Part 2

Structure, Function, and Engineering

An Example of Enzyme Catalysis: Serine Proteinases

11

In 1946 Linus Pauling first formulated the basic principle underlying enzyme catalysis, namely, that an enzyme increases the rate of a chemical reaction by binding and stabilizing the transition state of its specific substrate tighter than the ground state. However, for many years it was not generally appreciated that the high affinity of an enzyme for the transition state of a substrate plays a major role in determining substrate specificity as well as the rate of catalysis. In the past few years, kinetic studies of site-directed mutants, combined with x-ray structures, have made it possible to identify unambiguously the role of particular amino acids in both the substrate specificity and the catalytic reaction of an enzyme as well as providing information about the energetic basis of catalysis itself. The full consequences of Pauling's principle emerged only when it was found that mutants designed to change an enzyme's catalytic rate also changed its substrate specificity and vice versa.

In this chapter we shall illustrate some fundamental aspects of enzyme catalysis using as an example the serine proteinases, a group of enzymes that hydrolyze peptide bonds in proteins. We also examine how the transition state is stabilized in this particular case.

Proteinases form four functional families

Proteinases are widely distributed in nature, where they perform a variety of different functions. Viral genes code for proteinases that cleave the precursor molecules of their coat proteins, bacteria produce many different extracellular proteinases to degrade proteins in their surroundings, and higher organisms use proteinases for such different functions as food digestion, cleavage of signal peptides, and control of blood pressure and blood clotting. Many proteinases occur as domains in large multifunctional proteins, but others are independent smaller polypeptide chains. *In vivo* the activity of many proteinases is controlled by endogenous protein inhibitors that complex with the enzymes and block them. The three-dimensional structures of a large number of the smaller proteinases and of their complexes with protein inhibitors have been determined, and this wealth of data allows some general conclusions to be drawn.

All the well-characterized proteinases belong to one or other of four families: serine, cysteine, aspartic, or metallo proteinases. This classification is based on a functional criterion, namely, the nature of the most prominent functional group in the active site. Members of the same functional family are usually evolutionarily related, but there are exceptions to this rule. We

have chosen two **serine proteinases**, mammalian **chymotrypsin** and bacterial **subtilisin**, as representative examples to illustrate one of the catalytic mechanisms leading to proteolysis. Before describing the structures, mechanism, and engineering of these two enzymes, however, we shall define some basic enzymological concepts.

The catalytic properties of enzymes are reflected in K_m and k_{cat} values

Leonor Michaelis and Maud Menten laid the foundation for enzyme kinetics as early as 1913 by proposing the following scheme:

$$E + S \underset{}{\overset{K_d}{\rightleftharpoons}} ES \overset{k_{cat}}{\rightleftharpoons} E + P$$

Enzyme and substrate first reversibly combine to give an **enzyme–substrate** (ES) complex. Chemical processes then occur in a second step with a rate constant called k_{cat}, or the **turnover number**, which is the maximum number of substrate molecules converted to product per active site of the enzyme per unit time. The k_{cat} is, therefore, a rate constant that refers to the properties and reactions of the ES complex. For simple reactions k_{cat} is the rate constant for the chemical conversion of the ES complex to free enzyme and products.

These definitions are valid only when the concentration of the enzyme is very small compared with that of the substrate. Moreover, they apply only to the initial rate of formation of products: in other words, the rate of formation of the first few percent of the product, before the substrate has been depleted and products that can interfere with the catalytic reaction have accumulated.

The Michaelis–Menten scheme nicely explains why a maximum rate, V_{max}, is always observed when the substrate concentration is much higher than the enzyme concentration (Figure 11.1). V_{max} is obtained when the enzyme is saturated with substrate. There are then no free enzyme molecules available to turn over additional substrate. Hence, the rate is constant, V_{max}, and is independent of further increase in the substrate concentration.

The substrate concentration when the half maximal rate, ($V_{max}/2$), is achieved is called the K_m. For many simple reactions it can easily be shown that the K_m is equal to the dissociation constant, K_d, of the ES complex. The K_m, therefore, describes the affinity of the enzyme for the substrate. For more complex reactions, K_m may be regarded as the overall dissociation constant of all enzyme-bound species.

The quantity k_{cat}/K_m is a rate constant that refers to the overall conversion of substrate into product. The ultimate limit to the value of k_{cat}/K_m is therefore set by the rate constant for the initial formation of the ES complex. This rate cannot be faster than the diffusion-controlled encounter of an enzyme and its substrate, which is between 10^8 to 10^9 per mole per second. The quantity k_{cat}/K_m is sometimes called the s**pecificity constant** because it describes the specificity of an enzyme for competing substrates. As we shall see, it is a useful quantity for kinetic comparison of mutant proteins.

Enzymes decrease the activation energy of chemical reactions

The Michaelis complex, ES, undergoes rearrangement to one or several **transition states** before product is formed. Energy is required for these rearrangements. The input energy required to bring free enzyme and substrate to the highest transition state of the ES complex is called the **activation energy** of the reaction (Figure 11.2). In the absence of enzyme, spontaneous conversion of substrate to product also proceeds through transition states that require activation energy. The rate of a chemical reaction is strictly dependent on its

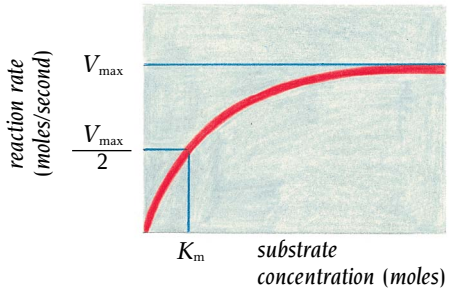


Figure 11.1 A plot of the reaction rate as a function of the substrate concentration for an enzyme catalyzed reaction. V_{max} is the maximal velocity. The Michaelis constant, K_m, is the substrate concentration at half V_{max}. The rate v is related to the substrate concentration, [S], by the Michaelis–Menten equation:

$$v = \frac{V_{max} \times [S]}{K_m + [S]}$$

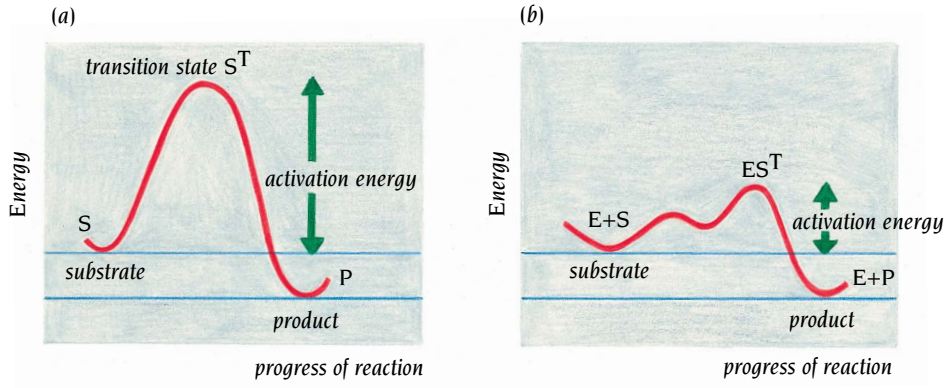


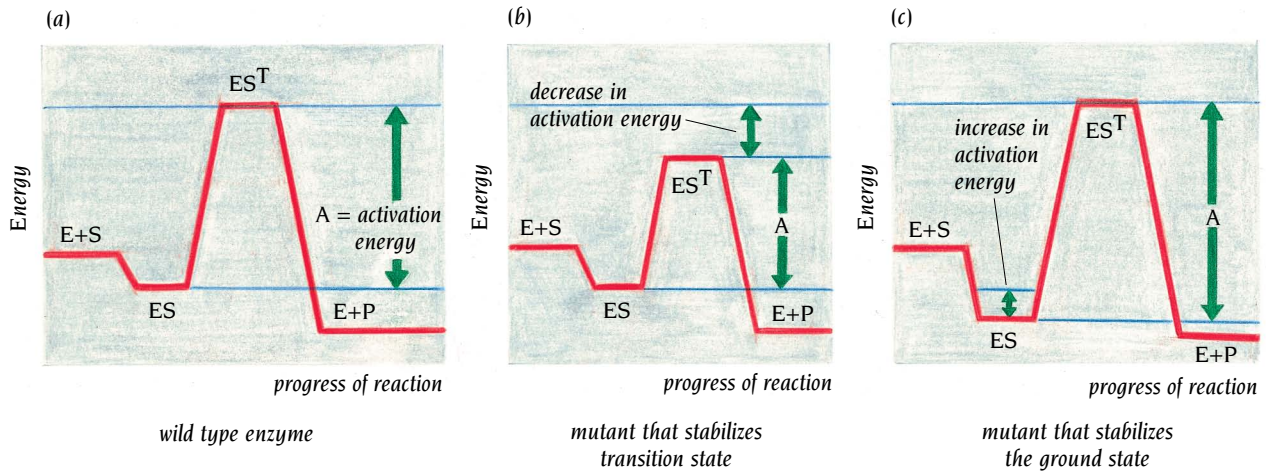
Figure 11.2 Enzymes accelerate chemical reactions by decreasing the activation energy. The activation energy is higher for a noncatalyzed reaction (a) than for the same reaction catalyzed by an enzyme (b). Both reactions proceed through one or several transition states, S^T. Only one transition state is shown in (a), whereas the two bumps in (b) represent two different transition states.

activation energy, and the more than 1 millionfold enhancement of rate achieved by enzyme catalysis results from the ability of the enzyme to decrease the activation energy of the reaction (Figure 11.2).

This decrease in activation energy is achieved by enzymes in several different ways: for example, by providing catalytically active groups for a specific reaction mechanism, by binding several substrates in an orientation appropriate to the reaction catalyzed, and, most importantly, by using the differential binding energy of the substrate in its transition state compared with its normal state. The activation energy for the conversion of ES to E + P is lower if the enzyme binds more tightly to the transition state of S than to its normal structure (Figure 11.3). The higher affinity of the enzyme for the transition state makes the transition energetically favorable and thus decreases the activation energy. If, on the other hand, the enzyme were to bind the unaltered substrate more strongly than the transition state, the decrease in binding energy on the formation of the transition state would increase the activation energy and catalysis would not be achieved (see Figure 11.3). It is therefore catalytically advantageous for the enzyme's active site to be complementary to the transition state of the substrate rather than to the normal structure of the substrate.

Since this differential binding energy relates to the complete substrate molecule, including groups that determine the substrate specificity, it is obvious that specificity and catalytic rate are interdependent. The importance of the differential binding energy for catalysis is nicely illustrated by the recent production of antibodies with catalytic activity. Such antibodies were raised against small hapten molecules that simulate a transition state structure for a specific chemical reaction, such as ester hydrolysis. These antibodies not only bound the transition state more tightly than the normal structure of the ester, but they also exhibited significant catalytic activity even though they had not been selected to have any catalytically competent residues in the binding site.

Figure 11.3 One of the most important factors in enzyme catalysis is the ability of an enzyme to bind substrate more tightly in its transition state than in its ground state. The difference in binding energy between these states lowers the activation energy of the reaction. This is illustrated by energy profiles for an enzyme in its wild-type form (a), for a mutant that stabilizes the substrate in its transition state and therefore decreases the activation energy from ES to the transition state ES^T giving higher rates (b), and for a mutant that stabilizes the substrate in its ground state giving lower rates (c). (Adapted from A. Fersht, *Enzyme Structure and Mechanism*, 2nd ed. pp. 314–315. New York: W.H. Freeman, 1984.)



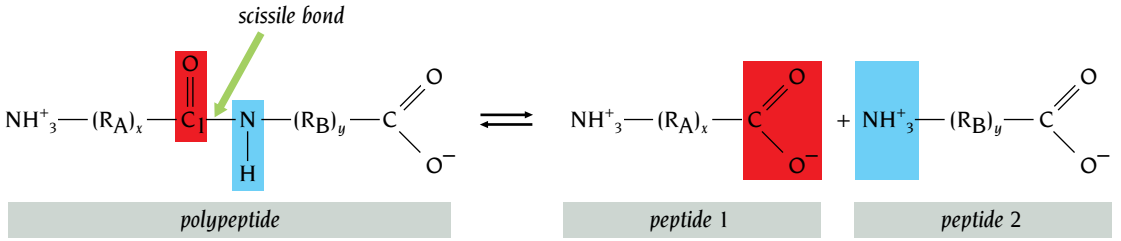


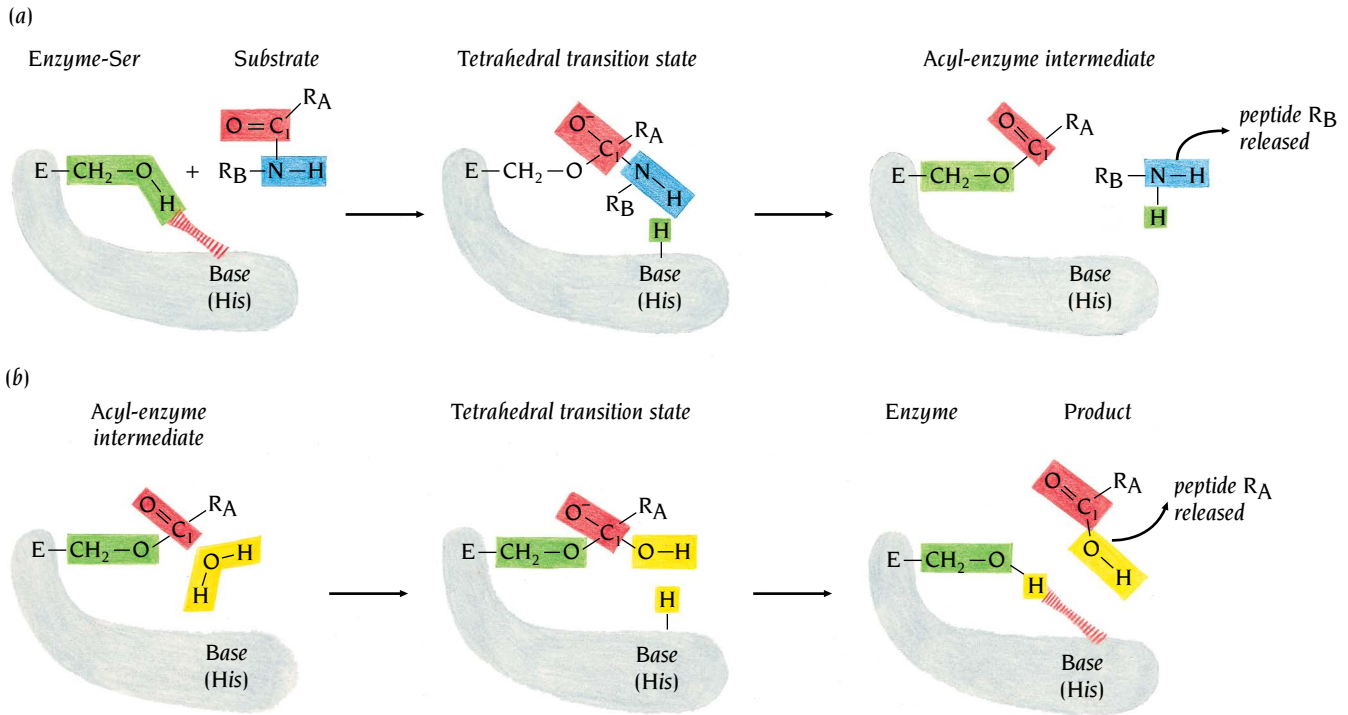
Figure 11.4 Serine proteinases catalyze the hydrolysis of peptide bonds within a polypeptide chain. The bond that is cleaved is called the scissile bond. $(R_A)_x$ and $(R_B)_y$ represent polypeptide chains of varying lengths.

Serine proteinases cleave peptide bonds by forming tetrahedral transition states

The serine proteinases have been very extensively studied, both by kinetic methods in solution and by x-ray structural studies to high resolution. From all these studies the following reaction mechanism has emerged.

A serine proteinase cleaves peptide bonds within a polypeptide to produce two new smaller peptides (Figure 11.4). The reaction proceeds in two steps. The first step produces a covalent bond between C_1 of the substrate and the hydroxyl group of a reactive Ser residue of the enzyme (Figure 11.5a). Production of this acyl-enzyme intermediate proceeds through a negatively charged transition state intermediate where the bonds of C_1 have tetrahedral geometry in contrast to the planar triangular geometry in the peptide group. During this step the peptide bond is cleaved, one peptide product is attached to the enzyme in the acyl-enzyme intermediate, and the other peptide product rapidly diffuses away. In the second step of the reaction, deacylation, the acyl-enzyme intermediate is hydrolyzed by a water molecule to release the second peptide product with a complete carboxy terminus and to restore the Ser-hydroxyl of the enzyme (Figure 11.5b). This step also proceeds through a negatively charged tetrahedral transition state intermediate (Figure 11.5b). What are the structural requirements for the enzyme to perform these reactions?

Figure 11.5 (a) Formation of an acyl-enzyme intermediate involving a reactive Ser residue of the enzyme is the first step in hydrolysis of peptide bonds by serine proteinases. First, a transition state is formed where the peptide bond is cleaved in which the C_1 carbon has a tetrahedral geometry with bonds to four groups, including the reactive Ser residue of the enzyme and a negatively charged oxygen atom. (b) Deacylation of the acyl-enzyme intermediate is the second step in hydrolysis. This is essentially the reverse of the acylation step, with water in the role of the NH_2 group of the polypeptide substrate. The base shown in the figure is a His residue of the protein that can accept a proton during the formation of the tetrahedral transition state.



