versus step-number in the sequence: roll angles of 45° can be seen at steps 0 and 10 in this diagram, but at other steps there is zero roll.

It might be better, however, to spread these roll angles over a greater number of base-pair steps within the circle. While a single roll angle of 45° per helix turn is satisfactory from a geometrical point of view, it is not satisfactory from a physical point of view, because the opening of a single base-pair step by 45°, as in Fig. 4.8(a), would expose a lot of water-insoluble base surfaces to the solvent. A slight improvement might be made by dividing the 45° into two parts, and thus locating a roll angle of 22.5° at two places within each double-helical turn. This scheme is shown in Fig. 4.8(b), and it has also been plotted in Fig. 4.9(b): roll angles of +22.5° are located at steps 0 and 10, while roll angles of −22.5° are located at steps −5, 5 and 15. The sign of the roll angle changes from plus to minus every five base-pairs, because the helix rotates by 180° over that distance. So the two angles of 22.5° open up the same side of the helix, and make equal contributions to the curvature, despite the change in sign.

Another way of bending DNA into the required curve might be to have five steps with a roll of about 14° followed by five with zero roll, and then five more with $R = 14°$, etc. This new scheme is shown in Figs 4.8(c) and 4.9(c). In effect, we are converting successive half-turns of DNA into different uniform configurations; and it is not difficult to see that this will produce a curve. The model shown in Fig. 4.8(c) is sometimes called a 'junction' model for DNA curvature, because 'bends' seem to appear at the junctions between the successive portions of DNA, each having uniform but different roll: but it is important not to forget that curvature depends on the roll angles at all of the steps.

Perhaps the best scheme of all for making curved DNA would be to spread the roll over the entire helix, at every base-pair step, rather than placing big roll angles at only a few steps. Such an idealized model is shown in Fig. 4.8(d), and in the corresponding diagram of Fig. 4.9(d). If we let the roll angles vary as a cosine wave, starting at step 0 and ending at step 10, then we shall need to vary these roll angles by no more than plus or minus 9° to get a curvature of 45° per turn. You can see that there are much less abrupt changes in roll from step to step in Fig. 4.8(d) than in any of (a), (b), or (c). This may seem a bit like magic: how can we know what curvature will result from such an assortment of different roll angles as in this example?

The roll angles in Fig. 4.9(d) vary as a cosine wave of amplitude 9°, or as $R_n = 9° \cos(36°n)$, where $n = 0, 1, 2, \ldots, 9$ identifies each step in any double-helical turn. For example, the roll at step 0 is $R_0 = 9°$, while the roll at step 1 is $R_1 = 9° \cos(36°) = 7.3°$, and the roll at step 5

is $R_5 = 9° \cos(180°) = -9°$. This new arrangement is similar overall to the various schemes shown before in (a), (b), and (c), but it is more subtle in the way in which it assigns roll to different parts of the helix. Now we come to a crucial point: not all of the roll angles contribute fully to curvature, because some of them cause the DNA to bend in the wrong way, out of the plane of the paper in Fig. 4.8. Let us look again at the helix shown in Fig. 4.8(a). In this picture, a positive roll angle at step 2, between steps 0 and 5, would bend the DNA down into the paper, rather than to the right as desired. But a roll angle at step 0 bends the DNA to the right, while a roll at step 5, not shown in the drawing, would also bend the DNA to the left or to the right, as illustrated in Fig. 4.8(b). It can be shown that the contribution to rightward curvature by any step is its roll angle R_n, multiplied by the cosine of the total helix twist, or $R_n \cos(36°n)$. Thus, with arrangement (d) we get for step 0 a big contribution to curvature of $9° \cos(0°) = 9°$, and for step 5 we also get a big contribution of $-9° \cos(180°) = 9°$; but for step 2 we get only the small contribution of $9° \cos^2(72°) = 0.9°$, because the roll points mostly in the wrong direction.

