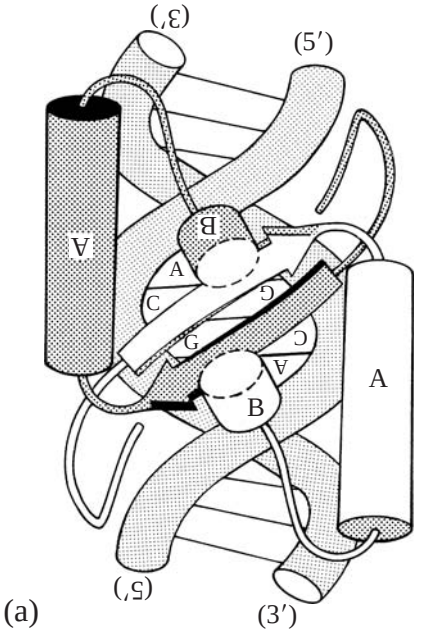
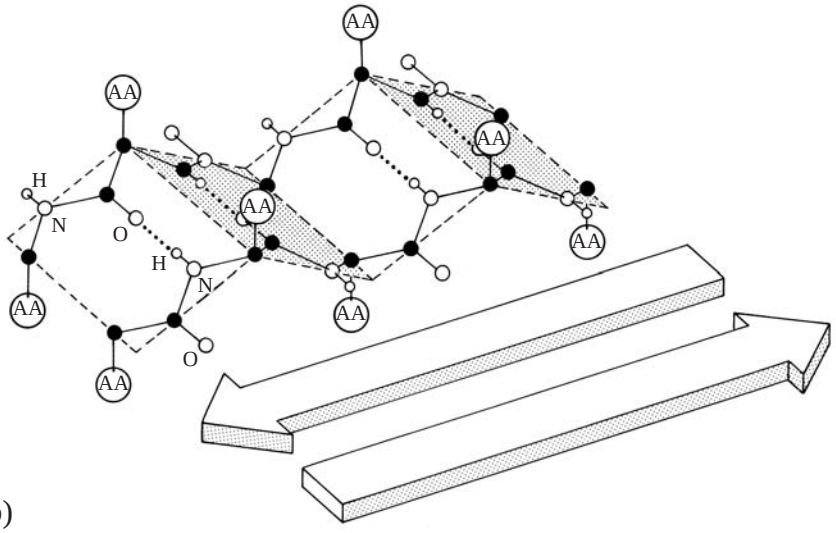
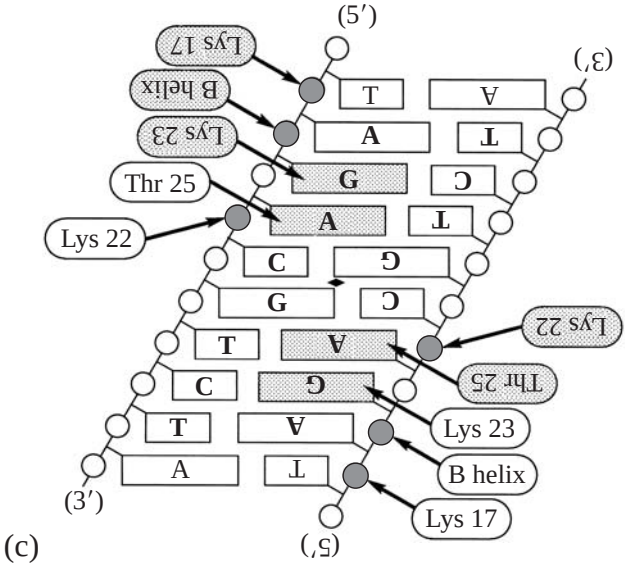