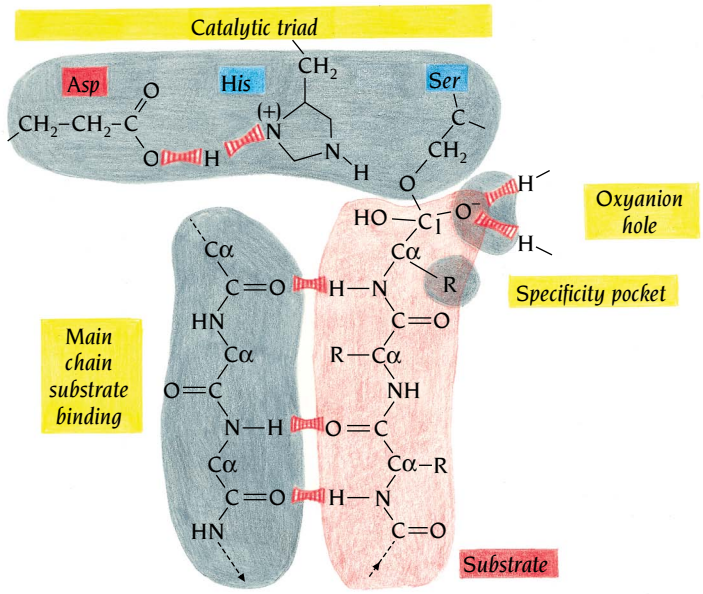


Figure 11.6 A schematic view of the presumed binding mode of the tetrahedral transition state intermediate for the deacylation step. The four essential features of the serine proteinases are highlighted in yellow: the catalytic triad, the oxyanion hole, the specificity pocket, and the unspecific main-chain substrate binding.

Four important structural features are required for the catalytic action of serine proteinases

The serine proteinases have four important structural features that facilitate this mechanism of catalysis (Figure 11.6).

1. The enzyme provides a general base, a His residue, that can accept the proton from the hydroxyl group of the reactive Ser thus facilitating formation of the covalent tetrahedral transition state. This His residue is part of a **catalytic triad** consisting of three side chains from Asp, His, and Ser, which are close to each other in the active site, although they are far apart in the amino acid sequence of the polypeptide chain (Figure 11.6).
2. Tight binding and stabilization of the tetrahedral transition state intermediate is accomplished by providing groups that can form hydrogen bonds to the negatively charged oxygen atom attached to C_1. These groups are in a pocket of the enzyme called the **oxyanion hole** (see Figure 11.6). The positive charge that develops on the His residue after it has accepted a proton also stabilizes the negatively charged transition state. These features also presumably destabilize binding of substrate in the normal state.
3. Most serine proteinases have no absolute substrate specificity. They can cleave peptide bonds with a variety of side chains adjacent to the peptide bond to be cleaved (the scissile bond). This is because polypeptide substrates exhibit a nonspecific binding to the enzyme through their main-chain atoms, which form hydrogen bonds in a short antiparallel β sheet with main-chain atoms of a loop region in the enzyme (see Figure 11.6). One of these hydrogen bonds is long (3.6 Å) in enzyme-substrate complexes but short in complexes that simulate the transition state. This nonspecific binding therefore also contributes to stabilization of the transition state.
4. Even though these enzymes have no absolute specificity, many of them show a preference for a particular side chain before the scissile bond as seen from the amino end of the polypeptide chain. The preference of chymotrypsin to cleave after large aromatic side chains and of trypsin to cleave after Lys or Arg side chains is exploited when these enzymes are used to produce peptides suitable for amino acid sequence determination and fingerprinting. In each case, the preferred side chain is oriented so as to fit into a pocket of the enzyme called the **specificity pocket**.

Convergent evolution has produced two different serine proteinases with similar catalytic mechanisms

These four features all occur in an almost identical fashion in all members of the chymotrypsin superfamily of homologous enzymes, which includes among others chymotrypsin, trypsin, elastase, and thrombin. Reasonably, one might imagine that such a combination of four characteristic features had arisen only once during evolution to give an ancestral molecule from which all serine proteinases diverged. However, subtilisin, a bacterial serine proteinase with an amino acid sequence and, as we will see, a three-dimensional structure quite different from the mammalian serine proteinases, exhibits these same four characteristic features. Subtilisin is not evolutionarily related to the chymotrypsin family of enzymes; nevertheless, the atoms in subtilisin that participate in the catalytic triad, in the oxyanion hole, and in substrate binding are in almost identical positions relative to one another in the three-dimensional structure as they are in chymotrypsin and its relatives. Starting from unrelated ancestral proteins, convergent evolution has resulted in the same structural solution to achieve a particular catalytic mechanism. The serine proteinases, in other words, provide a spectacular example of convergent evolution at the molecular level, which we can best appreciate by explaining in detail the structures of chymotrypsin and subtilisin.

The chymotrypsin structure has two antiparallel β-barrel domains

In 1967 the group of David Blow at the MRC Laboratory of Molecular Biology, Cambridge, UK, determined the three-dimensional structure of chymotrypsin. This was one of the very first enzyme structures known at high resolution. Since then a large number of serine proteinase structures, complexed both with small peptide inhibitors and large endogenous polypeptide inhibitors, have been determined to high resolution mainly by the groups of Michael James, Edmonton, and Robert Huber, Munich.

The polypeptide chain of chymotrypsinogen, the inactive precursor of chymotrypsin, comprises 245 amino acids. During activation of chymotrypsinogen residues 14–15 and 147–148 are excised. The remaining three polypeptide chains are held together by disulfide bridges to form the active chymotrypsin molecule.

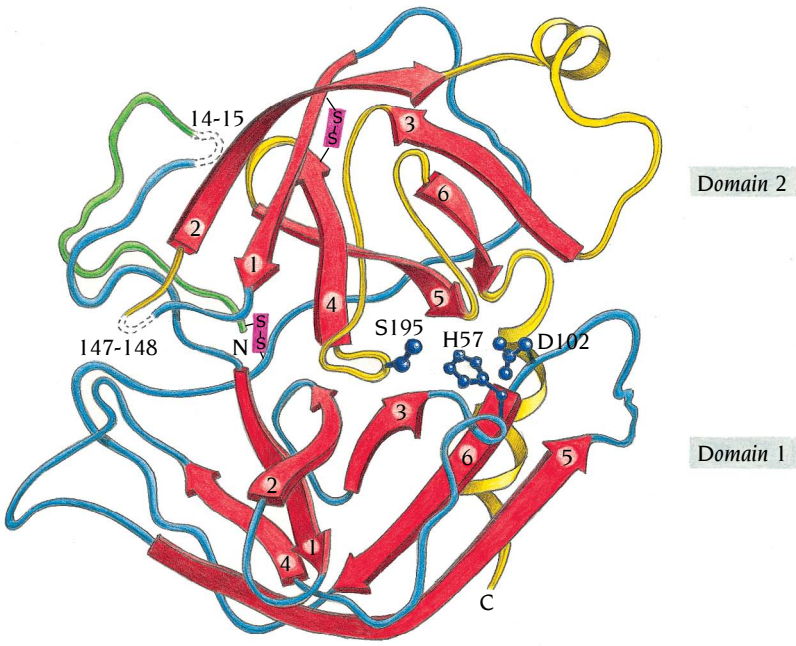
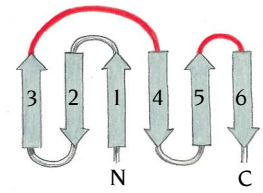


Figure 11.7 Schematic diagram of the structure of chymotrypsin, which is folded into two antiparallel β domains. The six β strands of each domain are red, the side chains of the catalytic triad are dark blue, and the disulfide bridges that join the three polypeptide chains are marked in violet. Chain A (green, residues 1–13) is linked to chain B (blue, residues 16–146) by a disulfide bridge between Cys 1 and Cys 122. Chain B is in turn linked to chain C (yellow, residues 149–245) by a disulfide bridge between Cys 136 and Cys 201. Dotted lines indicate residues 14–15 and 147–148 in the inactive precursor, chymotrypsinogen. These residues are excised during the conversion of chymotrypsinogen to the active enzyme chymotrypsin.

The polypeptide chain is folded into two domains (Figure 11.7), each of which contains about 120 amino acids. The two domains are both of the antiparallel β-barrel type, each containing six β strands with the same topology (Figure 11.8). Even though the actual structure looks complicated, the topology is very simple, a Greek key motif (strands 1–4) followed by an antiparallel hairpin motif (strands 5 and 6).



The active site is formed by two loop regions from each domain

The active site is situated in a crevice between the two domains. Domain 1 contributes two of the residues in the catalytic triad, His 57 and Asp 102, whereas the reactive Ser 195 is part of the second domain (see Figure 11.7).

Inhibitors as well as substrates bind in this crevice between the domains. From the numerous studies of different inhibitors bound to serine proteinases we have chosen as an illustration the binding of a small peptide inhibitor, Ac-Pro-Ala-Pro-Tyr-COOH to a bacterial chymotrypsin (Figure 11.9). The enzyme–peptide complex was formed by adding a large excess of the substrate Ac-Pro-Ala-Pro-Tyr-CO-NH_2 to crystals of the enzyme. The enzyme molecules within the crystals catalyze cleavage of the terminal amide group to produce the products Ac-Pro-Ala-Pro-Tyr-COOH and NH_3^+. The ammonium ions diffuse away, but the peptide product remains bound as an inhibitor to the active site of the enzyme.

This inhibitor does not form a covalent bond to Ser 195 but one of its carboxy oxygen atoms is in the oxyanion hole forming hydrogen bonds to the main-chain NH groups of residues 193 and 195. The tyrosyl side chain is positioned in the specificity pocket, which derives its specificity mainly from three residues, 216, 226, and 189, as we shall see later. The main chain of

Figure 11.8 Topology diagrams of the domain structure of chymotrypsin. The chain is folded into a six-stranded antiparallel β barrel arranged as a Greek key motif followed by a hairpin motif.

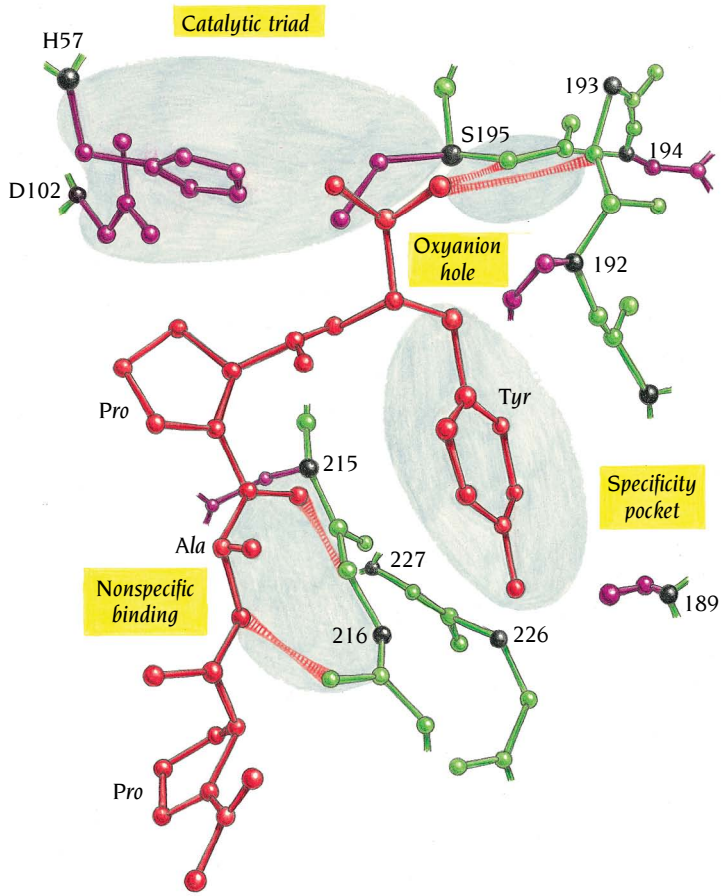


Figure 11.9 A diagram of the active site of chymotrypsin with a bound inhibitor, Ac-Pro-Ala-Pro-Tyr-COOH. The diagram illustrates how this inhibitor binds in relation to the catalytic triad, the substrate specificity pocket, the oxyanion hole and the nonspecific substrate binding region. The inhibitor is red. Hydrogen bonds between inhibitor and enzyme are striped. (Adapted from M.N.G. James et al., *J. Mol. Biol.* 144: 43–88, 1980.)

the inhibitor forms a short stretch of antiparallel β sheet with residues 215–216 of the enzyme forming hydrogen bonds to the NH and CO groups of residue 216.

A closer examination of these essential residues, including the catalytic triad, reveals that they are all part of the same two loop regions in the two domains (Figure 11.10). The domains are oriented so that the ends of the two barrels that contain the Greek key crossover connection (described in Chapter 5) between β strands 3 and 4 face each other along the active site. The essential residues in the active site are in these two crossover connections and in the adjacent hairpin loops between β strands 5 and 6. Most of these essential residues are conserved between different members of the chymotrypsin superfamily. They are, of course, surrounded by other parts of the polypeptide chains, which provide minor modifications of the active site, specific for each particular serine proteinase.

His 57 and Ser 195 are within loop 3–4 of domains 1 and 2, respectively. The third residue in the catalytic triad, Asp 102, is within loop 5–6 of domain 1. The rest of the active site is formed by two loop regions (3–4 and 5–6) of domain 2. As in so many other protein structures described previously, the barrels apparently provide a stable scaffold to position a few loop regions that constitute the essential features of the active site.

Did the chymotrypsin molecule evolve by gene duplication?

Although the two domains of chymotrypsin have similar three-dimensional structures there is no amino acid sequence identity between them. Nevertheless, based on the argument that three-dimensional structure is more conserved than amino acid sequence, it has been suggested that the members of the chymotrypsin superfamily evolved by gene duplication of a single ancestral proteinase domain. The putative ancestral domain, obviously, could not have had the catalytic triad in present-day serine proteinases since the contemporary triad is derived from both domains. However, this is less of an obstacle to the gene-duplication hypothesis than it seems at first sight. The ancestral domain could have been a barrel structure similar to the second domain of chymotrypsin, which contains most of the essential features of the active site, including the reactive serine residue. We also now know from experiments with genetically engineered mutants in which the triad has been abolished that the catalytic triad is not absolutely essential for catalytic activity. As we will see later, these mutants retain some proteinase activity. It is, therefore, quite possible that there was a single ancestral gene specifying a single domain with some catalytic activity. This activity could then have been enhanced by a gene-duplication event followed by mutation and evolution leading to the catalytic triad of today. The fact that the active-site residues that comprise the catalytic triad of chymotrypsin and its relatives are clustered in the same two loop regions of domains 1 and 2 supports such an evolutionary history.

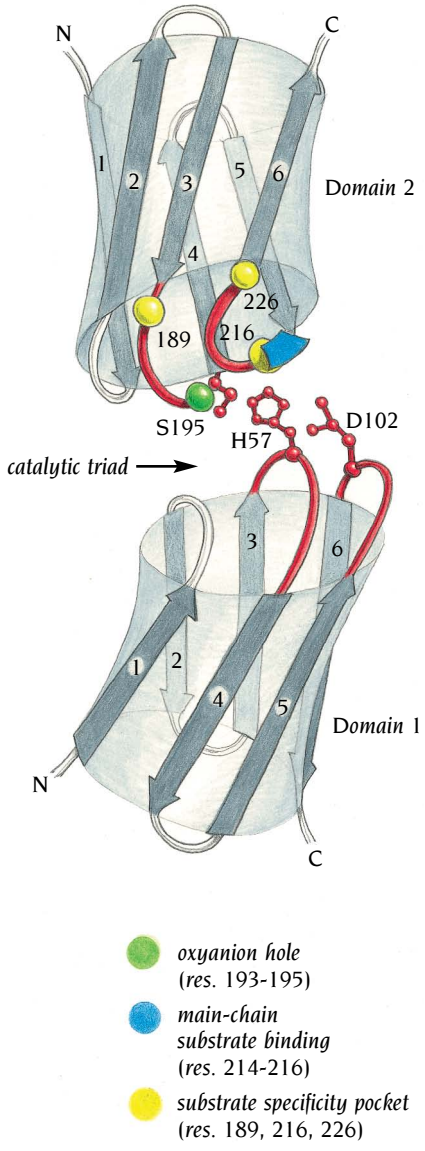


Figure 11.10 Topological diagram of the two domains of chymotrypsin, illustrating that the essential active-site residues are part of the same two loop regions (3–4 and 5–6, red) of the two domains. These residues form the catalytic triad, the oxyanion hole (green), and the substrate binding regions (yellow and blue) including essential residues in the specificity pocket.

Different side chains in the substrate specificity pocket confer preferential cleavage

The serine proteinases all have the same substrate, namely, polypeptide chains of proteins. However, different members of the family preferentially cleave polypeptide chains at sites adjacent to different amino acid residues. The structural basis for this preference lies in the side chains that line the substrate specificity pocket in the different enzymes.

This is nicely illustrated by members of the chymotrypsin superfamily: the enzymes chymotrypsin, trypsin, and elastase have very similar three-dimensional structures but different specificity. They preferentially cleave adjacent to bulky aromatic side chains, positively charged side chains, and small uncharged side chains, respectively. Three residues, numbers 189, 216, and 226, are responsible for these preferences (Figure 11.11). Residues 216

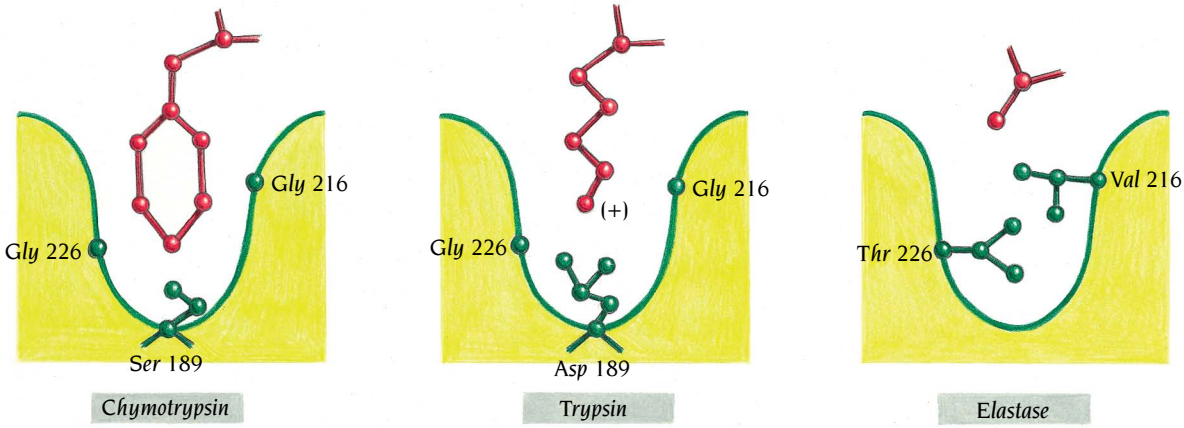


Figure 11.11 Schematic diagrams of the specificity pockets of chymotrypsin, trypsin and elastase, illustrating the preference for a side chain adjacent to the scissile bond in polypeptide substrates. Chymotrypsin prefers aromatic side chains and trypsin prefers positively charged side chains that can interact with Asp 189 at the bottom of the specificity pocket. The pocket is blocked in elastase, which therefore prefers small uncharged side chains.

and 226 line the sides of the pocket. In trypsin and chymotrypsin these are both glycine residues that allow side chains of the substrate to penetrate into the interior of the specificity pocket. In elastase they are Val and Thr, respectively, that fill up most of the pocket with hydrophobic groups (Figure 11.11). Consequently, elastase does not cleave adjacent to large or charged side chains but adjacent to small uncharged side chains instead.