It follows, then, that the overall curvature k due to roll at all steps $n = 0, 1, 2, \ldots, 9$ can be found by adding together 10 different terms of the kind $R_n \cos(36°n)$, or in this case

$$\sum_{n=0,9} 9° \cos^2(36°n) = 9° + 5.9° + 0.9° + 0.9° + 5.9° + 9° + 5.9° + 0.9° + 0.9° + 5.9° = 45°$$

In summary, the roll angles in Fig. 4.9(d) vary as $9° \cos\theta$, where θ is the total twist (or sum of base-step twists T) relative to step 0; and then you have to multiply them by $\cos\theta$ again before adding up to get the total curvature, because not all steps point in the same direction. This simple scheme is not quite exact, geometrically, but it works well for small roll angles, of less than or equal to about 10°; and that covers most practical cases.

It is not difficult to show that if a constant roll angle is added to every step, so that $R_n = 9° \cos(36°n) + (\text{constant})$, then the sum worked out above has exactly the same value as before. A constant change of roll at all steps will change the overall appearance of the DNA, just as in Fig. 3.14(c), but it imparts no curvature.

As we have seen, not every step contributes equally to curvature in our scheme. Steps 0, 1, 4, 5, 6, 9 (where the minor or major groove faces the center of curvature) contribute a lot, while steps 2, 3, 7, 8 don't contribute much at all. The reason for this behavior is that the

DNA is not flexible enough in a direction at right-angles to the roll axis (that is, about the 'front–back' axis of Fig. 3.7) to curve very much in that direction. Also, one might suspect that slide S and twist T vary along with roll R as the DNA curves, in accord with the relations among R, S, and T discussed above in Chapter 3; but this would make only small differences to the general pattern of behavior which we have described.

When studying the curvature of real DNA, which is made from a variety of base sequences, we cannot expect every kind of sequence to follow a smooth, wave-like pattern of roll angles exactly. Some steps, such as the pyrimidine–purine variety and others discussed in Chapter 3, are flexible enough to adopt either high roll or low roll with little difficulty. However, others such as AA/TT are known from studies of DNA in the crystal to prefer low roll, $R = 0°$; while steps such as GC/GC are thought to prefer high roll, $R = +5°$ to $+10°$. Essentially, what we must do for real DNA, when it is curved around a protein spool, is to find the best fit of a cosine wave to a given series of R values, at a period of 10 steps, even though some of these steps may not agree with the idealized cosine wave exactly. The spread of roll angles can range from $+9°$ to $-9°$, or from $+12°$ to $-6°$; it doesn't matter. Once we have obtained a 'best fit', then the amplitude of this best-fit cosine wave will be proportional to the absolute curvature of the sequence, while its left-to-right position or phase, as in Fig. 4.9(d), will tell which steps have the largest (or most positive) roll angle, and which steps have the smallest (or most negative). Where the roll angle is large and positive, the minor-groove edges of base-pairs will lie along the *outside* of the curved DNA, and where the roll angle is small (or most negative), the minor-groove edges will lie along the *inside* of the curve. This protocol is known to mathematicians as taking the 'Fourier transform' of the roll angles, to get the amplitude and phase of the curvature of the DNA helix.

You can see for yourself how this works by making a detailed plot along the lines of Fig. 4.9(c) for a series of steps such as those shown in Fig. 4.8(c). There, a batch of 5 steps (-2, -1, 0, 1, 2) with $R = 14°$ is followed by a batch of 5 steps with $R = 0°$, and then by another batch with $R = 14°$, and so on. You will find that the cosine wave that can be drawn most closely over these points has an amplitude (or half-height) of about 9°; which is just what is needed for a curvature of 45° per double-helical turn.