Figure 8.5 (a) A schematic drawing of the met repressor protein as bound specifically to DNA. The protein consists of two identical halves, each of which contains several α-helices (A and B) and part of a two-stranded β-sheet (broad arrows), which lies within the major groove. Other parts of the protein are not shown, for the sake of clarity. The identical halves are colored differently, and the upside-down lettering on some components corresponds to the two-fold rotational symmetry of the assembly. (b) A more detailed view of how a peptide chain folds into a β-sheet. Two peptide chains lie side-by-side (either parallel, or antiparallel as shown here), so that each peptide unit may make two hydrogen bonds (dotted lines) to nearby peptide units on other chains. Each peptide chain in a β-sheet can be drawn also as a broad arrow, as shown in the lower part of the diagram. (c) An unrolled view of the DNA sequence which is bound specifically to met repressor protein, showing which amino acids from the β-sheet contact which base-pairs or phosphates. The twofold symmetry of these contacts is also indicated by the upside-down lettering at the top, which is identical to the rightside-up lettering at the bottom.

extend the same β-sheet to the right, but which are represented in less detail by arrow-like strips, of the kind shown in Fig. 8.5(a).

When the met repressor protein binds to DNA, both strands of its β-sheet fit snugly into the major groove, so as to recognize base-pairs there. Close contacts between amino acids and base-pairs are shown in Fig. 8.5(c). The met repressor protein recognizes a conserved eight-base-pair sequence of DNA, although specific bonds are made by amino acids to only four of the eight individual bases in each part. Thus, adjacent G and A bases form hydrogen bonds with Lys (lysine) 23 and Thr (threonine) 25 from different strands of the β-sheet; while the outer two bases A and C in AGAC do not seem to contact any amino acids directly. Several DNA phosphates are





bound by various amino acids, namely Lys 17, 22, and a few peptides from α-helix B.

Thus, the specific recognition of DNA by the met repressor protein resembles closely the recognition of DNA by both 434 and bZIP proteins; except that in the case of met repressor, those amino acids which contact the base-pairs directly are held in place by a β-sheet rather than by an α-helix. The met repressor protein has two-fold rotational symmetry, and it recognizes a two-fold symmetric base sequence in the DNA. The same is true for the 434 repressor and bZIP proteins, and the DNA sequences which they recognize. But in addition to the interaction with DNA, two met repressor proteins

