to limit ourselves to studying a few, well characterized examples and to point out the salient principles.

Today, the best characterized information about specific protein–DNA interactions in a living cell lies within the detailed structures of many different protein–DNA complexes, at near-atomic resolution. Since 1990, hundreds of such large complexes between protein and DNA have been studied in detail by the X-ray diffraction of single crystals, and also by nuclear magnetic resonance spectroscopy of molecules in solution (see Chapter 9). Those structures now provide a highly useful and reliable set of data, by which we can understand in part the biological functions of DNA and protein in living cells.

One could perhaps organize a survey of such protein–DNA structures according to the scheme shown in Fig. 8.1, where 'repressors' fall into one class, 'activators' into another, and 'unwinding proteins' into another. Yet it seems that various proteins of widely different structure may lie within each of these biological classes. For example, repressors can recognize DNA by means of an α('alpha')-helix, by means of a β('beta')-sheet, or by other kinds of amino-acid 'motif'. Similarly, activators and unwinding proteins may recognize the DNA through a wide variety of protein structures. For example, in many eukaryotic regulators of transcription, short peptide loops are stabilized through the coordination of zinc metal: these are the

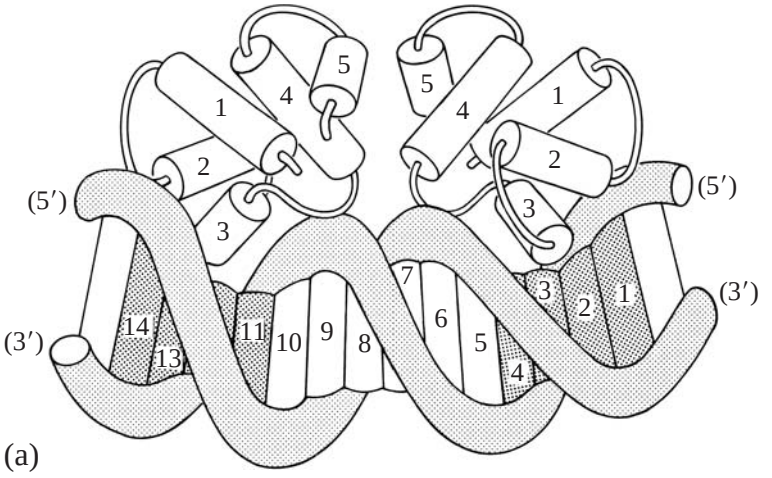


Figure 8.2 (a) A schematic drawing of the 434 repressor protein as it binds to two turns of DNA. The 434 protein consists of two identical halves, which are related by two-fold rotational symmetry. Within each half, a long peptide chain folds into five numbered α-helices, which are drawn as a series of short cylinders connected by flexible linkers. One of the α-helices, '3', lies deep within the major groove of DNA on each side of the complex, where its amino acids can contact directly the four base-pairs which are shaded.

'zinc fingers', and they occur in a very wide range of regulatory proteins of varied structure and function.

Hence, it seems more logical to organize our survey according to the general motif by which a protein recognizes DNA: whether it be by an α-helix, by two α-helices, by a β-sheet, or by a zinc-finger. Examples of each of these classes of 'reading head' will be described in detail below, while references to other important protein–DNA structures of a similar kind are listed in the bibliography. We shall also discuss some synthetic compounds which have been designed to recognize particular DNA sequences and to bind to them strongly.

Let us start our survey of protein–DNA interactions with the simplest kind of recognition, where an α-helix from a protein fits snugly into the major groove of a small piece of DNA, and makes direct contact with a few bases there. The best-known example of that kind of structure is shown in Fig. 8.2(a), where the 434 repressor