When we consider the DNA that is found in our chromosomes, we might expect to find AA/TT sequences where the DNA curvature requires low roll, and GC sequences where the curvature requires high roll. That is precisely what we do find, to a first approximation. Most of the DNA in our chromosomes curves strongly around

protein 'spools' (Fig. 1.5) for almost two turns of 80 base-pairs each, or for about 160 base-pairs in all, into a flat, left-handed supercoil that resembles part of a telephone cord. The proteins that make up any spool are called 'histones', and they will be discussed in Chapter 7. When we examine the DNA sequences that reside in these tight coils that wrap around the protein spools, we find that AA/TT sequences have a higher probability of being in a low-roll position than other DNA, and that GC sequences have a higher probability of being in a high-roll position; but not every AA/TT goes to low roll and not every GC goes to high roll. The preferences of all 160 base-pair steps in this DNA have to balance against one another to attain an optimal positioning of roll angles for *most* of the steps in the sequence; and so some individual steps may not fit into the overall pattern.

Some actual data are shown in Fig. 4.10, concerning the preferred locations of different base sequences in curved, chromosomal DNA. To construct these plots, we isolated the DNA from over one hundred different spool–DNA complexes, and then counted how many times N each kind of sequence might be found at any possible location on the histone spool. During the isolation of this DNA, it was trimmed in length from 160 base-pairs to 145 base-pairs by a special enzyme, in order to mark more precisely where each spool might begin or end. The plots in Fig. 4.10 show data that have been averaged over positions 1 to 72 and 73 to 144 of the spool–DNA complex, about the center at 72.5, because these data were found to be much the same on both sides of the center.

Figure 4.10(a) shows the number N of AA/TT steps that were found at each possible location along the path of the long, curved DNA in our many examples. These AA/TT steps may be seen to be located preferentially at positions 5, 15, 25, 36, 46, and 56 from either end of the DNA; such locations are known to be positions of low roll, where the minor groove faces inwards, from an X-ray analysis of the spool–DNA complex. Around the center of the DNA at position 72.5, the periodic pattern is broken, and peaks for AA/TT appear at positions 62 and 72. It turns out that the preference of AA/TT steps for these central positions may not be due to DNA curvature, but perhaps to specific contacts between certain amino-acid side-chains on the protein and the exposed edges of the A/T base-pairs; the explanation is still not certain. But we need not worry about this detail: the bulk of the DNA shows roll-angle preferences in accordance with our general scheme.

Figure 4.10(b) shows, similarly, the number N of GC steps at each possible location along the long, curved DNA. These GC steps may be seen to be located preferentially at positions 11, 21, 31, 41, 52, or about 5 steps away from the preferred locations of AA/TT steps. In