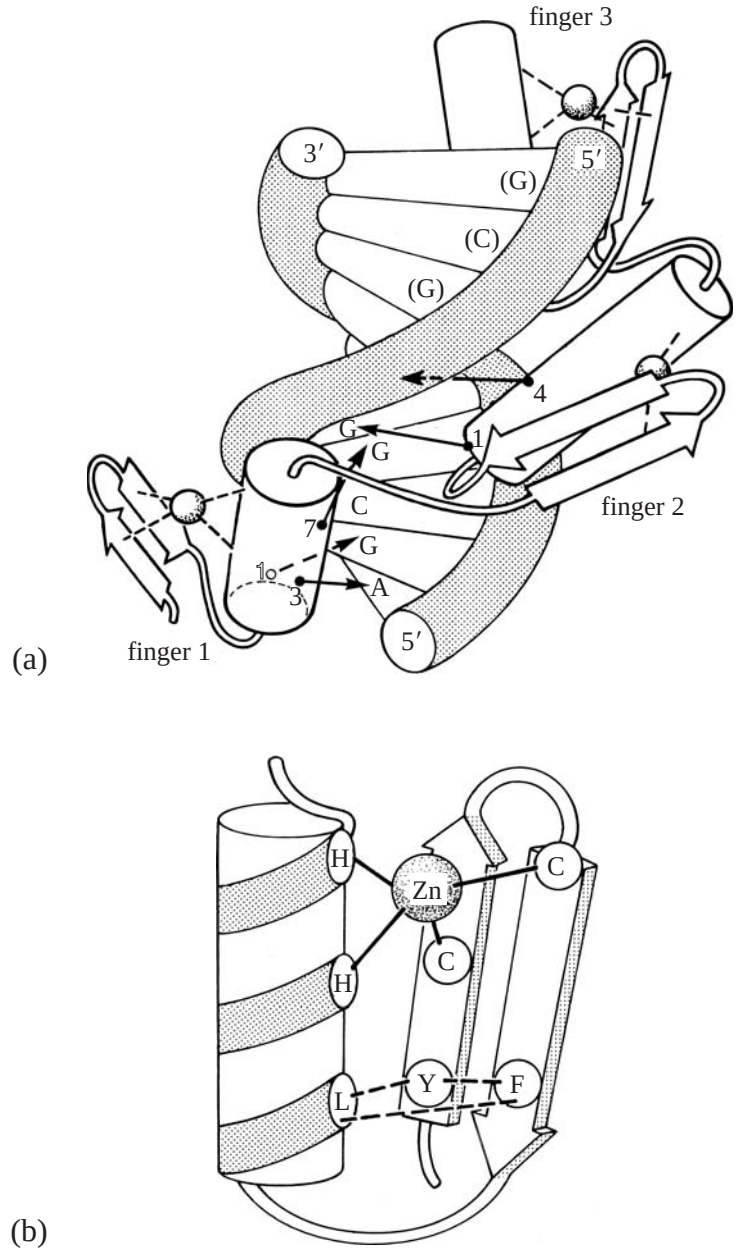


Figure 8.6 (a) A schematic drawing of the Zif268 protein as bound specifically to DNA, showing how its three zinc-finger modules coil in a right-handed sense, so as to follow the DNA major groove over almost a full turn. Each zinc-finger places an α-helix in the major groove at a steep angle of attack, so as to contact base-pairs there. Some bases are identified by letters, which are enclosed in parentheses on the minor-groove side. (b) A more detailed view of how a peptide chain folds into a zinc-finger. Part of the peptide chain wraps into an α-helix, while the other part wraps into a β-sheet. The zinc atom provides a link between them.

can also bind to each other if they to bind to adjacent sites on the DNA. This enhances the strength of DNA binding; and later we shall see several other cases where analogous protein-to-protein interactions occur on a repeated DNA target-sequence.

We mentioned in Chapter 4 the protein TBP (or 'TATA-binding-protein'), which unwinds and sharply bends the TATA promoter sequence in DNA: see Fig. 4.14. That special protein also uses a β-sheet to recognize DNA, as for the met repressor, except there the sheet fits into a greatly widened minor groove at the low-twist TATA sequence, rather than into a normal-width major groove as for met repressor. We shall return to TBP when we discuss how it interacts with another protein which is also bound to the DNA nearby.

The 'zinc-finger' family of proteins is another example of alternative structures used by proteins to recognize DNA. One example of such a structure was shown in Fig. 4.13, for the zinc-finger protein Zif268. There a series of zinc-fingers were pictured schematically as binding to a series of three-base-pair recognition sites on the DNA, by means of specific contacts between arginine amino acids and guanine bases in the major groove. The 'modular' arrangement permits small proteins to select long recognition sequences in DNA, despite using a small protein *motif*.

A more detailed drawing of the same specific complex between Zif268 and DNA is shown in Fig. 8.6(a). There we can see how three

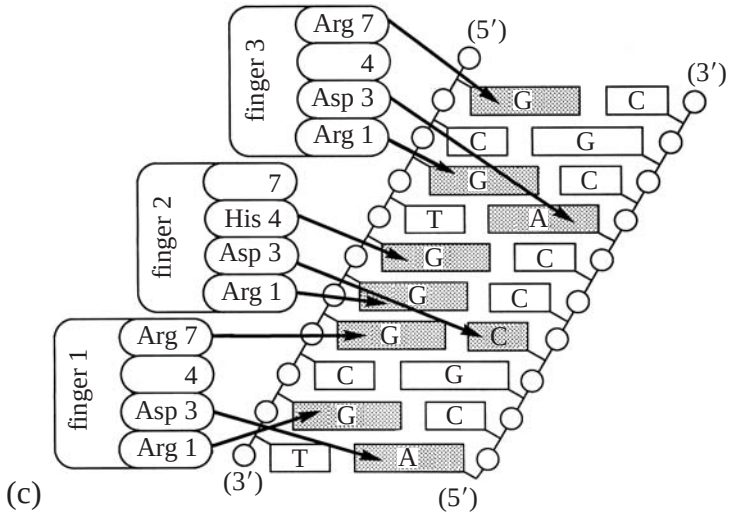


Figure 8.6 (c) An unrolled view of the DNA sequence which is bound to Zif268, showing which amino acids from fingers 1, 2 and 3 contact which bases. Note that Asp 3 of finger 2 and Arg 7 of finger 1 contact different bases of the same base-pair; in this way, the binding sites of neighboring fingers overlap (as they do in other cases of zinc-fingers).

successive zinc-fingers coil through space in a right-handed sense, so as to follow the path of the major groove. Each zinc-finger inserts an α-helix into the groove at the same, fairly steep, 'angle of attack', and then joins to the next finger along the chain through a flexible peptide linker. Zif268 is thought to function as an activator of transcription, because it is made by the cell in response to certain growth factors, which induce the cell to grow rapidly and divide.