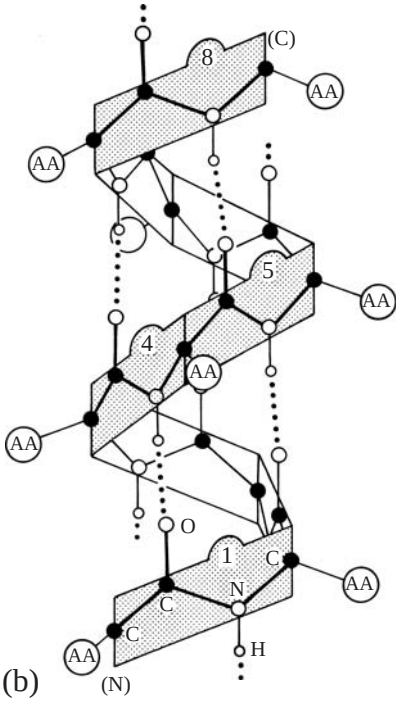


Figure 8.2 (b) A detailed explanation, at a much larger scale, of how a peptide chain folds into an α-helix. Each rigid peptide unit makes two hydrogen bonds (dotted lines) to other peptide units, three steps away along the chain, so as to fold the entire assembly into a regular right-handed spiral. Many different amino acids (AA) may be accommodated on this spiral structure, since the amino-acid side chains protrude out and away from the central peptide core. (The conventional numbering system for peptides starts at the N (amino) end and finishes at the C (carboxyl) terminus.)

protein – consisting of two identical molecules: a 'dimer' – is seen to bind to two successive turns of DNA. Within a small part of each turn, this protein inserts an α-helix, numbered '3' in the picture, into the major groove of the DNA, and is thereby able to recognize accurately the identity of the base-pairs there. A less detailed picture of the same thing was shown in Fig. 4.11, with regard to the bending of DNA in the very center of the complex, near base-pairs 5 to 10. Here we shall focus on specific recognition of DNA by the 434 protein at each end of the complex, near base-pairs 1 to 4 and 11 to 14, which are shaded in the picture.

Before proceeding to examine in detail the most important aspects of this protein–DNA complex, let us explain briefly some well-known features of protein structure, for the benefit of readers who are not already trained in biochemistry at a university level. In Chapter 1, we learned that every protein is manufactured as a long chain of peptides, with an amino acid attached to each peptide as prescribed by the DNA sequence, according to the three-base Genetic Code set out in Table 1.1. We also learned that there are 20 kinds of amino acid. We now need to recall that some of them are large and some are small; some carry an electric charge while others are neutral; and some are hydrophobic while others are hydrophilic.

Any long chain of peptides as found in most proteins will contain a broad mixture of amino acids, and it may be positively charged in one part of the chain but negatively charged in another; or hydrophobic in one part but hydrophilic in another. Different amino acids within the chain may then attract or repel one another in a complex way, so that a certain long sequence of amino acids can specify, for example, the entire detailed structure of the 434 protein shown in Fig. 8.2(a).

The process of folding any long polypeptide chain into its compact, final state must clearly be complex, as there will be a few, key interactions between the amino acid residues amongst an astronomically large number of possibilities. Still, one can understand certain key features, such as the stability and interaction of secondary structural elements, or '*motifs*', like β-strands or α-helices, by examining the folded structures. For example, as shown in Fig 8.2(a), each of the two identical halves of the 434 protein consists of five small cylinders of various lengths, which are joined at their ends by flexible linkers to other cylinders of the same kind. Each of these cylinders actually contains a short length of peptide chain, that is wrapped into a single-stranded spiral or α-helix, as shown with more detail in Fig. 8.2(b).

Within each α-helix, successive rigid peptides coil into a right-handed spiral having 3.6 peptides per helical turn. Thus, you can see in Fig. 8.2(b) that after two turns of the coil, peptide 8 lies almost

directly above peptide 1; and indeed it would lie *exactly* above it if the spiral were to contain 3.5 peptides per turn, instead of the actual 3.6. Each peptide is held weakly to its neighbors along the next turn of the chain by two hydrogen bonds, which run almost parallel to the helix axis and are shown here as dotted lines. Each of these bonds connects an oxygen atom (large open ball) in one peptide unit to a hydrogen atom (small open ball) in a neighboring peptide unit. Different amino acids, shown here as large balls labeled 'AA', may be attached along the outside of the central polypeptide core, something like large decorations on a small Christmas tree. This simple spiral structure for polypeptide chains was discovered in 1950 by the chemist Linus Pauling, when he was at home sick one day, and to occupy himself made models for proteins by drawing a polypeptide chain on a piece of paper and then rolling it up.

As mentioned above, the two identical parts of the 434 repressor protein are made from several such α-helices, which are numbered '1' to '5' in Fig. 8.2(a). The spaces between them are tightly filled by the amino acids that decorate their surfaces. Two of these α-helices, numbered '3', fit snugly into the major grooves of the DNA on either side of the complex. The amino acids which protrude from the α-helices '3' make direct contacts with base-pairs within the major groove of the DNA, in locations marked 1 to 4 or 11 to 14. The rest of the 434 protein holds these two recognition helices '3' the right distance apart and in the correct orientation, so that they fit well into both major grooves of the gently bent DNA, and can probe the edges of base-pairs there so as to determine their identities.