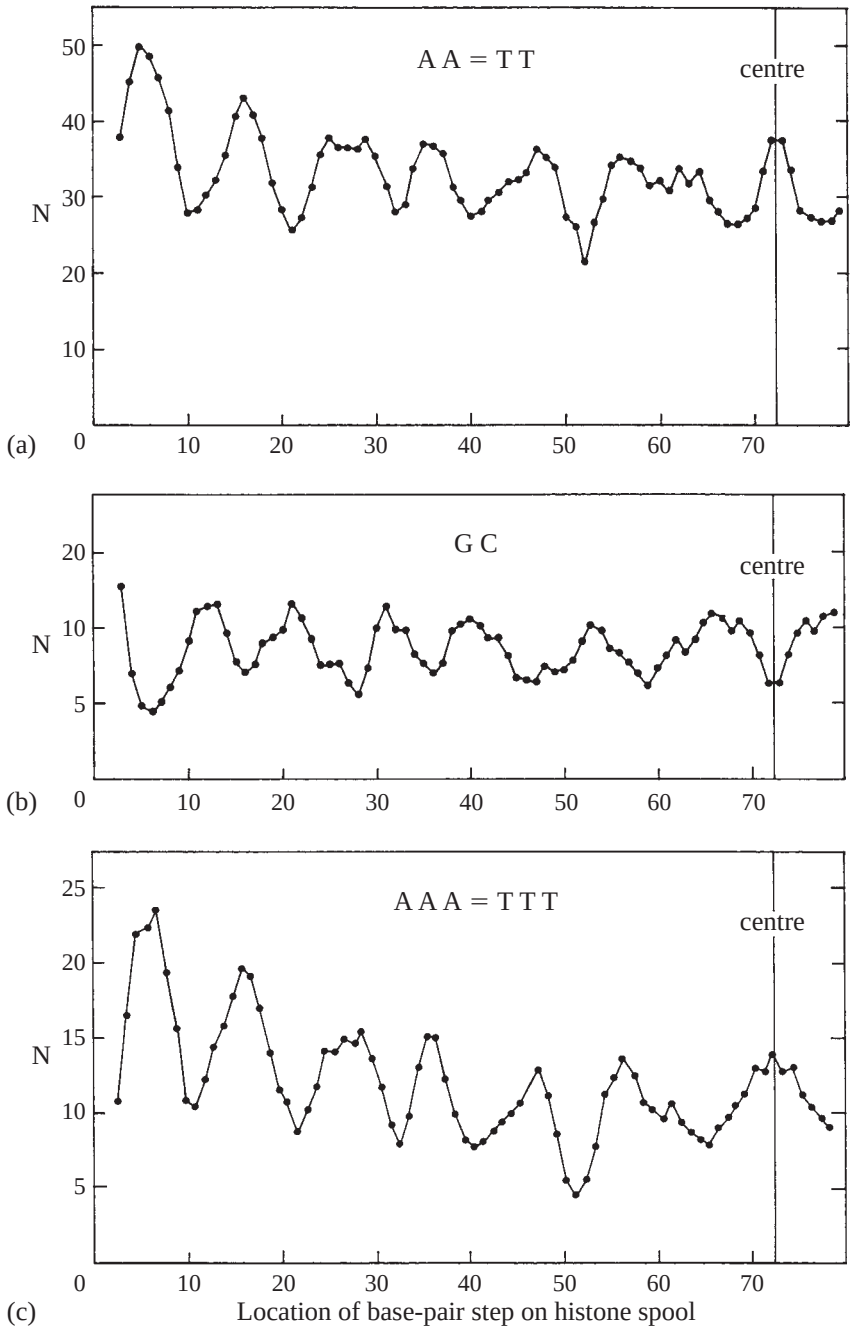

Figure 4.10 Number *N* of various short sequences in curved chromosomal DNA, plotted against their locations on the histone spool (see Fig. 1.5). The overall length of the DNA is 145 base-pairs, but the data could be averaged on both sides of the center at position 72.5, because they were found to be the same on both sides. Almost two hundred different DNA molecules were sequenced in order to construct these plots. From S.C. Satchwell *et al.* (1986) *Journal of Molecular Biology* **191**, 659–75.

other words, they prefer positions of high roll. The mean periodicity, or spacing of peaks, in both the AA/TT and GC patterns is close to 10.2 base-pairs, or one turn of double helix. The preferred helical periodicity of this DNA in other circumstances, when freed from the protein, is a slightly larger 10.6 base-pairs per turn.

Figure 4.10(c) shows an especially strong pattern of preferences for the trimer AAA/TTT: here two adjoining AA/TT sequences reinforce each other's individual preferences. Thus the AAA/TTT trimer has a strong preference to be located in a position of low roll rather than a position of high roll: the value of N differs by a factor of about 2 between these locations. This is quite a large factor, when you remember that the piece of DNA which is being curved around the histone spool consists of 144 steps or 145 base-pairs. Suppose this piece of DNA contains just one AAA/TTT trimer: then in 2 out of 3 cases, the AAA/TTT will be located in a position of low roll, due to its own influence in combination with the influences of other sequences. From this example, we can see that roll-angle preferences of particular short (i.e. two- and three-base-pair) sequences in DNA may be rather strong, provided they are not balanced by the opposing preferences of other sequences. Thus, certain AA/TT steps might predispose the DNA to adopt one position on the spool, whereas certain GC steps might prefer a different position. This is like the story of the husband and wife who always vote for different political parties, Republican and Democrat, or Conservative and Labour: in the end their votes cancel.