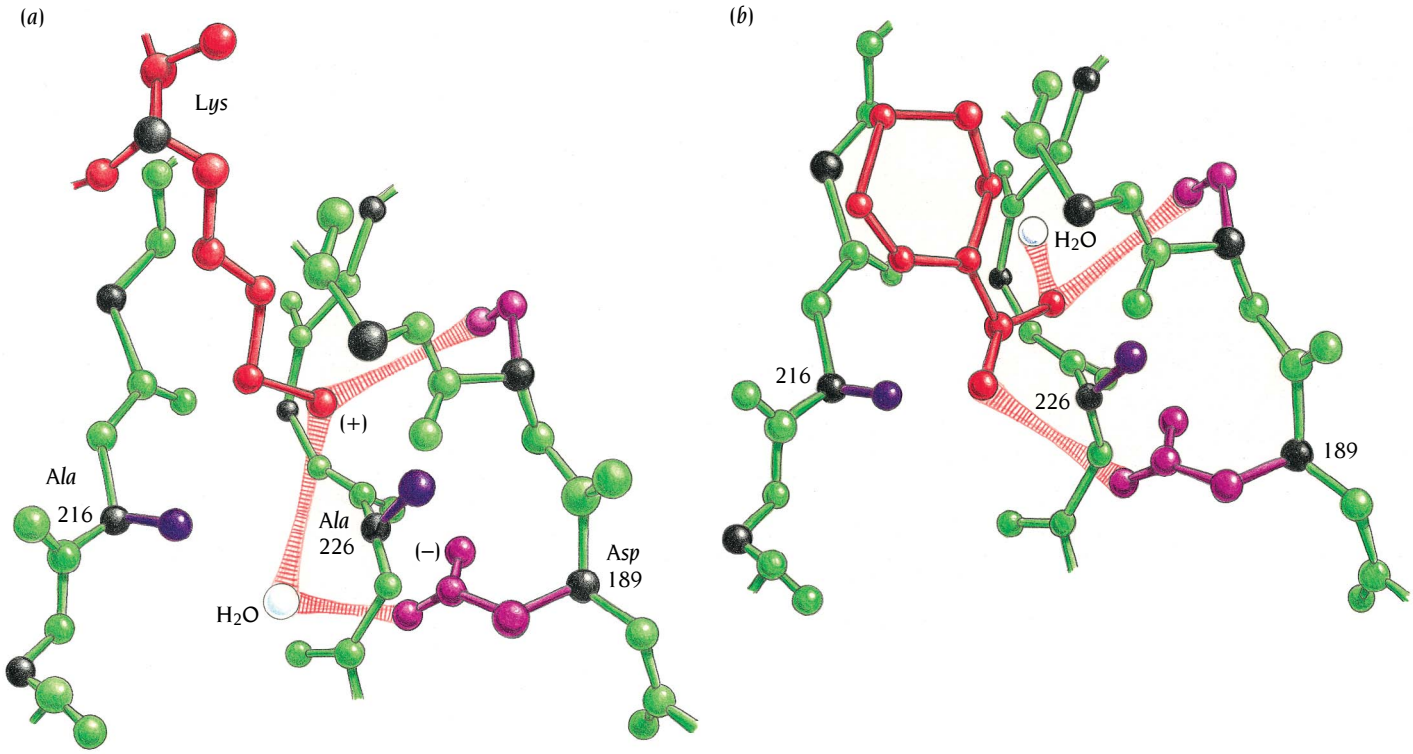
Residue 189 is at the bottom of the specificity pocket. In trypsin the Asp residue at this position interacts with the positively charged side chains Lys or Arg of a substrate. This accounts for the preference of trypsin to cleave adjacent to these residues. In chymotrypsin there is a Ser residue at position 189, which does not interfere with the binding of the substrate. Bulky aromatic groups are therefore preferred by chymotrypsin since such side chains fill up the mainly hydrophobic specificity pocket. It has now become clear, however, from site-directed mutagenesis experiments that this simple picture does not tell the whole story.

Engineered mutations in the substrate specificity pocket change the rate of catalysis

How would substrate preference be changed if the glycine residues in trypsin at positions 216 and 226 were changed to alanine rather than to the more bulky valine and threonine groups that are present in elastase? This question was addressed by the groups of Charles Craik, William Rutter, and Robert Fletterick in San Francisco, who have made and studied three such trypsin mutants: one in which Ala is substituted for Gly at 216, one in which the same substitution is made at Gly 226, and a third containing both substitutions.

Model building shows that both Arg- and Lys-containing substrates should be accommodated by the substrate specificity pocket after these Gly to Ala changes but that some details of the binding mode at the bottom of the pocket would be altered. The Ala 226 substitution would introduce a methyl group in the region where the end of the substrate's side chain binds (Figure 11.12) and would therefore be expected to accommodate Lys better than Arg, since the latter has a longer and more bulky side chain. Based on these steric arguments alone, one would therefore predict that the K_m for an Arg-containing substrate would be larger (less favorable binding) and that the K_m for Lys would be essentially unaltered. The specificity constant, k_{cat}/K_m, would decrease more for an Arg-containing substrate than for one with Lys.

Model building also predicts that the Ala 216 mutant would displace a water molecule at the bottom of the specificity pocket that in the wild type enzyme binds to the NH_3^+ group of the substrate Lys side chain (Figure 11.12). The extra CH_3 group of this mutant is not expected to disturb the binding of the Arg side chain. One would therefore expect that the K_m for Lys



substrates would increase and therefore k_{cat}/K_m would decrease more for Lys- than for Arg-containing substrates. For the double mutant where both Gly 216 and Gly 226 are changed to Ala, one would predict an increase in the K_m values for both Lys- and Arg-containing substrates.

The experimentally determined k_{cat} and K_m values for the wild-type enzyme and the mutants are shown in Table 11.1. The dramatic kinetic effects of these mutations are best illustrated with the Arg substrate. The three mutants have roughly similar K_m values 15–35 times higher than for the wild type, but the k_{cat} values decrease by factors of 10 to about 1000. The mutants were designed to change the specificity, but by far the largest changes occur in the catalytic rates. Apparently, these mutations affect the structure of the enzyme in additional ways, possibly by causing conformational changes outside the specificity pocket, so that the stabilization of the transition state is reduced and consequently the activation energies for the reactions are different.

The changes in the specificity constants, on the other hand, were as expected from the predictions. The ratio of the k_{cat}/K_m values for the Arg and Lys substrates (last column in Table 11.1) gives a measure of the relative specificities. This ratio decreases for the Ala 226 mutant and increases for the Ala 216 mutant as predicted. However, the changes in these values depend not

Figure 11.12 Schematic diagram of the specificity pocket of mutant trypsin with Ala (purple) at positions 216 and 226. (a) The position of a bound Lys side chain (red) in this pocket as observed in the structure of a complex between trypsin (green) and a peptide inhibitor. The NH_3^+ group of the Lys side chain interacts with the COO^- group of Asp 189 through a water molecule. (b) No structure is available for an Arg side chain in the substrate specificity pocket of trypsin. It is assumed that the complex of trypsin (green) with benzamide (red) is a good model for arginine binding in this pocket. One NH_2 group of benzamide interacts directly with the COO^- group of Asp 189 and the second NH_2 group interacts with a water molecule and the OH group of Ser 190. (Adapted from C.S. Craik et al., *Science* 228: 291–297, 1985.)

Table 11.1 Kinetic data for wild-type and mutant trypsins

Enzyme	Arg			Lys			(k_{cat}/K_m) Arg / (k_{cat}/K_m) Lys
	k_{cat}	K_m	k_{cat}/K_m	k_{cat}	K_m	k_{cat}/K_m	
Wild type	1	1	1	0.9	10	0.1	10
Gly 216, *Gly* 226→*Ala*	0.001	15	0.0001	0.0005	25	0.00002	25
Gly 226→*Ala*	0.01	35	0.0003	0.1	250	0.0005	0.5
Gly 216→*Ala*	0.7	30	0.02	0.2	280	0.001	20

The substrates used were D-Val-Leu-Arg-amino fluorocoumarin (Arg) and D-Val-Leu-Lys-amino fluorocoumarin (Lys). For clarity the K_m and k_{cat} values have been normalized to those of the wild-type enzyme for the Arg substrate.

only on changes in the K_m values, which reflect binding of substrate, but even more on changes in the k_{cat} values, which reflect catalytic rate. It can, therefore, be argued that the agreement with prediction is fortuitous.

The simple lesson to be learnt from these experiments is that critical amino acid residues can have pleiotropic roles in determining a protein's structure and therefore its function.

The Asp 189-Lys mutation in trypsin causes unexpected changes in substrate specificity

Asp 189 at the bottom of the substrate specificity pocket interacts with Lys and Arg side chains of the substrate, and this is the basis for the preferred cleavage sites of trypsin (see Figures 11.11 and 11.12). It is almost trivial to infer, from these observations, that a replacement of Asp 189 with Lys would produce a mutant that would prefer to cleave substrates adjacent to negatively charged residues, especially Asp. On a computer display, similar Asp–Lys interactions between enzyme and substrate can be modeled within the substrate specificity pocket but reversed compared with the wild-type enzyme.

The results of experiments in which the mutation was made were, however, a complete surprise. The Asp 189-Lys mutant was totally inactive with both Asp and Glu substrates. It was, as expected, also inactive toward Lys and Arg substrates. The mutant was, however, catalytically active with Phe and Tyr substrates, with the same low turnover number as wild-type trypsin. On the other hand, it showed a more than 5000-fold increase in k_{cat}/K_m with Leu substrates over wild type. The three-dimensional structure of this interesting mutant has not yet been determined, but the structure of a related mutant Asp 189-His shows the histidine side chain in an unexpected position, buried inside the protein.

As these experiments with engineered mutants of trypsin prove, we still have far too little knowledge of the functional effects of single point mutations to be able to make accurate and comprehensive predictions of the properties of a point-mutant enzyme, even in the case of such well-characterized enzymes as the serine proteinases. Predictions of the properties of mutations using computer modeling are not infallible. Once produced, the mutant enzymes often exhibit properties that are entirely surprising, but they may be correspondingly informative.

The structure of the serine proteinase subtilisin is of the α/β type

Subtilisins are a group of serine proteinases that are produced by different species of bacilli. These enzymes are of considerable commercial interest because they are added to the detergents in washing powder to facilitate removal of proteinaceous stains. Numerous attempts have therefore recently been made to change by protein engineering such properties of the subtilisin molecule as its thermal stability, pH optimum, and specificity. In fact, in 1988 subtilisin mutants were the subject of the first US patent granted for an engineered protein.

The subtilisin molecule is a single polypeptide chain of 275 amino acids with no similarities in the amino acid sequence to chymotrypsin. The three-dimensional structure of subtilisin BPN′ from *Bacillus amyloliquefaciens* was determined in 1969 by the group of Joseph Kraut in San Diego, California, and that of subtilisin Novo in 1971 by the group of Jan Drenth in Groningen, The Netherlands. The main feature of the subtilisin structure is a region of five parallel β strands (blue in Figure 11.13) surrounded by four helices, two on each side of the parallel β sheet. This α/β structure is thus quite different from the double antiparallel β-barrel structure of chymotrypsin (compare with Figure 11.7).

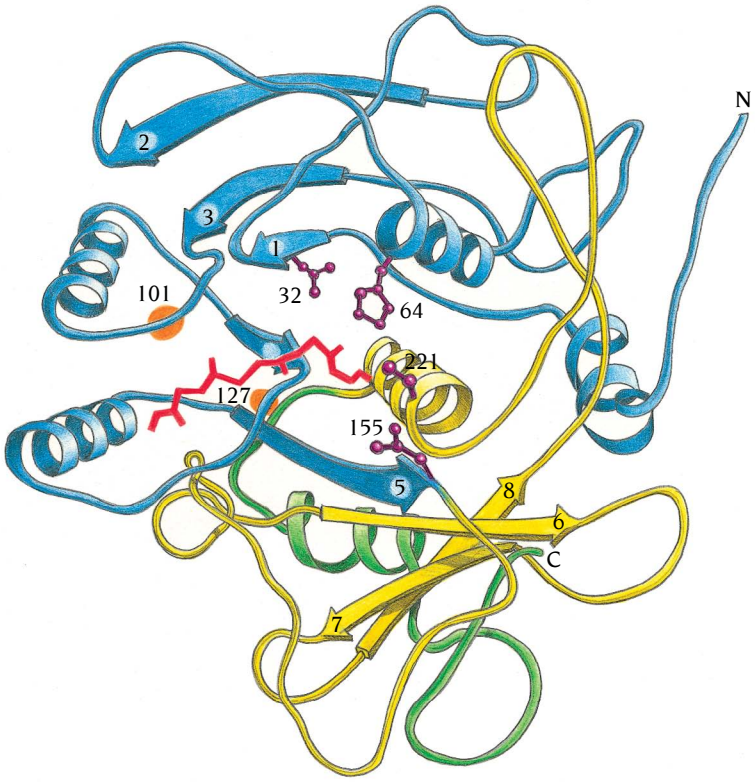


Figure 11.13 Schematic diagram of the three-dimensional structure of subtilisin viewed down the central parallel β sheet. The N-terminal region that contains the α/β structure is blue. It is followed by a yellow region, which ends with the fourth α helix of the α/β structure. The C-terminal part is green. The catalytic triad Asp 32, His 64, and Ser 221 as well as Asn 155, which forms part of the oxyanion hole are shown in purple. The main chain of part of a polypeptide inhibitor is shown in red. Main-chain residues around 101 and 127 (orange circles) form the nonspecific binding regions of peptide substrates.

The active sites of subtilisin and chymotrypsin are similar

The active site of subtilisin is outside the carboxy ends of the central β strands analogous to the position of the binding sites in other α/β proteins as discussed in Chapter 4. Details of this active site are surprisingly similar to those of chymotrypsin, in spite of the completely different folds of the two enzymes (Figures 11.14 and 11.9). A catalytic triad is present that comprises residues Asp 32, His 64 and the reactive Ser 221. The negatively charged oxygen atom of the tetrahedral transition state binds in an oxyanion hole,

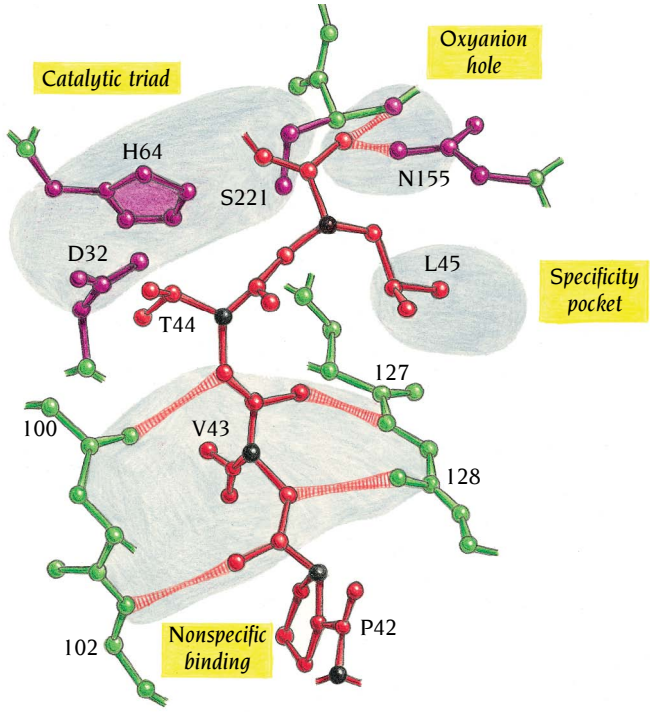


Figure 11.14 Schematic diagram of the active site of subtilisin. A region (residues 42–45) of a bound polypeptide inhibitor, eglin, is shown in red. The four essential features of the active site—the catalytic triad, the oxyanion hole, the specificity pocket, and the region for non-specific binding of substrate—are highlighted in yellow. Important hydrogen bonds between enzyme and inhibitor are striped. This figure should be compared to Figure 11.9, which shows the same features for chymotrypsin. (Adapted from W. Bode et al., *EMBO J.* 5: 813–818, 1986.)

forming hydrogen bonds with the side-chain amide group of Asn 155 and the main-chain nitrogen atom of Ser 221. Peptide substrates and inhibitors bind nonspecifically by forming a small antiparallel pleated sheet, which in subtilisin comprises three β strands (Figure 11.14). There is also a hydrophobic specificity pocket adjacent to the scissile bond.

All the four essential features of the active site of chymotrypsin are thus also present in subtilisin. Furthermore, these features are spatially arranged in the same way in the two enzymes, even though different framework structures bring different loop regions into position in the active site. This is a classical example of convergent evolution at the molecular level.

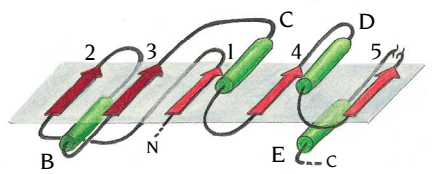


Figure 11.15 Topology diagram of the α/β region of subtilisin illustrating that β_2-α_B-β_3 has a different hand than the other β-α-β units.

A structural anomaly in subtilisin has functional consequences

There is one anomalous and puzzling feature of the subtilisin structure. We mentioned in Chapter 4 that virtually all β-α-β motifs were of the same hand, they were right-handed. Subtilisin contains the one exception to this general rule, which is illustrated in the topology diagram Figure 11.15. There are three β-α-β motifs in subtilisin, β_2-α_B-β_3, β_3-α_C-β_4, and β_4-α_D-β_5. If these motifs were of the same hand, the three α helices α_B, α_C, and α_D would be on the same side of the β sheet. However, α_B is beneath the sheet in the topology diagram in contrast to the other two helices because β_2-α_B-β_3 is left-handed. Why has this exception to the general rule of right-handed β-α-β motifs evolved?