Let us now examine the structure of a zinc-finger in even more detail. Within any finger, part of the peptide chain wraps into an α-helix and part into a β-sheet, as shown in Fig. 8.6(b). The α-helix is bound to the β-sheet by hydrophobic contacts between amino acids leucine, phenylalanine and tyrosine – here shown as discs labelled in the single-letter code as L, F and Y, respectively – and also by a zinc ion that binds to two histidine (H) amino acids from the α-helix, and two cysteine (C) amino acids from the β-sheet. The binding of zinc stabilizes the folding of the peptide so as to present an unique recognition surface to the DNA. The zinc-finger in fact is very small in comparison with other proteins, such as the 434 repressor shown in Fig. 8.2. Zinc-fingers may be found in a wide variety of DNA- and RNA-binding proteins from animals or plants, but only rarely in DNA-binding proteins from bacteria – perhaps because the ancestors of those simple organisms could not protect the zinc-cysteine bond from the damaging effects of a high-oxygen environment.

How does a zinc-finger recognize base sequences within the major groove of DNA? Specific examples of such recognition are shown in Fig. 8.6(c) for the three zinc-fingers of Zif268. Both fingers 1 and 3 recognize bases along one strand of sequence GCG, by means of hydrogen bonds between Arg (arginine) 1 and 7 and atoms on the major-groove edge of guanine G: see Fig. 4.12 for more detail. The base C in each triplet GCG is not recognized directly by either finger.

Finger 2 of the same protein recognizes bases of a sequence TGG along the same strand, rather than GCG. This finger lacks an amino acid Arg at position 7, and so it cannot easily recognize G as the first base; but finger 2 now contains an amino acid His (histidine) at position 4, that recognizes the edge of the second base as G by a hydrogen bond. All three fingers have Asp at position 3, that contacts an A or C on the complementary strand. Because of this Asp 3 contact, the binding sites of finger 1 and finger 2 overlap. All three fingers together recognize a ten-base-pair sequence GCGTGGGCGT, with more strength and specificity than for any one of these zinc-fingers considered separately.

Zinc fingers for their size would seem to be very efficient at recognizing a long series of base-pairs in the DNA, because they can

follow the right-handed, twisted path of the major groove. The end-to-end tethering of several modular zinc-finger units enables them to recognize target DNA sequences that are not necessarily symmetric – in contrast to the 434 repressor, met repressor and bZIP examples, where the 'reading heads' are held together by two-fold rotationally symmetric structures.

A final group of sequence-specific DNA-binding proteins from animals are the 'receptors' for small 'signalling-molecules' such as hormones. Once those proteins have bound a specific hormone molecule, they can stimulate the transcription of a target gene in the close vicinity of their specific DNA binding site. The cells of animals may contain many different types of signalling receptor, each of which will recognize some specific 'ligand' – a small molecule such as androgen, estrogen, vitamin D or vitamin A. One part of the receptor protein is specific for recognition of the small ligand,while another part is specific for recognition of the DNA. Crystal structures show how those receptor proteins use an α-helix, stabilized and oriented by zinc binding, to make specific base-pair contacts in the major groove.

Surprisingly, many of these hormone-receptor proteins bind to the *same* DNA sequence, yet they control different genes! In this case, gene-specificity arises from the recognition of the unique *spacing* and *orientation* between two separate, but nearly identical DNA sequences.

Let us first consider a well-known example: the estrogen receptor (Fig. 8.7(a)). Overall, the estrogen receptor is somewhat similar to the 434 repressor protein, in that it binds to two identical 'half-sites' of the DNA target, which run in *opposite directions*. Also, it consists of two identical protein molecules connected together in a head-to-head fashion. The entire estrogen receptor protein and its recognition half-sites are two-fold rotationally symmetric, just as for the 434 repressor; but its half-sites are each six base-pairs long, and they are separated by any three base-pairs.