When studying pieces of DNA from a chromosome that are longer than the 145 base-pairs examined above, often we see that the spool–DNA complexes align themselves into an ordered array that may extend for 1000 to 2000 base-pairs. In such cases, the roll-angle preferences within any small region of the DNA, say 100 base-pairs, would have to be very strong to dominate the alignment of the other 900 to 1900 steps. Alternatively, these 100 steps could 'break' the array to yield a less regular structure; and many irregular models of chromosome structure have lately been proposed. In any case, the fit of roll-angle preferences in the DNA on one histone 'spool' to those of the DNA wrapped around neighboring spools is a matter of some interest, but it is not yet understood.

These alignments of DNA along the surfaces of histone spools turn out to be very important in biology, to decide for example where the AIDS or HIV virus will insert itself into human chromosomes. Such preferences for a particular rotational setting of the DNA are so strong, in fact, that some people are now using the HIV insertion enzyme as a way to probe the structures of chromosomes inside living cells! Thus in general, proteins such as the HIV insertion enzyme, that bind even to a very small region of DNA, seem to follow the same rules of DNA curvature.

As another example, a protein from a bacterial virus, known as the '434 repressor' (because it stops RNA from being made at a site of DNA unwinding in 434 virus), binds to DNA in the following way: it probes deeply into the grooves at either end of a 14-base-pair region, and curves the helix moderately in the middle, without making contact with it there, as shown in Fig. 4.11. The minor groove faces the protein at the center, and so the base-pair steps around bases 6, 7, 8, 9 must adopt a low or negative roll for the DNA to be bent by the protein in a correct fashion, so that the two parts of the protein can fit together properly. Indeed, if you put a low-roll sequence of four bases such as AATT in positions 6, 7, 8, 9 of the DNA, the protein binds 300 times more tightly than if you put a high-roll sequence such as GGCC. In chromosomes, such preferences are averaged out over a long region of DNA so there they appear to be weak; but in small regions of DNA, these preferences are not averaged out, and so they act strongly. In fact, you can calculate from the kind of bases which are found in positions 6 to 9 of the binding site how tightly this 434 repressor protein will bind to DNA, by using roll-angle preferences taken from the spool–DNA data discussed above. The same physical phenomenon of curvature-preference in DNA is operative in both cases.

But how does the 434 protein bind to DNA in positions 1 to 5 and 10 to 14, where it contacts the base-pairs most closely? Is the flexibility of DNA over positions 6 to 9 sufficient to provide enough specificity for the protein to carry out its function in the cell, of

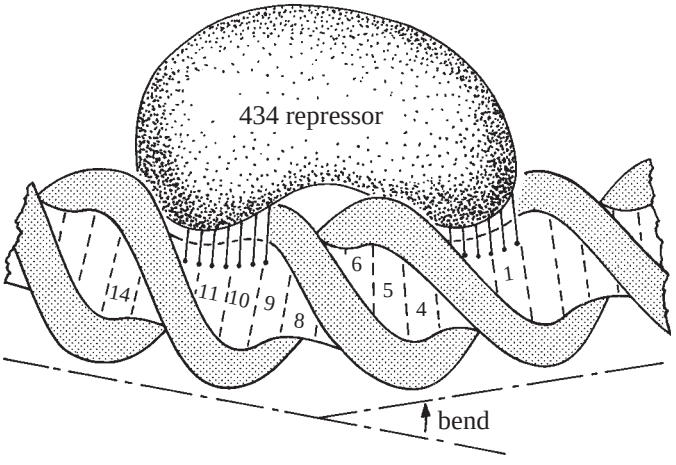

Figure 4.11 Schematic view of the binding of 434 repressor to DNA. This protein probes for the identities of base-pairs in positions 1 to 4 and 11 to 14 of its binding site, and also docks to the overall DNA conformation by testing the ease of curvature in positions 5 to 10. (More detail is shown in Fig. 8.2(a).)