The answer is quite clear. His 64, which is part of the catalytic triad, is in the first turn of helix α_B (Figure 11.13). This helix would be on the other side of the β sheet, far removed from the active site if the motif β_2-α_B-β_3 were right-handed. Therefore, to produce a proper catalytic triad of Asp 32, His 64, and Ser 221, helix α_B must be on the same side of the β sheet as Ser 221; consequently, the motif has evolved to be left-handed.

Transition-state stabilization in subtilisin is dissected by protein engineering

Two essential features are required to stabilize the covalent tetrahedral transition state in serine proteinases—the oxyanion hole, which provides hydrogen bonds to the negatively charged oxygen atom in the transition state, and the histidine residue of the catalytic triad, which provides a positive charge. The charge on this histidine is, in turn, stabilized by the aspartic acid side chain of the catalytic triad (Figure 11.6). The histidine residue also plays a second role in the catalytic mechanism by accepting a proton from the reactive serine residue and then donating that proton to the nitrogen atom of the leaving group. The effects on the catalytic rate of the different side chains involved in the catalytic triad and the oxyanion hole have been examined by P. Carter, J.A. Wells, and D. Estell at Genentech, USA, by analyses of mutants of subtilisin with one or several of these side chains have been changed.

Catalysis occurs without a catalytic triad

By changing Ser 221 in subtilisin to Ala the reaction rate (both k_{cat} and k_{cat}/K_m) is reduced by a factor of about 10^6 compared with the wild-type enzyme. The K_m value and, by inference, the initial binding of substrate are essentially unchanged. This mutation prevents formation of the covalent bond with the substrate and therefore abolishes the reaction mechanism outlined in Figure 11.5. When the Ser 221 to Ala mutant is further mutated by changes of His 64 to Ala or Asp 32 to Ala or both, as expected there is no effect on the catalytic reaction rate, since the reaction mechanism that involves the catalytic triad is no longer in operation. However, the enzyme still has an appreciable catalytic effect; peptide hydrolysis is still about 10^3–10^4 times the nonenzymatic rate. Whatever the reaction mechanism

used by these mutants, it is apparent that the remaining parts of the active site must bind more tightly to the substrate in its transition state than in its initial state, thereby giving a higher reaction rate than in the absence of enzyme.

Substrate molecules provide catalytic groups in substrate-assisted catalysis

The single mutation His 64–Ala decreases the reaction rate of subtilisin for most substrates by the same factor (approx. 10^6) as the mutation of Ser 221. This histidine (His 64), therefore, seems to be as important as Ser 221 for the formation of a covalent tetrahedral intermediate. However, model building suggested that it might be possible at least partly to compensate for the loss of this histidine in the catalytic triad of the mutant protein with a histidine side chain from a peptide substrate (Figure 11.16). Experiments confirmed this prediction and showed that the mutant His 64–Ala catalyzes hydrolysis of a peptide substrate about 400 times faster when the peptide has histidine at the appropriate position in its sequence. However, the rate is still several orders of magnitude below the rate of the wild-type enzyme, presumably because of the slightly different position and orientation of the histidine side chain. Nevertheless, the principle of **substrate-assisted catalysis** has been demonstrated: an essential group that is lacked by a mutant enzyme can be replaced by a similar group from the substrate. One consequence of substrate-assisted catalysis is that the mutant enzyme is highly specific for substrates containing the essential group. The His 64–Ala mutant of subtilisin, for example, has a specificity factor (ratios of k_{cat}/K_m) of about 200 for substrates containing histidine.

The single mutation Asp 32–Ala reduces the catalytic reaction rate by a factor of about 10^4 compared with wild type. This rate reduction reflects the role of Asp 32 in stabilizing the positive charge that His 64 acquires in the transition state. A similar reduction of k_{cat} and k_{cat}/K_m (2.5×10^3) is obtained for the single mutant Asn 155–Thr. Asn 155 provides one of the two hydrogen bonds to the substrate transition state in the oxyanion hole of subtilisin.

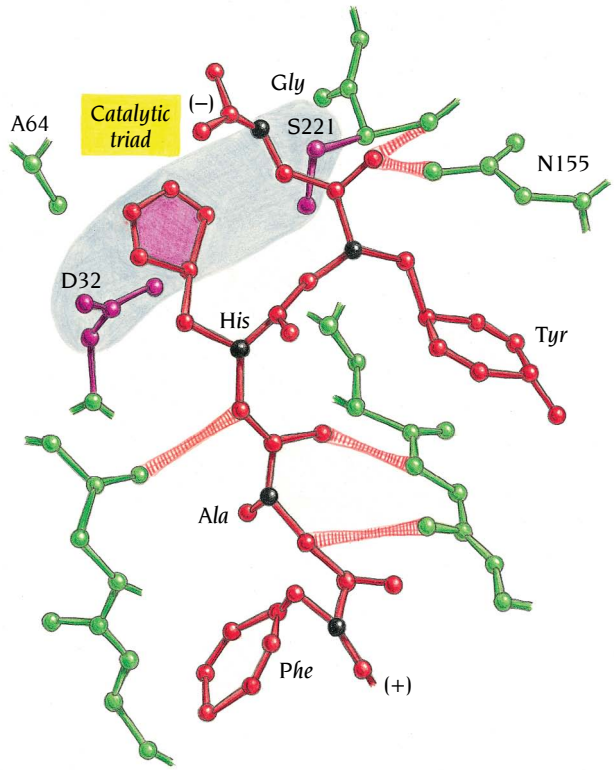


Figure 11.16 Substrate-assisted catalysis. Schematic diagram from model building of a substrate, NH_2-Phe-Ala-His-Tyr-Gly-COOH (red), bound to the subtilisin mutant His 64–Ala. The diagram illustrates that the His residue of the substrate can occupy roughly the same position in this mutant as His 64 in wild-type subtilisin (see Figure 11.14) and thereby partly restore the catalytic triad.

Model building shows that the OH group of Thr in the mutant is too far away to provide such a hydrogen bond. The loss of this feature of the stabilization of the transition state thus reduces the rate by more than a thousandfold.

The subtilisin mutants described here illustrate the power of protein engineering as a tool to allow us to identify the specific roles of side chains in the catalytic mechanisms of enzymes. In Chapter 17 we shall discuss the utility of protein engineering in other contexts, such as design of novel proteins and the elucidation of the energetics of ligand binding to proteins.

Conclusion

Enzymes increase the rate of chemical reactions by decreasing the activation energy of the reactions. This is achieved primarily by the enzyme preferentially binding to the transition state of the substrate. Catalytic groups of the enzyme are required to achieve a specific reaction path for the conversion of substrate to product.

Serine proteinases such as chymotrypsin and subtilisin catalyze the cleavage of peptide bonds. Four features essential for catalysis are present in the three-dimensional structures of all serine proteinases: a catalytic triad, an oxyanion binding site, a substrate specificity pocket, and a nonspecific binding site for polypeptide substrates. These four features, in a very similar arrangement, are present in both chymotrypsin and subtilisin even though they are achieved in the two enzymes in completely different ways by quite different three-dimensional structures. Chymotrypsin is built up from two β-barrel domains, whereas the subtilisin structure is of the α/β type. These two enzymes provide an example of convergent evolution where completely different loop regions, attached to different framework structures, form similar active sites.

The catalytic triad consists of the side chains of Asp, His, and Ser close to each other. The Ser residue is reactive and forms a covalent bond with the substrate, thereby providing a specific pathway for the reaction. His has a dual role: first, it accepts a proton from Ser to facilitate formation of the covalent bond; and, second, it stabilizes the negatively charged transition state. The proton is subsequently transferred to the N atom of the leaving group. Mutations of either of these two residues decrease the catalytic rate by a factor of 10^6 because they abolish the specific reaction pathway. Asp, by stabilizing the positive charge of His, contributes a rate enhancement of 10^4.

The oxyanion binding site stabilizes the transition state by forming two hydrogen bonds to a negatively charged oxygen atom of the substrate. Mutations that prevent formation of one of these bonds in subtilisin decrease the rate by a factor of about 10^3.

Mutations in the specificity pocket of trypsin, designed to change the substrate preference of the enzyme, also have drastic effects on the catalytic rate. These mutants demonstrate that the substrate specificity of an enzyme and its catalytic rate enhancement are tightly linked to each other because both are affected by the difference in binding strength between the transition state of the substrate and its normal state.

Selected readings

General

Blow, D.M. Structure and mechanism of chymotrypsin. *Acc. Chem. Res.* 9: 145–152, 1976.

Fersht, A. *Enzyme Structure and Mechanism*, 2nd ed. New York: W.H. Freeman, 1984.

Huber, R., Bode, W. Structural basis of the activation and action of trypsin. *Acc. Chem. Res.* 11: 114–122, 1978.

James, M.N.G. An x-ray crystallographic approach to enzyme structure and function. *Can. J. Biochem.* 58: 251–270, 1980.

Jencks, W.P. Binding energy, specificity, and enzymatic catalysis: the Circe effect. *Adv. Enzymol.* 43: 219–410, 1975.

Knowles, J.R. Tinkering with enzymes: what are we learning? *Science* 236: 1252–1258, 1987.

Kraut, J. How do enzymes work? *Science* 242: 533–540, 1988.

Kraut, J. Serine proteases: structure and mechanism of catalysis. *Annu. Rev. Biochem.* 46: 331–358, 1977.

Neurath, H. Evolution of proteolytic enzymes. *Science* 224: 350–357, 1984.

Pauling, L. Nature of forces between large molecules of biological interest. *Nature* 161: 707–709, 1948.

Steitz, T.A., Shulman, R.G. Crystallographic and NMR studies of the serine proteases. *Annu. Rev. Biochem. Biophys.* 11: 419–444, 1982.

Stroud, R.M. A family of protein-cutting proteins. *Sci. Am.* 231(1): 74–88, 1974.

Walsh, C. *Enzymatic Reaction Mechanisms*. New York: W.H. Freeman, 1979.

Warshel, A., et al. How do serine proteases really work? *Biochemistry* 28:3629–3637, 1989.

Wells, J.A., et al. On the evolution of specificity and catalysis in subtilisin. *Cold Spring Harbor Symp. Quant. Biol.* 52: 647–652, 1987.

Specific structures

Bode, W., et al. Refined 1.2 Å crystal structure of the complex formed between subtilisin Carlsberg and the inhibitor eglin c. Molecular structure of eglin and its detailed interaction with subtilisin. *EMBO J.* 5: 813–818, 1986.

Bode, W., et al. The refined 1.9 Å crystal structure of human α-thrombin: interaction with D-Phe-Pro-Arg chloromethylketone and significance of the Tyr-Pro-Pro-Trp insertion segment. *EMBO J.* 8: 3467–3475, 1989.

Bryan, P., et al. Site-directed mutagenesis and the role of the oxyanion hole in subtilisin. *Proc. Natl. Acad. Sci. USA* 83: 3743–3745, 1986.

Carter, P., Wells, J.A. Dissecting the catalytic triad of a serine protease. *Nature* 332: 564–568, 1988.

Carter, P., Wells, J.A. Engineering enzyme specificity by "substrate-assisted catalysis." *Science* 237: 394–399, 1987.

Craik, C.S., et al. Redesigning trypsin: alteration of substrate specificity. *Science* 228: 291–297, 1985.

Craik, C.S., et al. The catalytic role of the active site aspartic acid in serine proteases. *Science* 237: 909–913, 1987.

Cunningham, B.C., Wells, J.A. Improvement in the alkaline stability of subtilisin using an efficient random mutagenesis and screening procedure. *Prot. Eng.* 1: 319–325, 1987.

Drenth, J., et al. Subtilisin novo. The three-dimensional structure and its comparison with subtilisin BPN. *Eur. J. Biochem.* 26: 177–181, 1972.

Estell, D.A., Graycar, T.P., Wells, J.A. Engineering an enzyme by site-directed mutagenesis to be resistant to chemical oxidation. *J. Biol. Chem.* 260: 6518–6521, 1985.

Fehlhammer, H., Bode, W., Huber, R. Crystal structure of bovine trypsinogen at 1.8 Å resolution. II. Crystallographic refinement, refined crystal structure and comparison with bovine trypsin. *J. Mol. Biol.* 111: 415–438, 1977.

Fujinaga, M., et al. Crystal and molecular structures of the complex of α-chymotrypsin with its inhibitor turkey ovomucoid third domain at 1.8 Å resolution. *J. Mol. Biol.* 195: 397–418, 1987.

Graf, L., et al. Selective alteration of substrate specificity by replacement of aspartic acid 189 with lysine in the binding pocket of trypsin. *Biochemistry* 26: 2616–2623, 1987.

Grütter, M.G., et al. Crystal structure of the thrombin-hirudin complex: a novel mode of serine protease inhibition. *EMBO J.* 9: 2361–2365, 1990.

James, M.N.G., et al. Structures of product and inhibitor complexes of Streptomyces griseus protease A at 1.8 Å resolution. A model for serine protease catalysis. *J. Mol. Biol.* 144: 43–88, 1980.

Krieger, M., Kay, L.M., Stroud, R.M. Structure and specific binding of trypsin: comparison of inhibited derivatives and a model for substrate binding. *J. Mol. Biol.* 83: 209–230, 1974.

Matthews, B.W., Sigler, P.B., Henderson, R., Blow, D.M. Three-dimensional structure of tosyl-α-chymotrypsin. *Nature* 214: 652–656, 1967.

McLachlan, A.D. Gene duplications in the structural evolution of chymotrypsin. *J. Mol. Biol.* 128: 49–79, 1979.

Poulos, T.L., et al. Polypeptide halomethyl ketones bind to serine proteases as analogs of the tetrahedral intermediate. *J. Biol. Chem.* 251: 1097–1103, 1976.

Read, R.J., James, M.N.G. Refined crystal structure of *Streptomyces griseus* trypsin at 1.7 Å resolution. *J. Mol. Biol.* 200: 523–551, 1988.

Rühlman, A., et al. Structure of the complex formed by bovine trypsin and bovine pancreatic trypsin inhibitor. *J. Mol. Biol.* 77: 417–436, 1973.

Shotton, D.M., Watson, H.C. Three-dimensional structure of tosyl-elastase. *Nature* 225: 811–816, 1970.

Sigler, P.B., et al. Structure of crystalline α-chymotrypsin II. A preliminary report including a hypothesis for the activation mechanism. *J. Mol. Biol.* 35: 143–164, 1968.

Smith, S.O., et al. Crystal versus solution structures of enzymes: NMR spectroscopy of a crystalline serine protease. *Science* 244: 961–964, 1989.

Sprang, S., et al. The three-dimensional structure of Asn^{102} mutant of trypsin: role of Asp^{102} in serine protease catalysis. *Science* 237: 905–909, 1987.

Thomas, P.G., Russel, A.J., Fersht, A. Tailoring the pH dependence of enzyme catalysis using protein engineering. *Nature* 318: 375–376, 1985.

Tsukada, H., Blow, D.M. Structure of α-chymotrypsin refined at 1.68 Å resolution. *J. Mol. Biol.* 184: 703–711, 1985.

Wang, D., Bode, W., Huber, R. Bovine chymotrypsinogen A. X-ray crystal structure analysis and refinement of a new crystal form at 1.8 Å resolution. *J. Mol. Biol.* 185: 595–624, 1985.

Wells, J.A., et al. Designing substrate specificity by protein engineering of electrostatic interactions. *Proc. Natl. Acad. Sci. USA* 84: 1219–1223, 1987.

Wells, J.A., et al. Recruitment of substrate-specificity properties from one enzyme into a related one by protein engineering. *Proc. Natl. Acad. Sci. USA* 84: 5167–5171, 1987.

Wright, C.S., Alden, R.A., Kraut, J. Structure of subtilisin BPN′ at 2.5 Å resolution. *Nature* 221: 235–242, 1969.

Membrane Proteins

12

Cells and organelles within them are bounded by membranes, which are extremely thin (4.5 nm) films of lipids and protein molecules. The lipids form a bilayered sheet structure that is hydrophilic on its two outer surfaces and hydrophobic in between. Protein molecules are embedded in this layer, and in the simplest case they are arranged with three distinct regions: one hydrophobic transmembrane segment and two hydrophilic regions, one on each side of the membrane. Those proteins whose polypeptide chain traverses the membrane only once usually form functional globular domains on at least one side of the membrane (Figure 12.1a). Often these can be cleaved off by proteolytic enzymes. The hemagglutinin and neuraminidase of influenza virus (discussed in Chapter 5), G-proteins and receptors (discussed in Chapter 13), and HLA proteins (discussed in Chapter 15) are examples of such cleavage products that can be handled as functional soluble globular domains. The polypeptide chain of other transmembrane proteins passes through the membrane several times, usually as α helices but in some cases as β strands (Figure 12.1b,c). In these cases the hydrophilic regions on either side of the membrane are the termini of the chain and the loops between the membrane-spanning parts. Proteolytic cleavage of these hydrophilic regions produces a number of fragments, and function is not preserved. Some proteins do not traverse the membrane but are instead attached to one side either through α helices that lie parallel to the membrane surface (Figure 12.1d) or by fatty acids, covalently linked to the protein, that intercalate in the lipid bilayer of the membrane.