Next, let us consider the vitamin D receptor (Fig. 8.7(c)). Again, two identical protein molecules recognize two identical DNA half-sites of six base-pairs, separated by three base-pairs; but in this case those half-sites run in the *same* direction, and so the protein-to-protein contact is of a head-to-tail kind.

Finally, let us consider the thyroid receptor (Fig. 8.7(b)). Here the DNA recognition half-sites are almost the same as for the estrogen and vitamin D receptors; except now they are separated by *four* base-pairs, and they run in the same direction. Again, two protein molecules contact one another so that the 'reading heads' adopt the right relationship. Two other examples, shown in Fig. 8.7(d) and

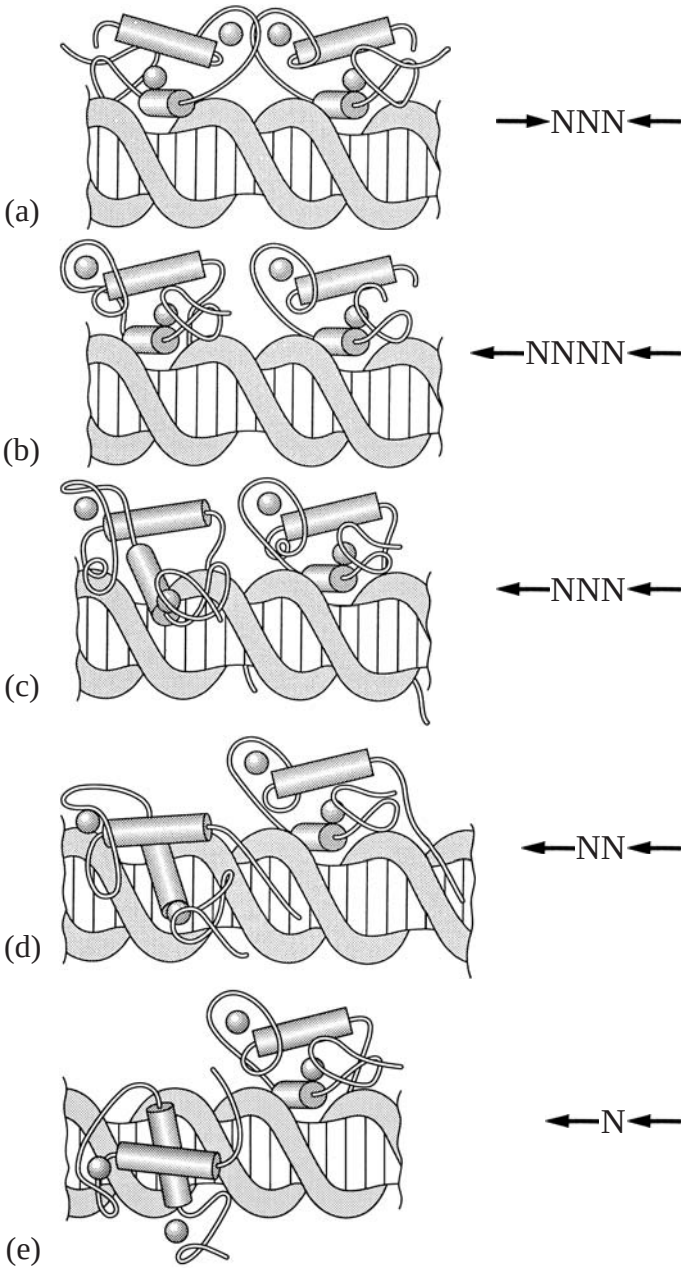


Figure 8.7 Examples of how protein-to-protein interactions are used to recognize the spatial pattern of DNA recognition sites. The steroid receptors, as represented by the estrogen receptor, bind to 'inverted' repeats, but other receptors bind to 'tandem' repeats. Receptor proteins for (a) estrogen, (b) thyroid, (c) vitamin D_3, (d) retinoic acid, (e) 9-cis-retinoic acid all bind to repeats of the same hexameric sequence, 'AGGTCA'. However, they do not confuse these sites, as each requires a unique pattern of protein-to-protein contacts. The schematics on the right show the recognition sites as directional arrows, separated by different numbers of base-pairs. These pictures show only the DNA-binding parts of the receptors; the hormone-binding parts are not shown.