repressing the synthesis of RNA from just one or a few genes on the DNA? For that purpose, the protein needs a specificity of binding to its particular target sequence over other sequences by a factor of 10 000, while the flexibility of DNA – as described above – provides a factor of 300 at most. The biological mechanism by which this protein blocks RNA synthesis is trivial: it simply binds to the same piece of DNA as that preferred by RNA polymerase, and so it physically prevents the polymerase from starting RNA chains. But the way in which the 434 repressor picks out a single DNA sequence (or just a few sequences) from many others in the bacterial chromosome is not easy to understand.

During the last few years, single crystals of many protein–DNA complexes have been analysed by X-ray methods, and this has made it possible to study in fine detail the close bonding of different parts of protein molecules to the bases of DNA. The general picture which has emerged from these studies is that the amino acids of a protein can make specific hydrogen bonds with exposed atoms on the sides of base-pairs, or along the 'floor' of the major or minor groove in the DNA. Such 'direct reading' of a DNA sequence by a small portion of protein will be a main subject in Chapter 8. But for the present, we may examine just one interaction of this sort as shown in Fig. 4.12, where a guanine base is making hydrogen bonds to an arginine amino acid. There are two contacts of hydrogen atoms on the arginine with oxygen or nitrogen atoms on the major-groove edge of the guanine ring. If the spatial patterns of electric charge on the two surfaces fit each other well, so that two or more hydrogen bonds can form, then there will be a highly specific local

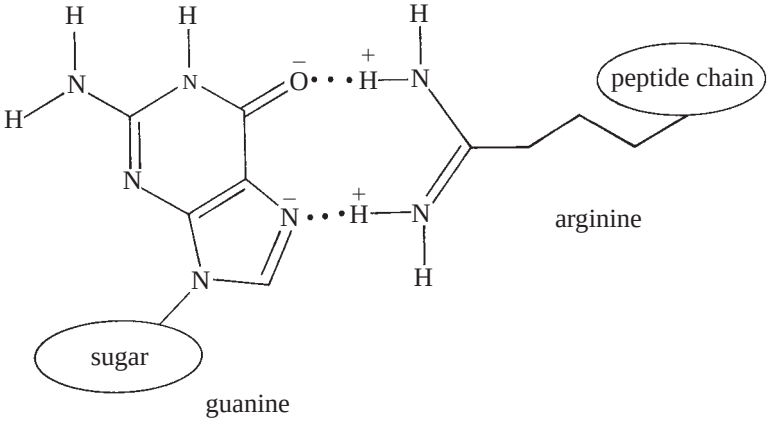

Figure 4.12 Hydrogen bonding between a guanine base in DNA and the arginine amino acid from a protein, as seen in many protein–DNA complexes (such as those shown in Figs 4.11 and 4.13). The paired cytosine base is not shown.

bonding arrangement between the amino acid and the base-pair. However, it is not completely true to say that the detailed patterns of amino acid-to-base hydrogen bonding, when summed over many different amino acids and bases, actually determine where a protein binds on DNA. For we have seen already that the flexibility of DNA plays a large part in its binding to the 434 protein, and to the histone–spool proteins, and to other proteins not discussed here; and this flexibility depends on the DNA sequence as well.

The picture we need in order to explain all aspects of protein–DNA recognition is one which involves consecutively two processes which we shall call 'docking' and 'probing'. By docking we mean the fitting together, on a large scale, of the protein and the DNA. If the fit at that scale is good, then the quality of hydrogen bonding between the contacting zones can be tested on a small scale; and only if this detailed probing is successful will the overall binding between the protein and the DNA be highly specific for some particular base sequence. In our example of the 434 repressor protein, the docking phase demands a capability of the DNA to bend in a certain way; if the DNA is too rigid, or prefers to bend in the wrong way, the docking cannot easily occur. But if the DNA has the right degree of flexibility, then the base-pairs in the contacting regions 1 to 5 and 10 to 14 can probe for the formation of enough hydrogen bonds with amino acids to let the second stage of binding take place.