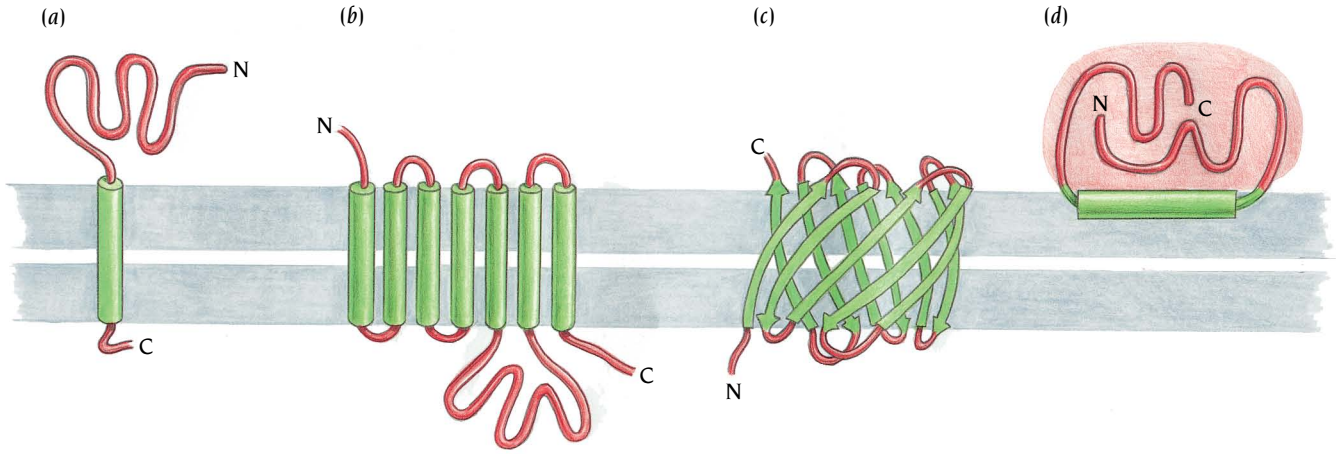


Figure 12.1 Four different ways in which protein molecules may be bound to a membrane. Membrane-bound regions are green and regions outside the membrane are red. Alpha-helices are drawn as cylinders and β strands as arrows. From left to right are (a) a protein whose polypeptide chain traverses the membrane once as an α helix, (b) a protein that forms several transmembrane α helices connected by hydrophilic loop regions, (c) a protein with several β strands that form a channel through the membrane, and (d) a protein that is anchored to the membrane by one α helix parallel to the plane of the membrane.

A biological membrane functions basically as a permeability barrier that establishes discrete compartments and prevents the random mixing of the contents of one compartment with those of another. However, biological membranes are more than passive containers. The embedded proteins serve as highly active mediators between the cell and its environment or the interior of an organelle and the cytosol. They catalyze specific transport of metabolites and ions across the membrane barriers. They convert the energy of sunlight into chemical and electrical energy, and they couple the flow of electrons to the synthesis of ATP. Furthermore, they act as signal receptors and transduce signals across the membrane. The signals can be, for example, neurotransmitters, growth factors, hormones, light or chemotactic stimuli. The transmembrane proteins of the plasma membrane are also involved in cell–cell recognition.

In this chapter we describe some examples of structures of membrane-bound proteins known to high resolution, and outline how the elucidation of these structures has contributed to understanding the specific function of these proteins, as well as some general principles for the construction of membrane-bound proteins. In Chapter 13 we describe some examples of the domain organization of receptor families and their associated proteins involved in signal transduction through the membrane.

Membrane proteins are difficult to crystallize

Membrane proteins, which have both hydrophobic and hydrophilic regions on their surfaces, are not soluble in aqueous buffer solutions and denature in organic solvents. However, if detergents, such as octylglucoside, are added to an aqueous solution, these proteins can be solubilized and purified in their native conformation. The hydrophobic part of the detergent molecules binds to the protein's hydrophobic surfaces, while the detergents' polar headgroups face the solution (Figure 12.2a). In this way the protein–detergent complex acquires an essentially hydrophilic surface with the hydrophobic parts buried inside the complex.

Such solubilized protein–detergent complexes are the starting material for purification and crystallization. For some proteins, the addition of small amphipathic molecules to the detergent-solubilized protein promotes crystallization, probably by facilitating proper packing interactions between the molecules in all three dimensions in a crystal (Figure 12.2b). Therefore, many different amphipathic molecules are added in separate crystallization experiments until, by trial and error, the correct one is found.

Despite considerable efforts very few membrane proteins have yielded crystals that diffract x-rays to high resolution. In fact, only about a dozen such proteins are currently known, among which are porins (which are outer membrane proteins from bacteria), the enzymes cytochrome c oxidase and prostaglandin synthase, and the light-harvesting complexes and photosynthetic reaction centers involved in photosynthesis. In contrast, many other membrane proteins have yielded small crystals that diffract poorly, or not at all, using conventional x-ray sources. However, using the most advanced synchrotron sources (see Chapter 18) it is now possible to determine x-ray structures from protein crystals as small as 20 μm wide which will permit more membrane protein structures to be elucidated.

Novel crystallization methods are being developed

These difficulties have prompted a search for novel techniques for crystallization of membrane proteins. Two approaches have given promising results; one using antibodies to solubilize the proteins and the second using continuous lipidic phases as crystallization media. Complexes with specific antibodies have larger polar surfaces than the membrane protein itself and are therefore likely to form crystals more easily in an aqueous enviroment. A recent example of an antibody–membrane protein complex utilized an F_v

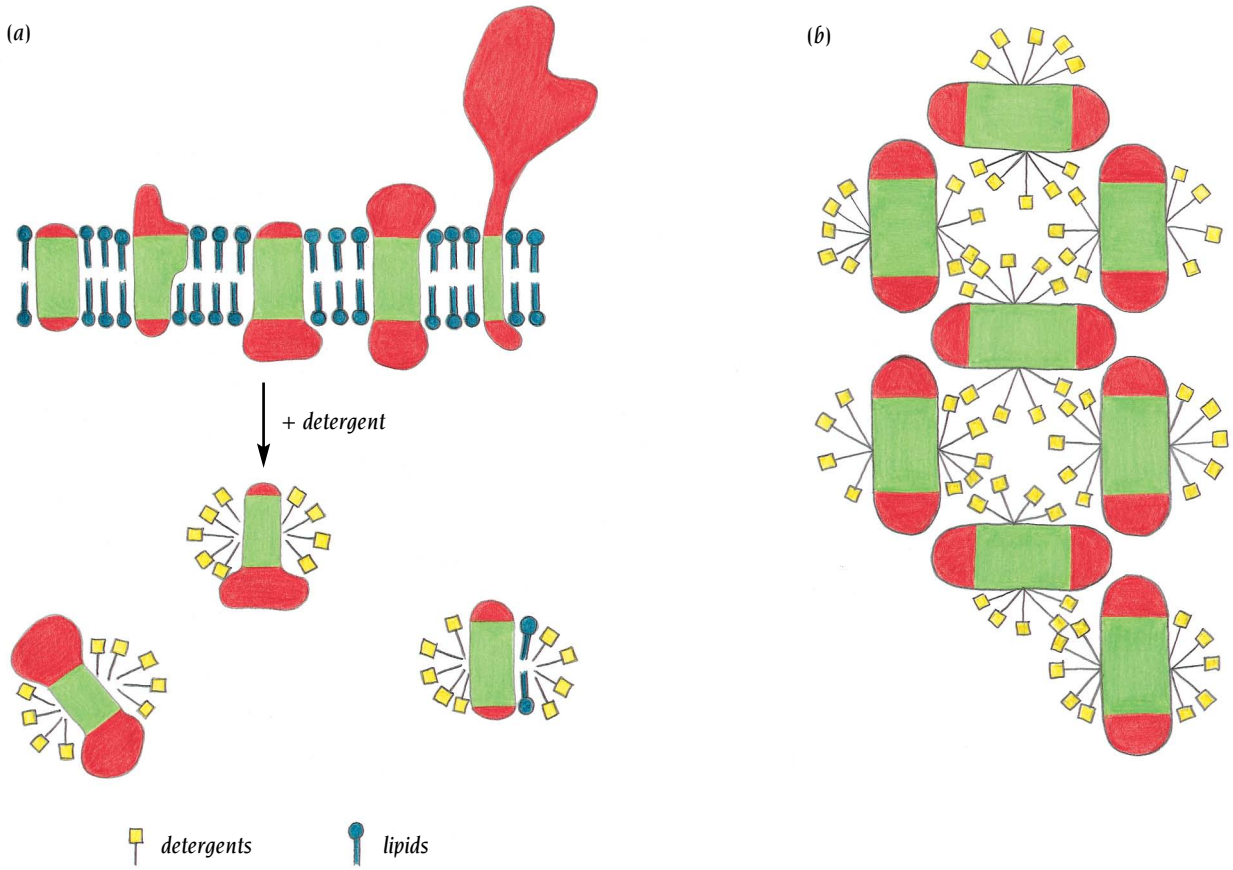


Figure 12.2 (a) Schematic drawing of membrane proteins in a typical membrane and their solubilization by detergents. The hydrophilic surfaces of the membrane proteins are indicated by red. (b) A membrane protein crystallized with detergents bound to its hydrophobic protein surface. The hydrophilic surfaces of the proteins and the symbols for detergents are as in (a). (Adapted from H. Michel, *Trends Biochem. Sci.* 8: 56–59, 1983.)

fragment (see Chapter 15) to crystallize a bacterial cytochrome c oxidase. In these crystals the major packing interactions are formed by the polar surfaces of the complex.

A continuous lipidic cubic phase is obtained by mixing a long-chain lipid such as monoolein with a small amount of water. The result is a highly viscous state where the lipids are packed in curved continuous bilayers extending in three dimensions and which are interpenetrated by communicating aqueous channels. Crystallization of incorporated proteins starts inside the lipid phase and growth is achieved by lateral diffusion of the protein molecules to the nucleation sites. This system has recently been used to obtain three-dimensional crystals 20 x 20 x 8 μm in size of the membrane protein bacteriorhodopsin, which diffracted to 2 Å resolution using a microfocus beam at the European Synchrotron Radiation Facility.

Two-dimensional crystals of membrane proteins can be studied by electron microscopy

The first really useful information about the structure of membrane proteins came not from x-ray crystallography but from high-resolution electron microscopy of **two-dimensional crystals**. Two-dimensional crystals can be thought of as crystalline membranes in which the membrane protein is arranged on a two-dimensional lattice. Naturally abundant membrane proteins sometimes form two-dimensional crystals *in vivo* or in isolated native membranes, in particular when some components such as lipids or other proteins are selectively extracted. A different way of making two-dimensional crystals is by reconstitution of detergent-solubilized membrane proteins into bilayers, which provide a natural, membrane-like environment for the protein. When the detergent is removed from a lipid–protein detergent mixture by dialysis or absorption, the hydrophobic effect causes the hydrophobic fatty acid tails of the lipids to associate with each other, and

with the hydrophobic surface of the membrane protein. In this way, the protein is incorporated into lipid sheets or vesicles. In favorable conditions it can then form a crystalline lattice.

Given the difficulty of obtaining three-dimensional crystals of membrane proteins, it is not surprising that the electron microscope technique is now widely used to study large membrane-bound complexes such as the acetylcholine receptor, rhodopsin, ion pumps, gap junctions, water channels and light-harvesting complexes, which crystallize in two dimensions.

Bacteriorhodopsin contains seven transmembrane α helices

The purple membrane of *Halobacterium halobium* contains ordered sheets of **bacteriorhodopsin**, a protein of 248 amino acid residues which binds retinal, the same photosensitive pigment that is used to capture light in our eyes. Bacteriorhodopsin uses the energy of light to pump protons across the membrane. Richard Henderson and Nigel Unwin at LMB, Cambridge, UK, pioneered high-resolution three-dimensional reconstruction of tilted low-dose electron microscopy images using such two-dimensional crystals. The 7-Å model of bacteriorhodopsin (Figure 12.3a) that they obtained in this way in 1975 provided the first glimpse of how membrane proteins are constructed. The fundamental observation that this protein has a number of **transmembrane α helices** (Figure 12.3b) has had a great impact on subsequent theories and experiments on membrane proteins; it also provided the first

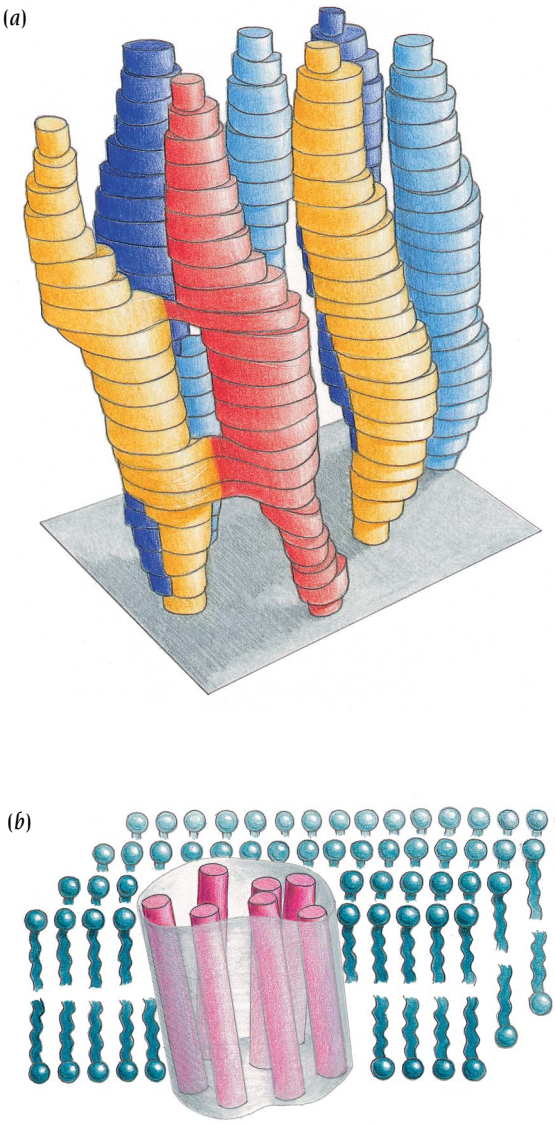


Figure 12.3 Two-dimensional crystals of the protein bacteriorhodopsin were used to pioneer three-dimensional high-resolution structure determination from electron micrographs. An electron density map to 7 Å resolution (a) was obtained and interpreted in terms of seven transmembrane helices (b). In 1990 the resolution was extended to 3 Å, which confirmed the presence of the seven α helices (c). This structure also showed how these helices were connected by loop regions and where the retinal molecule was bound to bacteriorhodopsin. (Courtesy of R. Henderson.)

experimental evidence behind the now extensively used methods to predict transmembrane helices from amino acid sequences.

This electron microscopy reconstruction has since been extended to high resolution (3 Å) where the connections between the helices and the bound retinal molecule are visible together with the seven helices (Figure 12.3c). The helices are tilted by about 20° with respect to the plane of the membrane. This is the first example of a high-resolution three-dimensional protein structure determination using electron microscopy. The structure has subsequently been confirmed by x-ray crystallographic studies to 2 Å resolution.

Bacteriorhodopsin is a light-driven proton pump

Halobacteria have the simplest biological system for the conversion of light to chemical energy. Under conditions of low oxygen tension and intense illumination the cells synthesize bacteriorhodopsin. When the bound retinal absorbs a photon it undergoes an isomerization from *trans* to *cis* (Figure 12.4) and, as a consequence, protons are pumped from the cytosol to the extracellular space, creating a proton gradient. This gradient is used to generate ATP and to transport ions and molecules across the membrane. The mechanism by which bacteriorhodopsin acts as a **proton pump** has been studied by many biophysical methods over several decades and the results, in conjunction with Henderson's structural studies, have given the following simplified scheme for proton pumping.

Retinal is bound in a pocket of bacteriorhodopsin about equidistant from the two sides of the membrane (Figure 12.5). The pigment forms a Schiff base with a lysine residue, Lys 216; in other words, it is covalently linked to the nitrogen atom of the lysine side chain that is protonated and therefore has a positive charge (see Figure 12.5). This positive charge is positioned in a channel of the protein that extends across the membrane and through which protons are pumped from the cytosolic to the extracellular side. The channel is narrow on the cytosolic side and lined with hydrophobic residues with the exception of Asp 96, which has been shown by studies of mutant proteins to be essential for proton pumping. In contrast, the channel is wide and hydrophilic on the extracellular side and includes a second essential acidic residue, Asp 85.

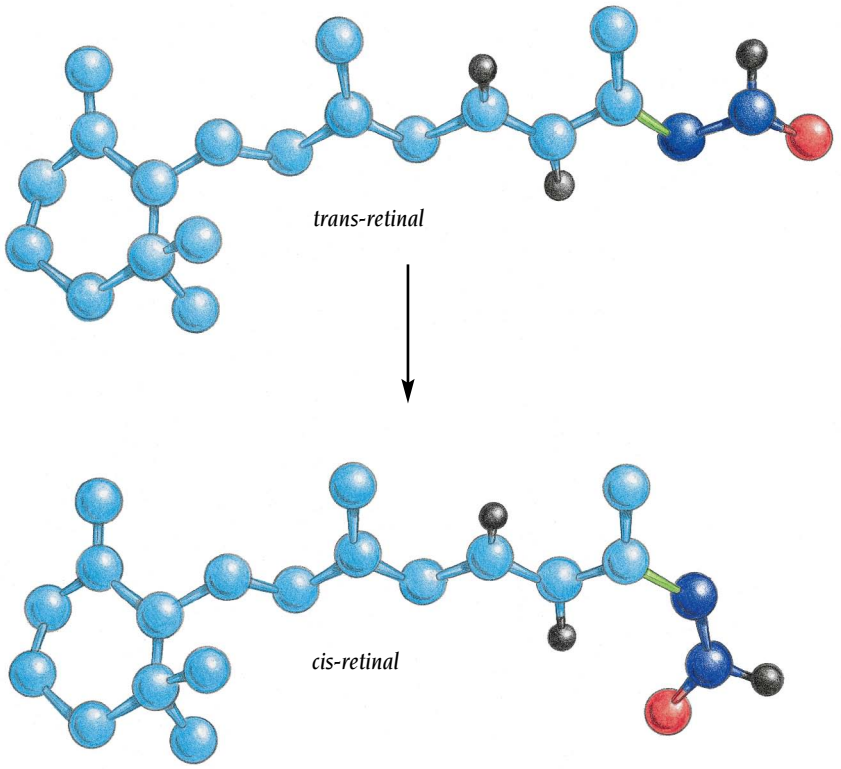


Figure 12.4 The light-absorbing pigment retinal undergoes a conformational change called isomerization, when it absorbs light. One part of the molecule (dark blue and red) rotates 180° around a double bond between two carbon atoms (green). The geometry of the molecule is changed by this rotation from a *trans* form to a *cis* form. Carbon atoms are blue, hydrogen atoms gray and the oxygen atom red.

