.

FURTHER READING

Brown TA (2010) Gene Cloning and DNA Analysis: An Introduction. Oxford: Wiley-Blackwell. *Includes details of DNA cloning and sequencing.*

Hershey AD & Chase M (1952) Independent functions of viral protein and nucleic acid in growth of bacteriophage. *J. Gen. Physiol.* 36, 39–56. *One of the original papers that showed that genes are made of DNA.*

Maddox B (2002) Rosalind Franklin: The Dark Lady of DNA. London: HarperCollins. *A biography of one of the key people involved in discovery of the double-helix structure, who sadly died just a few years later.*

McCarty M (1985) The Transforming Principle: Discovering That Genes Are Made of DNA. London: Norton. *Personal account by one of the scientists who worked with Avery.*

Olby R (1974) The Path to the Double Helix. London: Macmillan. *A scholarly account of the research that led to the discovery of the double helix.*

Prober JM, Trainor GL, Dam RJ et al. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* 238, 336–341. *The chain termination method for DNA sequencing, as it is used today.*

Rich A & Zhang S (2003) Z-DNA: The long road to biological function. *Nat. Rev. Genet.* 4, 566–572.

Sanger F, Nicklen S & Coulson AR (1977) DNA sequencing with chain terminating inhibitors. *Proc. Natl. Acad. Sci. USA* 74, 5463–5467. *The first description of chain termination sequencing.*

Watson JD (1968) The Double Helix. London: Atheneum. *The most important discovery of twentieth-century biology, written as a soap opera.*

Watson JD & Crick FHC (1953) Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature* 171, 737–738. *The scientific report of the discovery of the double-helix structure.*