By binding tightly to DNA at just a few sequences in a cell, the 434 protein acts as a repressor of RNA synthesis for certain genes in a bacterial virus called 434. Its specific binding to DNA helps to decide whether the virus lyses (i.e. ruptures) and kills the bacterium which it infects, or just grows peacefully along with it. Thus, the biological action of 434 protein lies in its ability to recognize just one or a few DNA sequences from all others in a viral or bacterial chromosome.

How do those two α-helices '3' recognize a preferred DNA sequence, once an overall 'docking' of the protein onto DNA has been made? The sequence of base-pairs to which a 434 repressor binds most tightly is shown in Fig. 8.3(a). There we can see that each α-helix '3' binds to a base sequence ACAA in positions 1 to 4, or to its equivalent TTGT in positions 11 to 14. A more detailed view of the specific interaction is shown in Fig. 8.3(b), where we can see that four amino acids from helix '3', namely numbers 27, 28, 29, and 33, bind to each of the four base-pairs A1–T1′, C2–G2′, A3–T3′, and A4–T4′.

Further details of these close contacts between protein and DNA are shown in Fig. 8.3(c) and (d). Figure 8.3(c) shows which amino

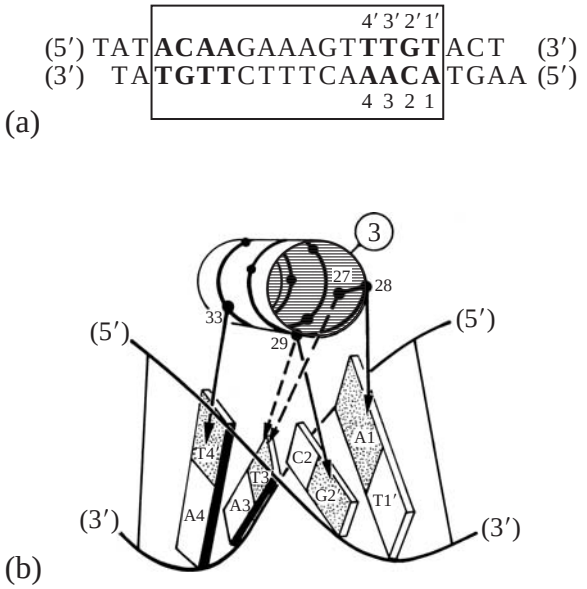


Figure 8.3 (a) A DNA base sequence which is recognized specifically by the 434 repressor protein. The four base-pairs on either end of this 14-base-pair sequence are shown in bold letters: they correspond to the four shaded base-pairs on each end of the DNA shown in Fig. 8.2(a), which are contacted directly by amino acids from α-helix '3'. (b) A more detailed view of how amino acids 27, 28, 29, and 33 from α-helix '3' contact different base-pairs within the major groove, at the sequence ACAA or TGTT shown in bold in (a). Hydrogen bonds between amino acids and base-pairs are drawn as continuous arrows, while hydrophobic contacts are drawn as dashed arrows.

acids connect to which base-pairs, while Fig. 8.3(d) shows which bases connect to which amino acids on the unrolled α-helix. For example, Thr (threonine) 27 connects to base T3′ by a hydrophobic contact in the major groove; Gln (glutamine) 28 connects to base A1 by a hydrogen bond in the major groove; Gln 29 connects to bases G2′ and T3′ by a hydrogen bond to G2′ and by a hydrophobic contact to T3′; while Gln 33 connects to base T4′ by a hydrogen bond. All contacts made by hydrogen bonds, for example N–H to O or O–H to O, are shown schematically here as solid lines, while a hydrophobic contact (for example, CH_3 of the amino acid Thr to CH_3 of base T) is drawn as a dashed line. The detailed chemical formulae of the 20 amino acids may be found in any biochemistry text; and a few were shown in Chapter 4.