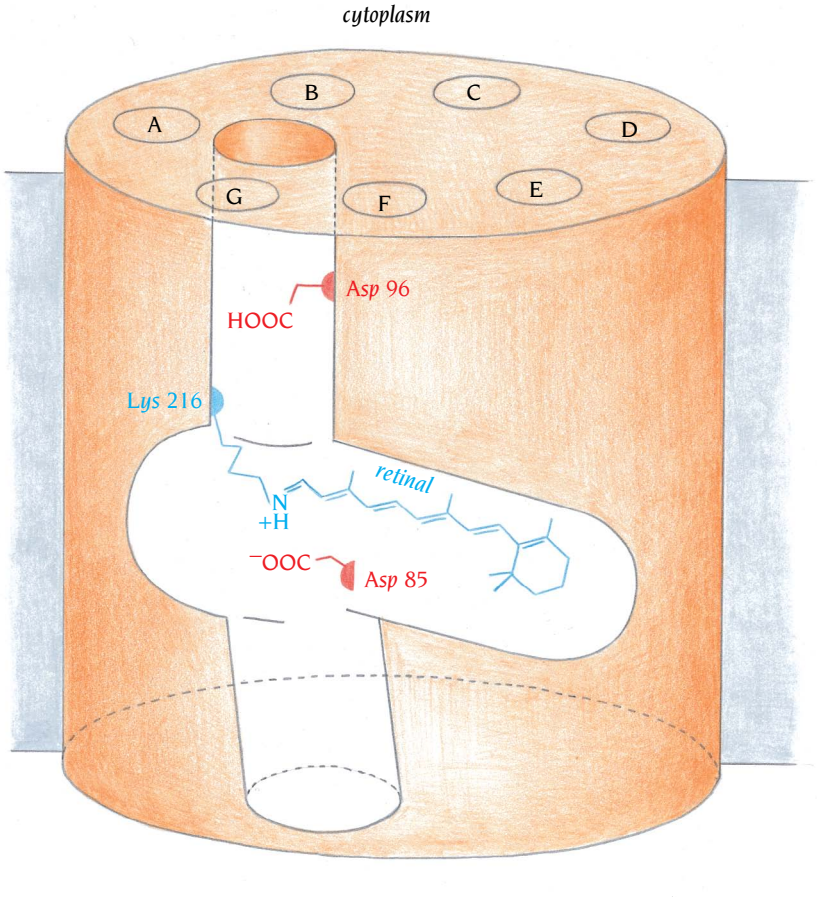


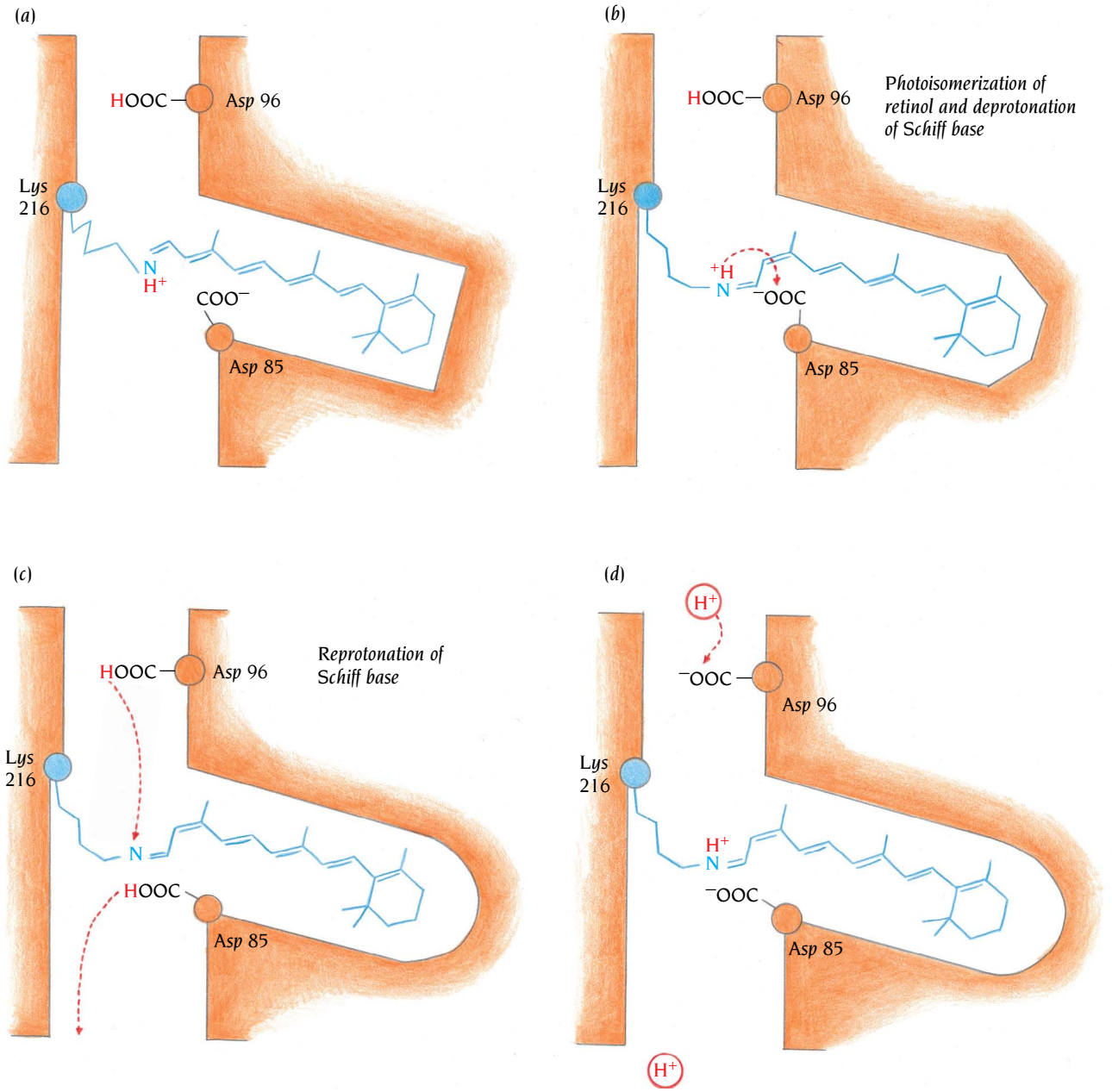
Figure 12.5 Schematic diagram of the bacteriorhodopsin molecule illustrating the relation between the proton channel and bound retinal in its *trans* form. A to E are the seven transmembrane helices. Retinal is covalently bound to a lysine residue. The relative positions of two Asp residues, which are important for proton transfer, are also shown. (Adapted from R. Henderson et al., *J. Mol. Biol.* 213: 899–929, 1990.)

In the unisomerized *trans* state of retinal, Asp 85 is close to the positive charge of the Schiff base (Figure 12.6a). The structural change of the retinal molecule due to the *trans* to *cis* photoisomerization causes the Schiff base to change its position relative to Asp 85, which induces transfer of the Schiff base proton to the aspartate group (Figure 12.6b). Once the Schiff base–Asp 85 ion pair is converted to a neutral pair by this proton transfer the protein undergoes a conformational change from the T state to the relaxed R state (see Chapter 6). X-ray diffraction and electron microscopy studies have shown that this conformational change involves a reorganization of some of the transmembrane helices that bind retinal, with the consequence that the Schiff base is moved from the extracellular part of the channel to the cytoplasmic part, away from Asp 85 towards Asp 96. Asp 85 then delivers a proton through the hydrophilic part of the channel to the extracellular space and Asp 96 reprotonates the Schiff base, which subsequently reverts to the *trans* state and the protein changes its conformation back to the T-state ready for another cycle of photoisomerization-induced proton transfer (Figure 12.6c,d). In short, light causes a chemical change at the active site that alters the conformation of the protein, which in turn drives protons from the cytosolic side of the membrane to the extracellular side.

Porins form transmembrane channels by β strands

Gram-negative bacteria are surrounded by two membranes, an inner plasma membrane and an outer membrane. These are separated by a periplasmic space. Most plasma membrane proteins contain long, continuous sequences of about 20 hydrophobic residues that are typical of transmembrane α helices such as those found in bacteriorhodopsin. In contrast, most outer membrane proteins do not show such sequence patterns.

This enigma was resolved in 1990 when the x-ray structure of an outer membrane protein, **porin**, showed that the transmembrane regions were β



strands and not α helices. Sequence comparisons have since shown that most if not all bacterial outer membrane porins have transmembrane β strands. Porins are among the most abundant proteins in bacteria. Each *Escherichia coli* cell contains about 100,000 copies of porin molecules in its outer membrane. Each porin forms an open water-filled channel that allows passive diffusion of nutrients and waste products across the outer membrane. These channels are restricted in size, and this excludes larger, potentially toxic compounds from entering the cell.

Porin channels are made by up and down β barrels

The first x-ray structure of a porin was determined by the group of Georg Schulz and Wolfram Welte at Freiburg University, Germany, who succeeded in growing crystals of a porin from *Rhodobacter capsulatus* that diffracted to 1.8 Å resolution. Since then the x-ray structures of several other porin molecules have been determined and found to be very similar to the *R. capsulatus* porin despite having no significant sequence identity.

Each subunit of the trimeric porin molecule from *R. capsulatus* folds into a 16-stranded up and down antiparallel β barrel in which all β strands form

Figure 12.6 Schematic diagram illustrating the proton movements in the photocycle of bacteriorhodopsin. The protein adopts two main conformational states, tense (T) and relaxed (R). The T state binds *trans*-retinal tightly and the R state binds *cis*-retinal. (a) Structure of bacteriorhodopsin in the T state with *trans*-retinal bound to Lys 216 via a Schiff base. (b) A proton is transferred from the Schiff base to Asp 85 following isomerization of retinal and a conformational change of the protein. (c) Structure of bacteriorphodopsin in the R state with *cis*-retinal bound. A proton is transferred from Asp 96 to the Schiff base and from Asp 85 to the extracellular space. (d) A proton is transferred from the cytoplasm to Asp 96. (Adapted from R. Henderson et al., *J. Mol. Biol.* 213: 899–929, 1990.)

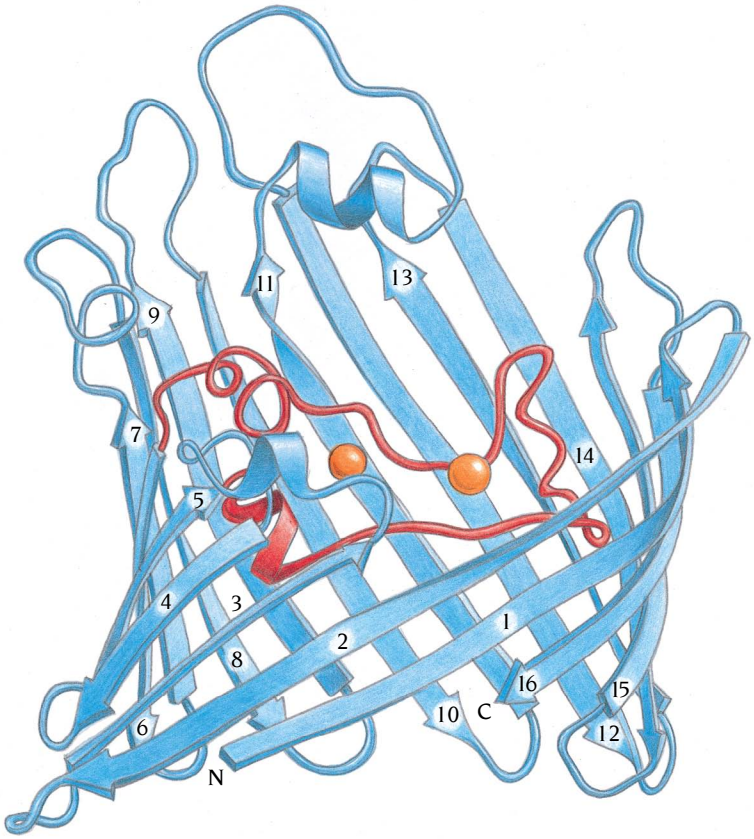


Figure 12.7 Ribbon diagram of one subunit of porin from *Rhodobacter capsulatus* viewed from within the plane of the membrane. Sixteen β strands form an antiparallel β barrel that traverses the membrane. The loops at the top of the picture are extracellular whereas the short turns at the bottom face the periplasm. The long loop between β strands 5 and 6 (red) constricts the channel of the barrel. Two calcium atoms are shown as orange circles. (Adapted from M.S. Weiss and G.E. Schulz, *J. Mol. Biol.* 227: 493–509, 1992.)

hydrogen bonds to their neighbors (Figure 12.7). All other known porin barrels also contain 16 β strands except a maltoporin from *E. coli*, which contains 18 β strands. In contrast to the β barrels we have previously discussed, which have a tightly packed hydrophobic core, the porin barrels contain a central channel because of the large number of β strands. The channel is, however, partially blocked by a long loop region between β strands 5 and 6 that projects into the channel. This arrangement creates a narrow region in the middle of the channel, called the eyelet, about 9 Å long and 8 Å in diameter, which defines the size of solute molecules that can traverse the channel (see Figure 12.8).

The eyelet is lined almost exclusively with positively and negatively charged groups that are arranged on opposite sides of the channel, causing a transversal electric field across the pore. One His, two Lys and three Arg residues form the positive side and four Glu and seven Asp residues are on the negative side. The large local surplus of negative charges is partially compensated by two bound calcium atoms. This asymmetric arrangement of charges no doubt contributes to the selection of molecules that can pass through the channel.

Since the outside of the barrel faces hydrophobic lipids of the membrane and the inside forms the solvent-exposed channel, one would expect the β strands to contain alternating hydrophobic and hydrophilic side chains. This requirement is not strict, however, because internal residues can be hydrophobic if they are in contact with hydrophobic residues from loop regions. The prediction of transmembrane β strands from amino acid sequences is therefore more difficult and less reliable than the prediction of transmembrane α helices.

Each porin molecule has three channels

The complete porin molecule is a stable trimer of three identical subunits, three each with a functional channel (Figure 12.8). About one-third of the

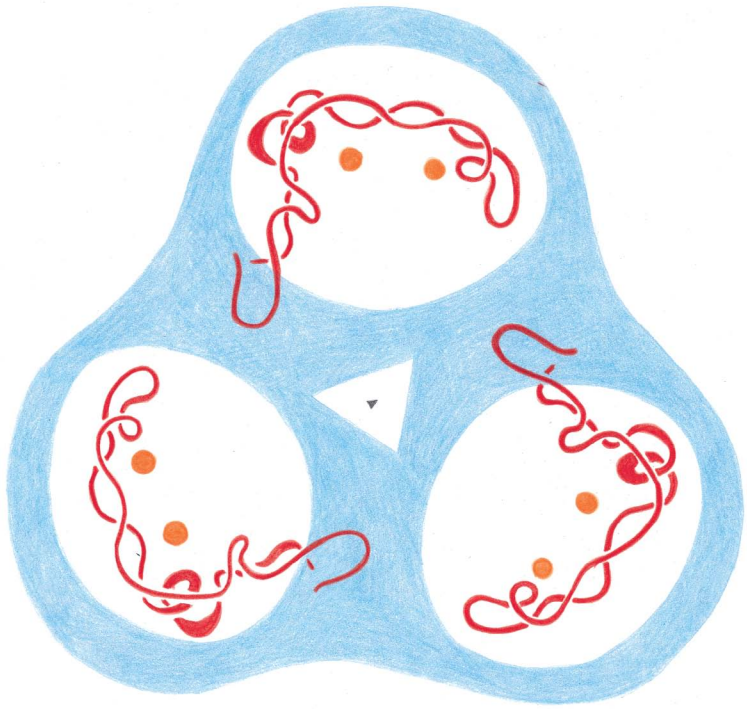


Figure 12.8 Schematic diagram of the trimeric porin molecule viewed from the extracellular space. Blue regions illustrate the walls of the three porin barrels, the loop regions that constrict the channel are red and the calcium atoms are orange.

barrel's outer surface is involved in subunit interactions with the other two subunits, comprising polar interactions from loop regions and hydrophobic interactions from side chains of the β strands. The bottom part of the trimer facing the periplasmic space has a flat and smooth surface made up of the short loop regions at this end of the three β barrels (as shown in Figure 12.7). In contrast the upper part has long loop regions and is funnel-shaped, with the channels of the three barrels providing three outlets from the single funnel. The inner sides of the funnel are lined with hydrophilic residues that are in contact with solvent from the extracellular space.

The journey of an extracellular solute molecule through the channel may now be depicted in the following way. Large molecules are prevented from entering by the size of the funnel of the trimer. Further screening occurs at the entrance of the channel in the individual barrels. Finally, a molecule small enough to enter the central channel then encounters the eyelet, where the charged side chains determine the size limitations and the ion selectivity of the pore. After this narrow passage the molecule is effectively released into the periplasmic space. It should, however, be borne in mind that this picture describes only one state of the channel. Triggers such as an electric potential or a change in osmotic pressure can modify the structure of the channel and therefore its permeability, but as yet we have no structural information on such changes.

The outer surface of the trimeric porin molecule shows a pronounced partitioning with respect to hydrophobicity. Polar side chains of the loop regions are abundant at the top of the trimer, followed by a hydrophobic band with a width of 25 Å that encircles the molecule. This band presumably forms the region that is embedded in the core of the outer membrane, which has a thickness of about 25 Å. The top and the bottom of this hydrophobic band are largely composed of aromatic residues, Phe and Tyr, whereas the central region is composed of small to medium-sized aliphatic residues. This suggests that aromatic rings are energetically favored at the interface between the inner lipid part of the membrane and the hydrophilic regions facing the extracellular and periplasmic spaces. A similar distribution of aromatic and aliphatic residues is also present in other membrane proteins such as the photosynthetic reaction center and bacteriorhodopsin.

Ion channels combine ion selectivity with high levels of ion conductance

Outer membrane channel-forming proteins such as porins, which have relatively large and permissive pores, would have disastrous effects if they directly connected the inside of a cell to the extracellular space. Therefore, most channel proteins in the plasma membranes of plant and animal cells have narrow and highly selective pores that are concerned specifically with the transport of inorganic ions, and so are referred to as **ion channels**. The function of such channels is to allow specific inorganic ions, mainly K^+, Na^+, Ca^{2+} and Cl^- to diffuse rapidly across the lipid bilayer and therefore balance differences in electric charge between the two sides of the membrane, the membrane potential. The membrane potential of resting cells is determined largely by K^+, which is actively pumped into the cell by an ATP-driven Na^+, K^+ pump, but which can also move freely in or out through **K^+ leak channels** in the plasma membrane.