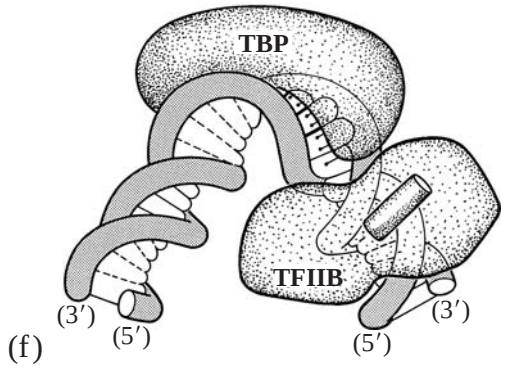


Figure 8.7 (f) The symmetry of the TATA–TBP complex is broken by interactions with a second protein, TFIIB, which contacts both TBP and upstream DNA. The TATA element has a rough two-fold rotational symmetry, and the TBP could in principle align with it in either of two directions; however, binding of TFIIB orients the complex by protein-to-protein and protein-to-DNA contacts. This complex signposts the direction for the start of transcription.

(e), remain similar to the vitamin D and thyroid receptors, except the half-sites are separated by two or one base-pairs, respectively.

To conclude, let us now return briefly to our discussion of the binding of TBP to an eight-base-pair TATA-containing sequence. We drew a picture of TBP in Fig. 4.14 as if the entire arrangement – both protein and DNA – had two-fold rotational symmetry. But to be precise, although the TBP protein has approximate two-fold rotational symmetry, the DNA sequence to which it binds need not have such symmetry; for example, its sequence might be TATAAAAT. Now TBP usually has a 'partner protein' on the promoter DNA for any specific gene, called TFIIB. That partner protein binds specifically to just one end of TBP, and simultaneously to a G–C-rich DNA sequence that precedes the TATA *motif*. Such a unique, well-directed arrangement ensures that the transcription of the promoter-associated DNA takes place in the correct direction.

In summary, we have seen how Nature has devised many distinct means of recognizing a specific DNA sequence, using a variety of protein structural elements. It turns out that typically two-thirds of the atomic contacts in protein-DNA complexes arise from a close fit of similar surfaces, while roughly one-third arise from hydrogen-bonded contacts, that may form either directly between amino acids and the DNA, or else through intervening water molecules.

The direct recognition of a DNA sequence by hydrogen bonds to either amino acids or water is often called 'direct readout'. By

contrast, 'indirect readout' results from some subtle recognition of a DNA-base sequence, through the ability of the DNA to change its shape, for example by curving or twisting (as described in Chapter 4) in order to fit the protein better. Since the greater proportion of protein-to-DNA contacts are hydrophobic, we can see that 'indirect readout' could make an important contribution to the specificity of binding of any protein to DNA.