Other important contacts between the 434 protein and DNA are made between amino acids and DNA backbone phosphates, as shown also in Fig. 8.3(c). For example, Ser (serine) 30 makes a hydrogen bond to a phosphate just beyond base T4′, while

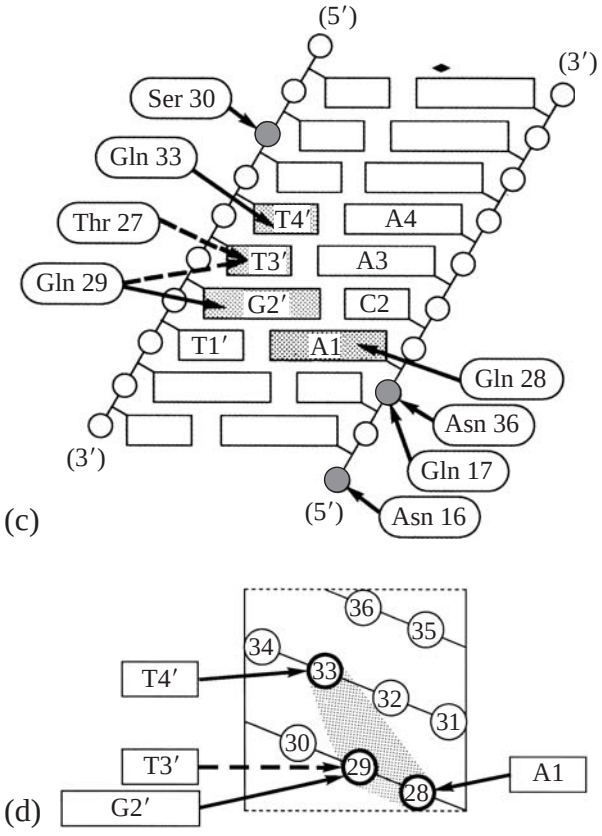


Figure 8.3 (c) Here the DNA helix has been 'unrolled' to display the major groove, and to show clearly which amino acids from the 434 protein contact which base-pairs or phosphates from the DNA. The small diamond at the top marks the center of two-fold rotational symmetry. (d) Here the protein α-helix '3' has been unrolled to show clearly which base-pairs from the DNA contact which amino acids on it. For example, base T4′ contacts amino acid 33, while base T3′ contacts amino acid 29.

Asn (asparagine) 16 and 36 and Gln 17 make hydrogen bonds to phosphates below base A1. Some of these specific contacts to phosphates come from amino acids in α-helix '2', while others come from the flexible linker which connects α-helices '3' and '4': see Fig. 8.2(a) for an overall three-dimensional view of such interactions.

Although the details shown in Fig. 8.3 are dauntingly complicated, we should note here that many simplifications have actually been made in drawing the pictures. In particular, we have omitted several water molecules that are found by X-ray crystallography at high resolution in the 'crevices' between the DNA and protein surfaces, and which provide many more, indirect hydrogen-bonded contacts between the two.

In summary, the specific interaction between 434 repressor protein and DNA seems to involve a sophisticated mixture of chemistry and three-dimensional geometry, which is known in general by the term 'stereochemistry'. Many different amino acids must first fit together to make a large protein, which is perfectly complementary in its shape to the surface of the DNA formed by base-pairs and phosphates; and then both surfaces must match closely so far as hydrogen bonds and hydrophobic contacts are concerned. Loss of just a few hydrogen bonds or hydrophobic contacts from an optimized protein–DNA complex will usually result in a large loss