The K^+ leak channels are highly selective for K^+ ions by a factor of 10,000 over Na^+. Two aspects of ion conduction by these K^+ channels have intrigued biophysicists. First, what is the chemical basis for this selectivity since both K^+ and Na^+ are featureless spheres? Steric occlusion cannot account for the selectivity since Na^+ is smaller than K^+ (0.95 Å and 1.35 Å radii respectively). Second, how can K^+ channels be so selective and at the same time exhibit a throughput of 10^8 ions per second, which approaches the diffusion-limited rate? The selectivity implies strong interactions between K^+ and the pore and intuitively one would assume that the off velocity of K^+ binding would be low, and consequently the rate of release would be low. A recent x-ray structure determination of a K^+ channel from the bacterium *Streptomyces lividans* by the group of Roderick MacKinnon at Rockefeller University, New York, has revealed the structural basis for combining ion selectivity with a high rate of ion conduction in these and other ion channels.

The K^+ channel is a tetrameric molecule with one ion pore in the interface between the four subunits

The polypeptide chain of the bacterial K^+ channel comprises 158 residues folded into two transmembrane helices, a pore helix and a cytoplasmic tail of 33 residues that was removed before crystallization. Four subunits

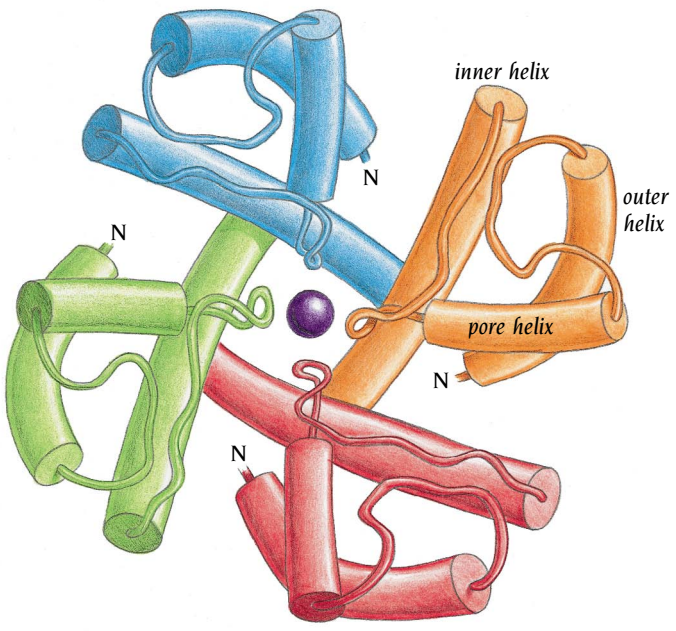


Figure 12.9 Schematic diagram of the structure of a potassium channel viewed perpendicular to the plane of the membrane. The molecule is tetrameric with a hole in the middle that forms the ion pore (purple). Each subunit forms two transmembrane helices, the inner and the outer helix. The pore helix and loop regions build up the ion pore in combination with the inner helix. (Adapted from S.A. Doyle et al., *Science* 280: 69–77, 1998.)

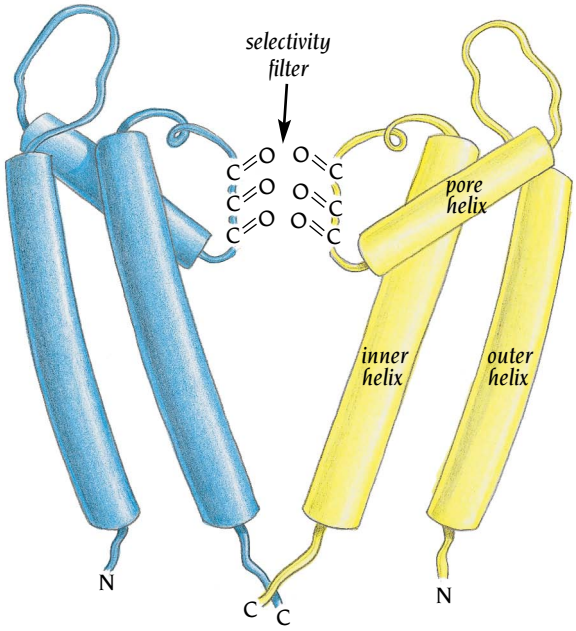


Figure 12.10 Diagram showing two subunits of the K^+ channel, illustrating the way the selectivity filter is formed. Main-chain atoms line the walls of this narrow passage with carbonyl oxygen atoms pointing into the pore, forming binding sites for K^+ ions. (Adapted from D.A. Doyle et al., *Science* 280: 69–77, 1998.)

arranged around a central fourfold symmetry axis form the K^+ channel molecule (Figure 12.9). The subunits pack together in such a way that there is a hole in the center which forms the ion pore through the membrane.

The C-terminal transmembrane helix, the inner helix, faces the central pore while the N-terminal helix, the outer helix, faces the lipid membrane. The four inner helices of the molecule are tilted and kinked so that the subunits open like petals of a flower towards the outside of the cell (Figure 12.10). The open petals house the region of the polypeptide chain between the two transmembrane helices. This segment of about 30 residues contains an additional helix, the pore helix, and loop regions which form the outer part of the ion channel. One of these loop regions with its counterparts from the three other subunits forms the narrow selectivity filter that is responsible for ion selectivity. The central and inner parts of the ion channel are lined by residues from the four inner helices.

The ion pore has a narrow ion selectivity filter

The overall length of the ion pore is 45 Å and its diameter varies along its length (Figure 12.11). As expected for a K^+ channel, there is a surplus of

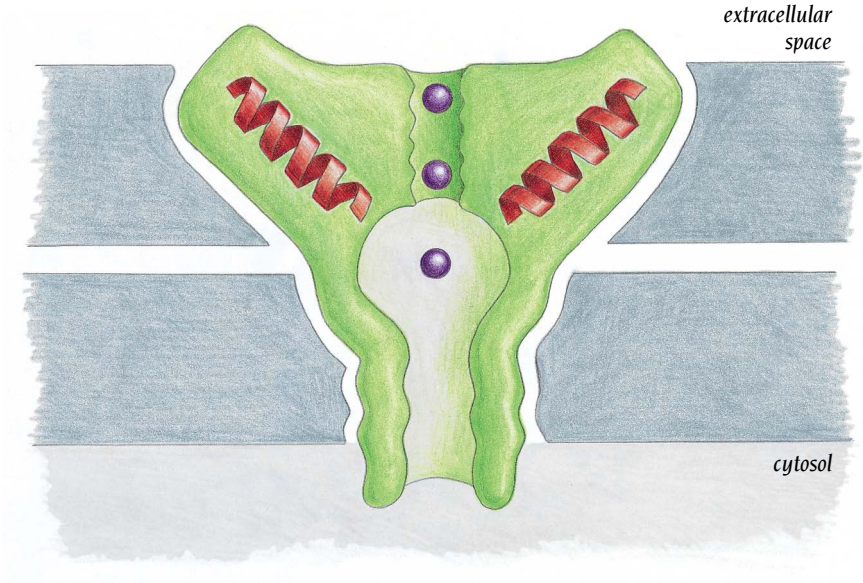


Figure 12.11 Schematic diagram of the ion pore of the K^+ channel. From the cytosolic side the pore begins as a water-filled channel that opens up into a water-filled cavity near the middle of the membrane. A narrow passage, the selectivity filter, links this cavity to the external solution. Three potassium ions (purple spheres) bind in the pore. The pore helices (red) are oriented such that their carboxyl end (with a negative dipole moment) is oriented towards the center of the cavity to provide a compensating dipole charge to the K^+ ions. (Adapted from D.A. Doyle et al., *Science* 280: 69–77, 1998.)

negative charges at both ends of the pore, which attract positively charged ions. From the cytosolic side, the pore begins as a channel 18 Å long, which opens into a wider cavity of about 10 Å diameter near the middle of the membrane. A narrow passage, the selectivity filter, links this cavity to the external solution. Main-chain atoms from all four subunits line the walls of this passage with carbonyl oxygen atoms pointing into the channel (see Figure 12.10). Three metal-binding sites have been identified in the ion pore (green in Figure 12.11), two within the selectivity filter and one in the cavity.

The structure of the selectivity filter has two essential features. First, the main-chain atoms create a stack of sequential oxygen rings along the passage, providing several closely spaced binding sites of the required dimensions for coordinating naked, dehydrated K^+ ions. The K^+ thus have only a small distance to diffuse from one site to the next within the selectivity filter. Second, the side chains of the residues that provide these binding sites point away from the channel and pack against the side chains from the pore helices. This packing firmly fixes the positions of the main-chain atoms, including the oxygen atoms that bind K^+. Since the side chains involved in these packing interactions are invariant in all known K^+ channels it is reasonable to assume that the carbonyl oxygen atoms are fixed in positions with the correct dimensions to provide strong binding sites for K^+. The resolution of the structure determination is, however, too low to establish details of these binding sites.

On the basis of the structure, MacKinnon has suggested a plausible mechanism for the ion selectivity and conductivity of the channel. When an ion, which in solution has a water hydration shell, enters the selectivity filter it dehydrates. Binding to the carbonyl oxygen atoms in the filter compensates the energetic cost of dehydration. The dimensions of the binding sites are such that a K^+ ion fits in the filter precisely so that the energetic costs and gains are well balanced, but the firm packing of the side chains prevents the carbonyl oxygen atoms from approaching close enough to compensate for the cost of dehydration of a Na^+ ion.

In the crystal, the selectivity filter has two K^+ ions, one bound at each end of the filter, separated by a distance of 7.5 Å. This is the same distance as the average distance between K^+ ions in 4 M KCl. There is thus a high local concentration of K^+ ions in the filter, implying that the filter attracts and concentrates K^+ ions. However, since there are no negative ions within the filter to balance these positive charges there is also a repulsive force between the two K^+ ions. Since in the crystal there is no concentration gradient of K^+ ions across the channel there is no net flow of the ions. When, however, channel molecules are embedded in cell membranes across which there is a K^+ ion concentration gradient, the ions flow through the channel. They are forced through by a combination of the mutual repulsion of the ions in the channel and the membrane potential. Within the filter the ions cascade from one carbonyl atom to the next.

All K^+ channels are tetrameric molecules. There are two closely related varieties of subunits for K^+ channels, those containing two membrane-spanning helices and those containing six. However, residues that build up the ion channel, including the pore helix and the inner helix, show a strong sequence similarity among all K^+ channels. Consequently, the structural features and the mechanism for ion selectivity and conductance described for the bacterial K^+ channel in all probability also apply for K^+ channels in plant and animal cells.

The bacterial photosynthetic reaction center is built up from four different polypeptide chains and many pigments

The crystallographic world was stunned when at a meeting in Erice, Sicily, in 1982, Hartmut Michel of the Max-Planck Institute in Martinsried, Germany, displayed the x-ray diagram shown in Figure 12.12. Not only was this the first x-ray picture to high resolution of a membrane protein, but the crystal was

formed not from a small protein of trivial function but from a large complex of polypeptide chains that represents a class of proteins having a function of central importance for life on earth. The protein complex was a **photosynthetic reaction center** from the photosynthetic purple bacterium *Rhodopseudomonas viridis*, which converts the energy of captured sunlight into electrical and chemical energy in the first steps of photosynthesis by pumping protons from one side of a membrane to the other. The structure has subsequently been resolved to 2.5 Å by H. Michel in collaboration with Hans Deisenhofer and Robert Huber at the same institute.

The interiors of rhodopseudomonad bacteria are filled with photosynthetic vesicles, which are hollow, membrane-enveloped spheres. The photosynthetic reaction centers are embedded in the membrane of these vesicles. One end of the protein complex faces the inside of the vesicle, which is known as the periplasmic side; the other end faces the cytoplasm of the cell. Around each reaction center there are about 100 small membrane proteins, the antenna pigment protein molecules, which will be described later in this chapter. Each of these contains several bound chlorophyll molecules that catch photons over a wide area and funnel them to the reaction center. By this arrangement the reaction center can utilize about 300 times more photons than those that directly strike the special pair of chlorophyll molecules at the heart of the reaction center.

The reaction center is built up from four polypeptide chains, three of which are called L, M, and H because they were thought to have light, medium, and heavy molecular masses as deduced from their electrophoretic mobility on SDS-PAGE. Subsequent amino acid sequence determinations showed, however, that the H chain is in fact the smallest with 258 amino acids, followed by the L chain with 273 amino acids. The M chain is the largest polypeptide with 323 amino acids. This discrepancy between apparent relative masses and real molecular weights illustrates the uncertainty in deducing molecular masses of membrane-bound proteins from their mobility in electrophoretic gels.

The L and M subunits show about 25% sequence identity and are therefore homologous and evolutionarily related proteins. The H subunit, on the other hand, has a completely different sequence. The fourth subunit of the reaction center is a cytochrome that has 336 amino acids with a sequence that is not similar to any other known cytochrome sequence.

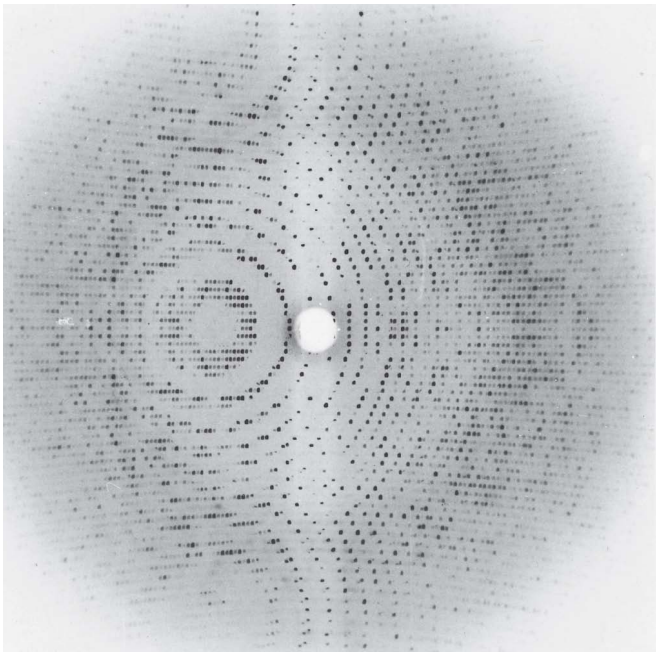


Figure 12.12 X-ray diffraction pattern from crystals of a membrane-bound protein, the bacterial photosynthetic reaction center. (Courtesy of H. Michel.)

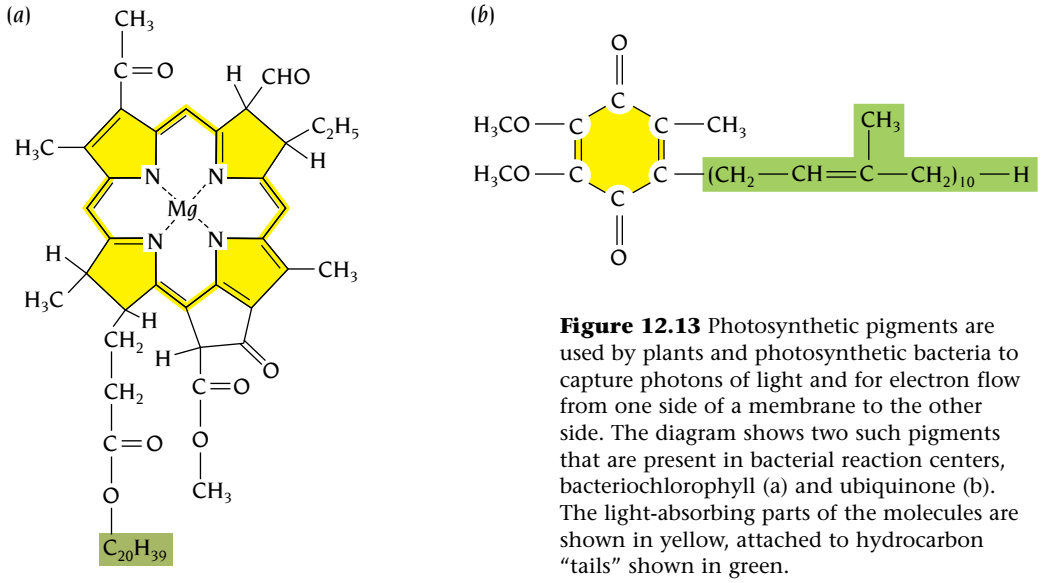


Figure 12.13 Photosynthetic pigments are used by plants and photosynthetic bacteria to capture photons of light and for electron flow from one side of a membrane to the other side. The diagram shows two such pigments that are present in bacterial reaction centers, bacteriochlorophyll (a) and ubiquinone (b). The light-absorbing parts of the molecules are shown in yellow, attached to hydrocarbon "tails" shown in green.

In addition to these polypeptide chains, the reaction center contains a number of pigments. There are four bacteriochlorophyll molecules (Figure 12.13a), two of which form the strongly interacting dimer called "the special pair." Furthermore, there is one Fe atom, a carotenoid, two quinone molecules (Figure 12.13b) and two bacteriopheophytin molecules, which are chlorophyll molecules without the central Mg^{2+} atom. Finally, the cytochrome subunit has four bound heme groups. The crystal structure shows how the polypeptide chains bind these pigments into a functional unit allowing electrons to flow from one side of the membrane to the other.

The L, M, and H subunits have transmembrane α helices

The L and the M subunits are firmly anchored in the membrane, each by five hydrophobic transmembrane α helices (yellow and red, respectively, in Figure 12.14). The structures of the L and M subunits are quite similar as expected from their sequence similarity; they differ only in some of the loop regions. These loops, which connect the membrane-spanning helices, form rather flat hydrophilic regions on either side of the membrane to provide interaction areas with the H subunit (green in Figure 12.14) on the cytoplasmic side and with the cytochrome (blue in Figure 12.14) on the periplasmic side. The H subunit, in addition, has one transmembrane α helix at the carboxy terminus of its polypeptide chain. The carboxy end of this chain is therefore on the same side of the membrane as the cytochrome. In total, eleven transmembrane α helices attach the L, M, and H subunits to the membrane.