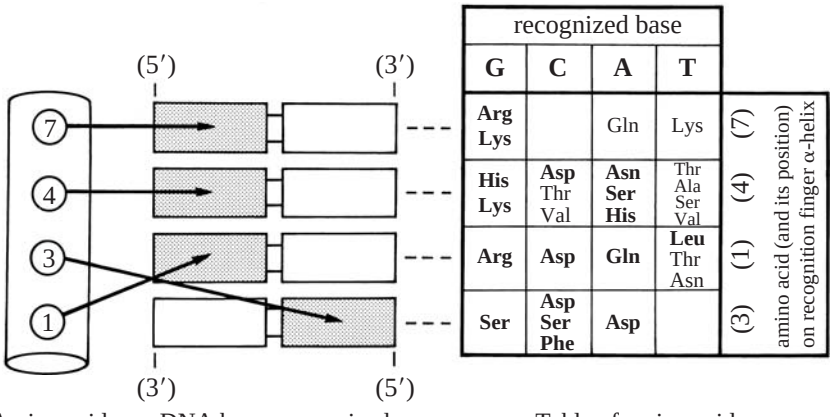
Now certain common modes of DNA recognition can be used by diverse proteins, say where either an α-helix or a β-sheet may be buried in the major groove to make an optimal match of the two surfaces. Might these rough similarities suggest some kind of 'recognition code', for predicting what specific sequence of DNA will be recognized by a particular sequence of amino acids? The simple pattern of recognition of three-base-pair DNA sequences by individual zinc-fingers in Zif268 suggests a special recognition code for the zinc-finger class of proteins. Indeed, Fig. 8.6(c) shows that either three or four successive bases of DNA may be recognized individually by amino acids at positions 7, 4, 3 and 1 on the zinc-finger α-helix, that dips into the major groove. Thus, an arginine (Arg) at position 7 recognizes a G base at the 5′ end of four base-pairs GXXX, while a histidine (His) at position 4 recognizes a G in the next base-pair, XGXX. Also, an Arg at position 1, on the very tip of the α-helix, recognizes a G in the third base-pair of XXGX, while an aspartic acid (Asp) at position 3 'reaches over' and recognizes either an A or a C from the opposite strand in the fourth base-pair, XXXA or XXXC.

Based on these promising ideas, several attempts have been made to design novel zinc-finger proteins, and to deliver them to therapeutic targets. For example, Andrew Jamieson and colleagues have made many random mutations in finger 1 of the protein Zif268, and have found that two different sequences, GTG or TCG, can replace the original GCG if Arg 1 and Arg 7 are changed to other amino acids. In a more extensive study, Yen Choo and Aaron Klug have made many random mutations so as to select for a series of three different zinc-fingers, that will bind specifically to the sequence GCAGAAGCC, which is present in certain cancer cells but not much in normal cells. They first selected for different zinc-fingers which would bind preferentially to each of the three short sequences GCA, GAA and GCC; then they combined three chosen zinc-fingers into a long protein molecule, which can indeed recognize all nine base-pairs together. Carl Pabo and colleagues have also used similar methods to find coding patterns that match the sequence of DNA to the four most relevant amino acids (1, 3, 4 and 7) of the fingers. Some of their results are shown in Fig. 8.8. For

example, the base G in each of the four targeted positions may be recognized by Arg, His, or Lys at α-helical positions 7, 4 and 1.

Could we also design variants of the other proteins described above, such as 434 repressor or the hormone receptors, so that by replacing just a few amino acids, they might pick out slightly different sequences of DNA? In the case of zinc fingers, it seems that the recognition α-helix always lies in much the same orientation relative to the major groove of the DNA, which makes possible the simple scheme of Fig. 8.8. But the situation for those other proteins is far more complicated, because their α-helices can adopt a wide range of inclinations when they dock onto the major groove of DNA. Any conceivable recognition scheme would therefore have to specify the overall geometry by which the critical α-helix is located within the major groove. Finally, as another severe difficulty, it has been found that the interactions of amino acids with base-pairs are not in general of a one-to-one kind, as they are in Fig. 8.8; instead they are context-dependent in ways that are presently difficult to predict.

Since the task of designing novel proteins seems so daunting, many workers have chosen instead to design novel *chemical*



recognized base				
G	**C**	**A**	**T**	
Arg **Lys**		Gln	Lys	(7)
His **Lys**	**Asp** Thr Val	**Asn** **Ser** **His**	Thr Ala Ser Val	(4)
Arg	**Asp**	**Gln**	**Leu** Thr Asn	(1)
Ser	**Asp** **Ser** **Phe**	**Asp**		(3)

amino acid (and its position) on recognition finger α-helix

Table of amino acids at specific positions on recognition helix which recognize specific bases

Figure 8.8 A picture-table which summarizes the frequently observed contacts between DNA and designed zinc-fingers, as determined by Choo and Klug and by Miller and Pabo. The pattern of contacts between amino acids and bases is shown in the simple schematic on the left. Each finger can potentially contact four bases, and the binding sites of consecutive fingers may overlap, as they do for instance in Fig. 8.6(c). The table on the right indicates which bases are recognized by which amino acids in the four recognition positions. Bold type denotes contacts seen in crystal structures. (See Table 1.1 for the three-letter code for amino acids.)