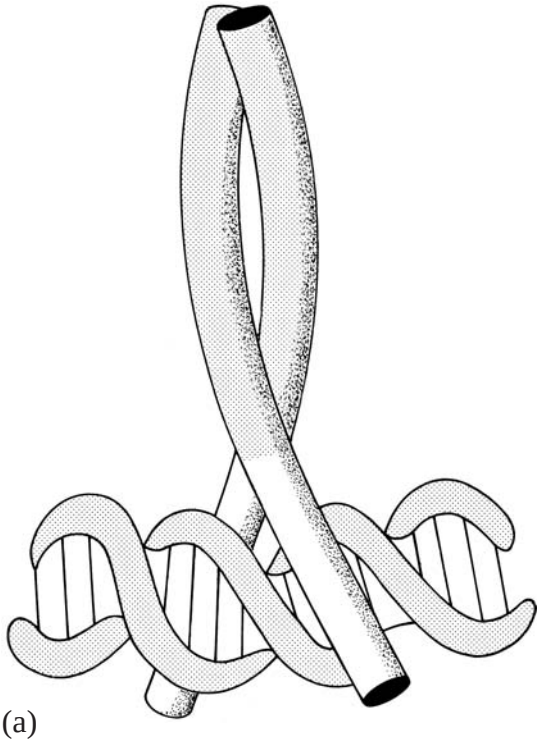
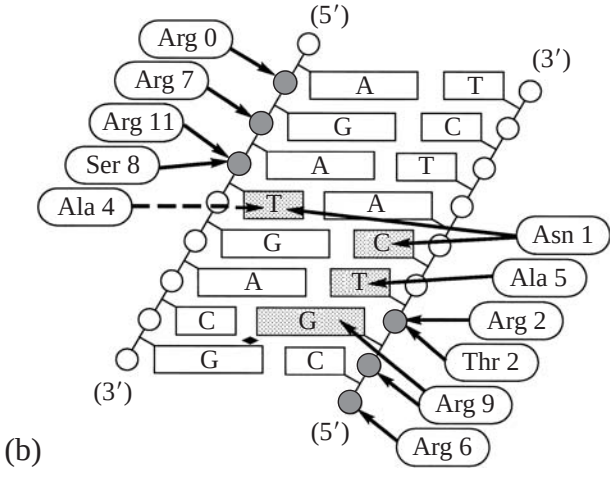


Figure 8.4 (a) A schematic drawing of the bZIP activator protein as bound to DNA. The bZIP protein consists of two related halves, not always identical. Each of these halves folds into a long α-helix that coils gently around the other half in what is known as a 'coiled coil' or 'leucine zipper' arrangement, here shown shaded. One end of each α-helix protrudes into the major groove of the DNA. (b) An unrolled view of a DNA sequence which is bound specifically by the bZIP protein, showing which amino acids from the α-helix contact which base-pairs or phosphates. The small diamond near the bottom indicates the center of two-fold rotational symmetry. (c) An unrolled view of the α-helix from bZIP which is bound within the major groove, showing which bases contact which amino acids. Amino acids encircled by a broken line make contact with phosphates.

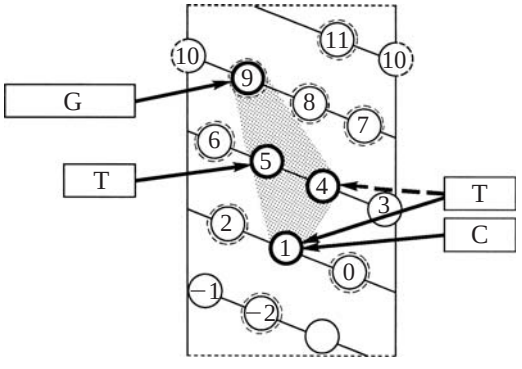
of specificity for the chosen DNA sequence, in comparison with others to which the protein might dock.

Even before the protein and DNA come together, each 'polar group' (e.g., N–H, O–H, N or O) will make hydrogen bonds to surounding water molecules. Those hydrogen bonds to water will be mostly replaced in the protein–DNA complex by hydrogen bonds made directly between protein and DNA. It is essentially that concerted replacement of many different water molecules by protein or DNA, which causes the complex to become stable.

Let us consider next a second geometry by which a protein can bind specifically to DNA: by inserting *two* α-helices into the major groove of any DNA molecule, within just one double-helical turn. A well-known example is the complex between a short piece of DNA and a protein domain called 'bZIP', as shown schematically in Fig. 8.4(a). Two copies of this α-helical bZIP domain associate to form a dimer; and this is part of a larger protein known as GCN4, that is an activator of genes in yeast. In contrast to the 'bean' shape



(b)



(c)

of the 434 repressor (Figs 4.11 and 8.2(a)), the bZIP protein domain has the shape of a letter 'Y', and its two arms are able to probe a single turn of the major groove. In this way it can engage just one double-helical turn of the DNA with two α-helices; while by contrast the bean-like 434 protein will engage two full double-helical turns with its two α-helices.