Another example of the same two-stage process of recognition between DNA and protein is the cutting of DNA by the enzyme DNAase I. This enzyme is very useful for the study of DNA structure, as we shall see in Chapter 9. It can only dock with DNA if the width and depth of the minor groove lie within a certain range; and then only after it has docked can it begin to probe the configuration of the sugar–phosphate chain, and make a cut if it finds a specific geometry there. Thus the enzyme must recognize not only a *global* feature (the groove dimensions) but also a *local* feature (the phosphate configuration) if it is to succeed in cutting the DNA. These two features together provide typically a factor of 1000 in the specificity of cutting some bonds over others.

A third interesting example of protein–DNA recognition was provided by the X-ray analysis of a complex between a protein called Zif268 and the sequence to which it binds on DNA. Zif268 is part of a larger protein that is known as an activator of transcription, or a 'transcription factor', because it somehow stimulates the synthesis of RNA from certain genes close to where it binds on the DNA. No one today is certain how this might be accomplished.

Nevertheless, the structure of the complex between Zif268 and DNA is of interest in its own right. The protein contains a series of

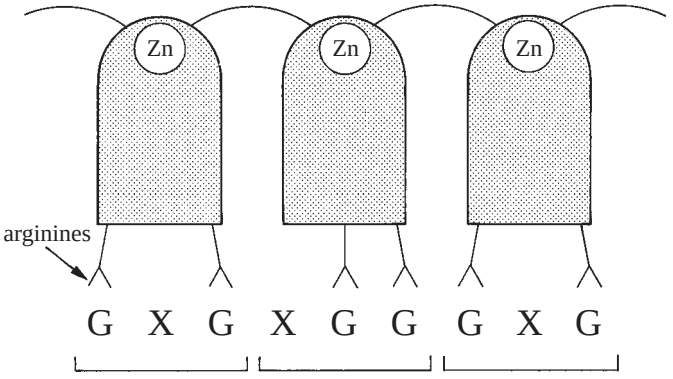

Figure 4.13 Schematic view of three 'zinc-fingers' recognising a particular DNA sequence, as in the X-ray structure of the Zif268 protein with DNA. Not shown are many contacts of the protein with DNA phosphates, or the moderately negative slide of the base-pairs. (More detail is shown in Fig. 8.6.)

small, modular units known as 'fingers', as shown in Fig. 4.13. Each unit contains a zinc atom which helps to fold the protein chain into a separate domain, and these domains are linked by short segments of the protein chain. Some proteins have as many as 10 to 20 fingers, but Zif268 has just three. When the Zif268 protein binds to DNA, it tries to attach itself to the DNA base-pairs through a series of arginine-to-guanine hydrogen bonds, indicated schematically in Fig. 4.13. Not shown in this diagram are many other hydrogen bonds which connect phosphate groups from one of the DNA sugar–phosphate chains to various other amino acids on the protein. In fact the three linked fingers form a spiral, which drapes itself along the sugar–phosphate chain on one side of the major groove, and thereby allows arginine amino acids to make firm contact with guanine bases in 6 out of 9 successive base-pairs. In other words, many contacts of the protein with the phosphates anchor it in place, so that it can contact the guanines.

How does such a zinc-finger protein 'dock' to the overall structure of DNA? In this example, the base-pairs of the DNA have moderately negative slide, and it may be that the negative slide imparts a particular spatial relationship to the phosphates and bases, which then constitutes a three-dimensional docking feature for any individual finger. Alternatively, the negative slide may act by increasing the distance of base-pairs from an overall helix axis, as shown in Figs 3.14(b) and 3.15, in which case the three fingers of the protein would have to follow a helical path of greater diameter than before. We shall return to the process of recognition of DNA sequences by zinc finger proteins in Chapter 8.