How are the two α-helices of bZIP held in the correct orientation and location by the rest of the bZIP protein, so as to recognize a specific sequence of DNA? Figure 8.4(a) shows that most of the bZIP protein forms a long pair of identical helices, which coil gently around one another in a left-handed sense. Those two α-helices adhere to each other very tightly, because their inner surfaces are covered by hydrophobic amino acids such as leucine and valine. In fact, the coiling of two α-helices as shown in Fig. 8.4(a) is sometimes called a 'leucine zipper', because many leucine amino acids form a kind of hydrophobic 'stripe' on each of the α-helices. In that picture, the helices are represented as smooth surfaces, but in fact the amino acids are more like knobbly protrusions on each α-helix, which fit into complementary holes between the knobs, on the partner helix. This 'knobs-into-holes' interaction then holds the two parts together like a zipper; and its existence was deduced by Francis Crick more than 50 years ago.

Why do the two α-helices from bZIP coil around one another in a left-handed sense? Why do they not lie side-by-side? Now we explained above in Fig. 8.2(b) that there are 3.6 amino acids for each turn of an α-helix. Suppose for a moment that this number were altered to 3.5. Then we should find that every seventh amino acid (where $2 \times 3.5 = 7.0$) would lie exactly in register with the first, in a view along the axis. In other words, amino acids numbered as 0, 7, 14, 21, … would all lie on a long 'stripe' exactly *parallel* to the helix axis. If the amino acids along those stripes were all hydrophobic in nature for two separate α-helices, then the helices would stick to one another just like two pencils lying side-by-side on a table.

But it turns out that there are actually 3.6 peptides per helical turn, and not 3.5 as in our hypothetical case. The small difference between 3.6 and 3.5 means that peptides which are seven apart on the chain now form a gentle left-handed spiral on the surface of the α-helix, as shown in Fig. 8.2(b). Consequently, the hydrophobic stripes of those two helices will match only if the helices coil around one another in a gentle left-handed fashion.

Once they are each located within the major groove of DNA, over a short stretch of just one double-helical turn, how do the two protruding α-helices of bZIP recognize a specific base-pair sequence there? Figure 8.4(b) and (c) shows the details of specific contacts

made between amino acids and DNA base-pairs, in the same manner as Fig. 8.3(c) and (d). Again, only one set of contacts is shown, because both the protein and the DNA are identical in the two halves: the entire arrangement has 'two-fold rotational symmetry'.

Each α-helix from the bZIP protein contacts four base-pairs of DNA, which are shown in Fig. 8.4(b): namely TGAC on one strand and its complement GTCA on the other. Thus, Ala (alanine) 4 makes a hydrophobic contact with T, while Asn (asparagine) 1 makes a hydrogen bond with T and C, and so on. You can also see where positively charged amino acids from the protein, such as Arg (arginine) 0, 7, and 11 on the left and Arg 2, 6 and 9 on the right, make contacts with negatively charged DNA phosphates along both strands. The mechanism of recognition for bZIP is thus similar to that shown above for 434 repressor, even though the relative locations and orientations of the two α-helices that are used for recognition are different in the two cases.

We will now broaden the scope of our survey, to include other examples where some protein does not deploy an α-helix to recognize base-pairs, but uses instead some other kind of structure or *motif*. For example, the met repressor protein shown in Fig. 8.5(a) inserts a β-sheet into the major groove of DNA, so as to recognize the base-pairs there. This met repressor protein is used by a bacterium to regulate the amount of the amino acid methionine which is made. It binds to the promoter of a gene which controls methionine synthesis, and represses that gene if a certain amount of methionine-derived chemical (called S-adenosyl-methionine) is present in the cell. From close inspection of Fig. 8.5(a), we can see that the met repressor consists of two identical peptide chains that interweave to build a single structural unit. There are α-helices labeled 'A' and 'B', which help to hold a two-stranded β-sheet (long arrows) deep within the major groove. Amino acids which protrude from the β-sheet may then contact DNA base-pairs, so that the met repressor protein can recognize a specific base sequence.

What is the detailed structure of a β-sheet, by way of comparison with that of the α-helix shown in Fig. 8.2(b)? Within any β-sheet, as drawn in Fig. 8.5(b), two or more polypeptide strands, here shown crimped up-and-down, lie side-by-side and close together in a plane. Each of the polypeptides is held weakly to its closest neighbor on a nearby strand by two hydrogen bonds (dotted lines), which connect an oxygen from one peptide unit (large open ball) to a hydrogen from a peptide unit on the other strand (small open ball). Moving along the strand, one finds that different amino acids (AA) protrude alternately above and below the plane of the β-sheet. The picture of Fig. 8.5(b) also shows two more strands which