No region of the cytochrome penetrates the membrane; nevertheless, the cytochrome subunit is an integral part of this reaction center complex, held through protein–protein interactions similar to those in soluble globular multisubunit proteins. The protein–protein interactions that bind cytochrome in the reaction center of *Rhodopseudomonas viridis* are strong enough to survive the purification procedure. However, when the reaction center of *Rhodobacter sphaeroides* is isolated, the cytochrome is lost, even though the structures of the L, M, and H subunits are very similar in the two species.

Alpha helices D and E from the L and M subunits (Figure 12.14) form the core of the membrane-spanning part of the complex. These four helices are tightly packed against each other in a way quite similar to the four-helix bundle motif in water-soluble proteins. Each of these four helices provides a histidine side chain as ligand to the Fe atom, which is located between the helices close to the cytoplasm. The role of the Fe atom is probably to

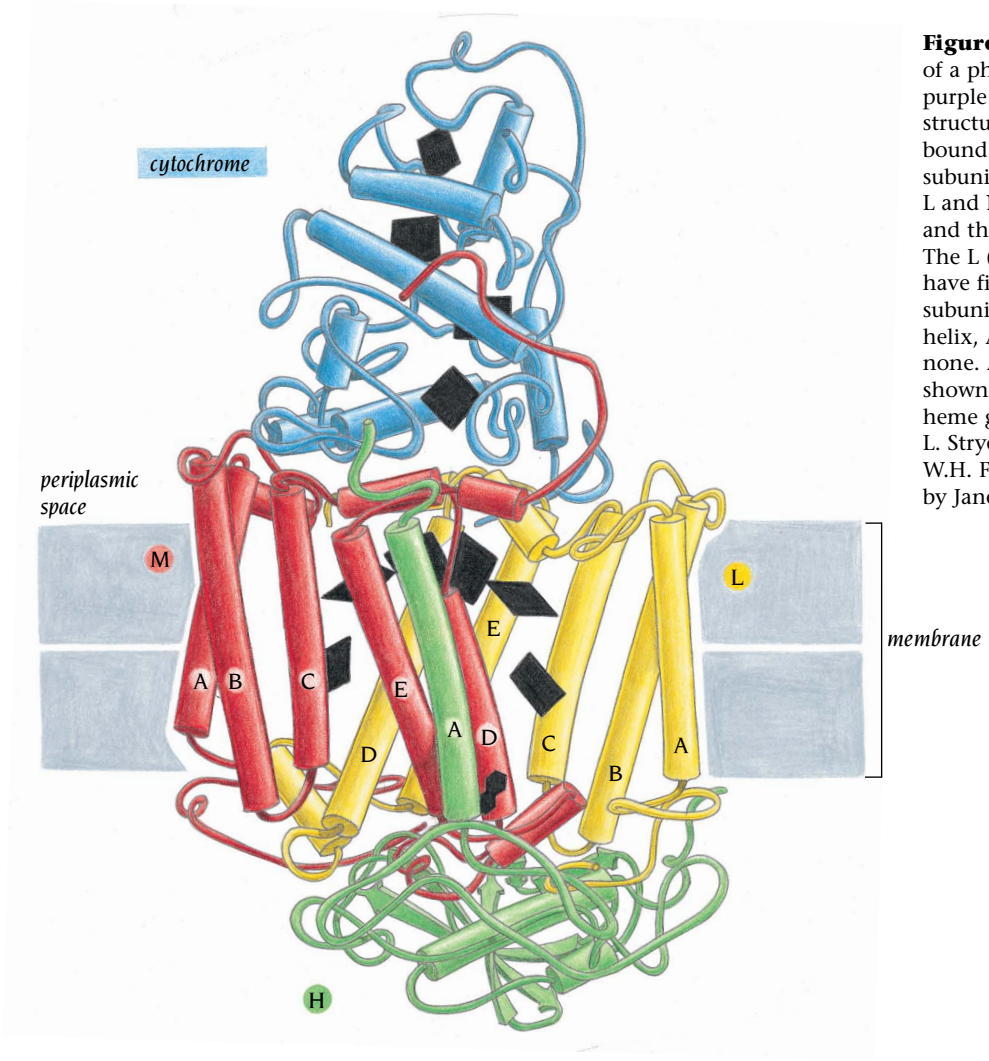


Figure 12.14 The three-dimensional structure of a photosynthetic reaction center of a purple bacterium was the first high-resolution structure to be obtained from a membrane-bound protein. The molecule contains four subunits: L, M, H, and a cytochrome. Subunits L and M bind the photosynthetic pigments, and the cytochrome binds four heme groups. The L (yellow) and the M (red) subunits each have five transmembrane α helices A–E. The H subunit (green) has one such transmembrane helix, AH, and the cytochrome (blue) has none. Approximate membrane boundaries are shown. The photosynthetic pigments and the heme groups appear in black. (Adapted from L. Stryer, *Biochemistry*, 3rd ed. New York: W.H. Freeman, 1988, after a drawing provided by Jane Richardson.)

stabilize the structure of the four-helix bundle. Since its removal does not change the rate of electron flow through the system, the Fe atom cannot have a functional role in electron transfer. The remaining three helices of each subunit are arranged around the core in a manner that is not found in water-soluble proteins. Presumably, their positions are at least partly determined by the positions of the loop regions outside the membrane and not by close packing of the α helices inside the membrane. It is interesting that none of the α helices are perpendicular to the assumed plane of the membrane; instead, they are all tilted at angles of about 20° to 25°, similar to the tilt of the transmembrane helices in bacteriorhodopsin (see Figure 12.3) and other transmembrane proteins.

The photosynthetic pigments are bound to the L and M subunits

The structurally similar L and M subunits are related by a pseudo-twofold symmetry axis through the core, between the helices of the four-helix bundle motif. The photosynthetic pigments are bound to these subunits, most of them to the transmembrane helices, and they are also related by the same twofold symmetry axis (Figure 12.15). The pigments are arranged so that they form two possible pathways for electron transfer across the membrane, one on each side of the symmetry axis.

This symmetry is important in bringing the two chlorophyll molecules of the "special pair" into close contact, giving them their unique function in initiating electron transfer. They are bound in a hydrophobic pocket close to the symmetry axis between the D and E transmembrane α helices of both

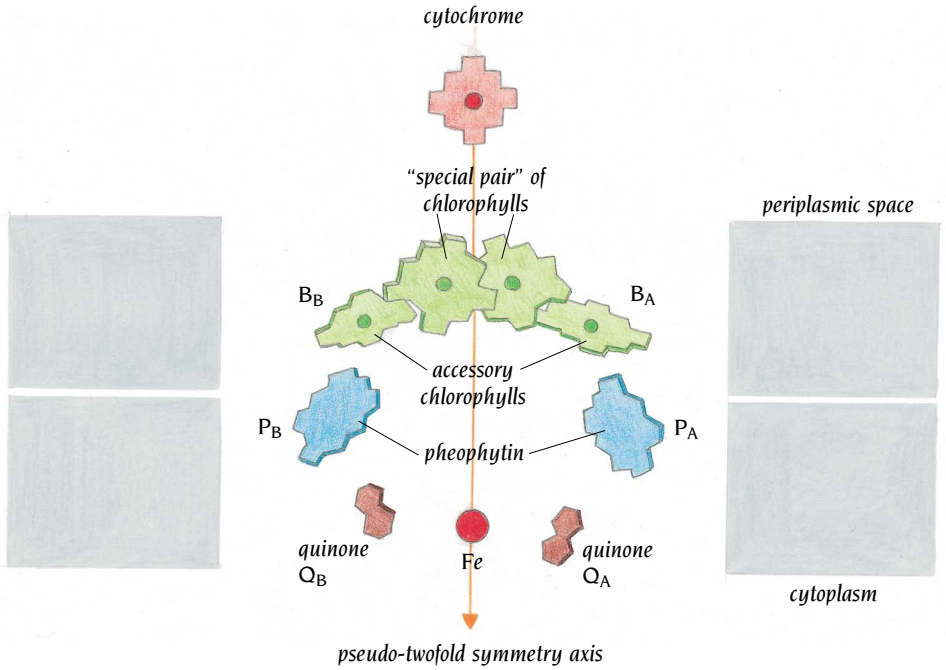


Figure 12.15 Schematic arrangement of the photosynthetic pigments in the reaction center of *Rhodopseudomonas viridis*. The twofold symmetry axis that relates the L and the M subunits is aligned vertically in the plane of the paper. Electron transfer proceeds preferentially along the branch to the right. The periplasmic side of the membrane is near the top, and the cytoplasmic side is near the bottom of the structure. (From B. Furugren, courtesy of the Royal Swedish Academy of Science.)

subunits. The pyrrol ring systems of these chlorophyll molecules are parallel and interact closely with each other; in particular, two pyrrol rings, one from each molecule, are stacked on top of each other, 3 Å apart (Figure 12.15). The twofold symmetry axis passes between these stacked pyrrol rings.

This pair of chlorophyll molecules, which as we shall see accepts photons and thereby excites electrons, is close to the membrane surface on the periplasmic side. At the other side of the membrane the symmetry axis passes through the Fe atom. The remaining pigments are symmetrically arranged on each side of the symmetry axis (Figure 12.15). Two bacteriochlorophyll molecules, the accessory chlorophylls, make hydrophobic contacts with the special pair of chlorophylls on one side and with the pheophytin molecules on the other side. Both the accessory chlorophyll molecules and the pheophytin molecules are bound between transmembrane helices from both subunits in pockets lined by hydrophobic residues from the transmembrane helices (Figure 12.16).

The functional reaction center contains two quinone molecules. One of these, Q_B (Figure 12.15), is loosely bound and can be lost during purification. The reason for the difference in the strength of binding between Q_A and Q_B is unknown, but as we will see later, it probably reflects a functional asymmetry in the molecule as a whole. Q_A is positioned between the Fe atom and one of the pheophytin molecules (Figure 12.15). The polar-head group is outside the membrane, bound to a loop region, whereas the hydrophobic tail is

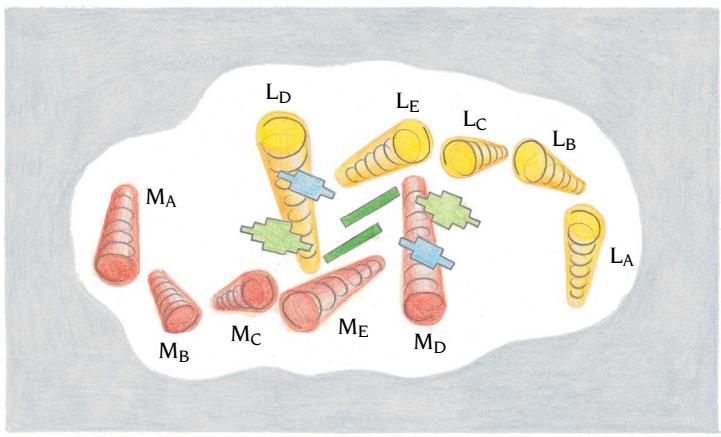


Figure 12.16 View of the reaction center perpendicular to the membrane illustrating that the pigments are bound between the transmembrane helices. The five transmembrane-spanning α helices of the L (yellow) and the M (red) subunits are shown as well as the chlorophyll (green) and pheophytin (blue) molecules.

bound to the pheophytin molecule and to hydrophobic side chains of transmembrane helices of the L subunit. If Q_B is lost during purification prior to crystallization, the corresponding binding site on the other side of the pseudo-twofold axis is empty. Certain weedkillers that are inhibitors of photosynthesis in plants can bind in the crystal to this empty binding site.

Reaction centers convert light energy into electrical energy by electron flow through the membrane

In photosynthesis light energy is converted to electrical energy by an electron flow that causes the separation of negatively and positively charged molecules. Many molecules can absorb photons and use the energy of this process to donate an electron to a nearby electron acceptor. The electron donor then becomes positively charged and the electron acceptor negatively charged. In most cases, however, the transfer of electrons back from the acceptor to the donor is as fast as the forward reaction and the absorbed energy is lost, usually as fluorescent radiation. The arrangement of photoreaction centers in both bacteria and green plants results in a very fast forward reaction and a slow back reaction; therefore, the electric charges induced by the absorbed light energy stay separated. This separation of charge represents a storage of energy because energy would be released if the charges were able to come together. This is the basic primary process of photosynthesis, the detailed mechanism of which we do not yet understand. However, by interpreting spectroscopic and genetic experiments in terms of the structure of the bacterial reaction center we may come to understand this process and be able to re-create this primary biological function—that of a solar-powered electrolytic battery.

In the bacterial reaction center the photons are absorbed by the special pair of chlorophyll molecules on the periplasmic side of the membrane (see Figure 12.14). Spectroscopic measurements have shown that when a photon is absorbed by the special pair of chlorophylls, an electron is moved from the special pair to one of the pheophytin molecules. The close association and the parallel orientation of the chlorophyll ring systems in the special pair facilitates the excitation of an electron so that it is easily released. This process is very fast; it occurs within 2 picoseconds. From the pheophytin the electron moves to a molecule of quinone, Q_A, in a slower process that takes about 200 picoseconds. The electron then passes through the protein, to the second quinone molecule, Q_B. This is a comparatively slow process, taking about 100 microseconds.

There are two pheophytin molecules, one on each side of the twofold axis, that in principle could accept the electrons (see Figure 12.15). However, only one pathway, on the right side of the symmetry axis that is shown in Figure 12.15, is used for electron transfer. Electrons do not pass through the chain of pigments on the left side, which appear to have no role in the charge separation. The best guess as to why these pigments are present is that the L and M chains have evolved from an ancestral chain that formed symmetrical homodimers in which both pigment chains were utilized. Presumably, the present-day reaction centers are more efficient for charge separation than the ancestral homodimers. Q_A is most stable when excited by two electrons, and if each electron arrived randomly at Q_A and Q_B the energy stored in Q_A after absorption of one electron might be dissipated before a second electron was absorbed.

One apparent discrepancy between the spectroscopic data and the crystal structure is that no spectroscopic signal has been measured for participation of the accessory chlorophyll molecule B_A in the electron transfer process. However, as seen in Figure 12.15, this chlorophyll molecule is between the special pair and the pheophytin molecule and provides an obvious link for electron transfer in two steps from the special pair through B_A to the pheophytin. This discrepancy has prompted recent, very rapid measurements of the electron transfer steps, still without any signal from B_A. This means either

that the electron bypasses B_A and is transferred directly over the very long distance of about 25 Å from the special pair to pheophytin, or that the transfer through B is too rapid to detect with current technology, less than 0.01 picosecond. Neither of these conclusions is compatible with current theories for electron transfer. However, the components for electron transfer are embedded in the protein environment of subunits L and M with special properties that are not taken into account in the theories.

While this electron flow takes place, the cytochrome on the periplasmic side donates an electron to the special pair and thereby neutralizes it. Then the entire process occurs again: another photon strikes the special pair, and another electron travels the same route from the special pair on the periplasmic side of the membrane to the quinone, Q_B, on the cytosolic side, which now carries two extra electrons. This quinone is then released from the reaction center to participate in later stages of photosynthesis. The special pair is again neutralized by an electron from the cytochrome.

The charge separation that stores the energy of the photons is now complete: two positive charges have been left on the cytochrome side of the membrane, and two electrons have traveled through the membrane to a quinone molecule on the cytosolic side. In the photosynthetic reaction centers charge separation is remarkably efficient for capturing light energy. The forward electron transfer from the reaction center to Q_A is more than eight orders of magnitude faster than the back reaction. This large difference allows the reaction center to capture the energy of between 98 and 100% of the photons it absorbs. As a solar cell, it is extraordinarily efficient: the energy stored in separated charges is about half of the energy inherent in the photons. The rest of the energy is lost in other ways, some of which are the reactions that drive the electrons along the pathway of photosynthetic pigment molecules.

Antenna pigment proteins assemble into multimeric light-harvesting particles

If the special pair of chlorophyll molecules of the reaction center was the only photon acceptor in the bacterial membrane, only a tiny fraction of the incoming sunlight would be captured and converted to chemical energy. All photosynthetic organisms have therefore evolved a system of **light-harvesting complexes,** which surrounds the reaction centers and increases the photon capturing area. The reaction centers receive practically all their light energy from such complexes. Detailed structural information is now available for the arrangement of the light-absorbing pigment molecules around the reaction center in photosynthetic bacteria. The pigments are firmly bound to small

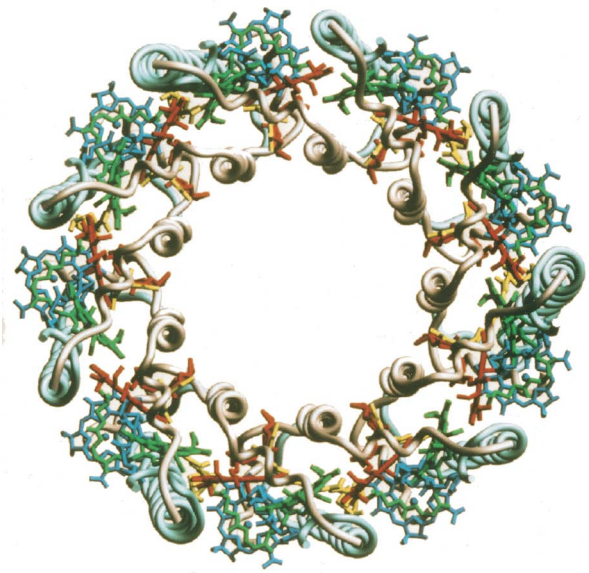


Figure 12.17 Computer-generated diagram of the structure of light-harvesting complex LH2 from *Rhodopseudomonas acidophila*. Nine α chains (gray) and nine β chains (light blue) form two rings of transmembrane helices between which are bound nine carotenoids (yellow) and 27 bacteriochlorophyll molecules (red, green and dark blue). (Courtesy of M.Z. Papiz.)

hydrophobic protein molecules that are embedded in the membrane and which assemble into two types of multimeric complexes, called LH1 and LH2. The crystal structures of two LH2 complexes have been determined to high resolution, one from the purple bacterium *Rhodopseudomonas acidophila* by the group of Richard Cogdell and Neil Isaacs, Glasgow University, UK, and the other from *Rhodospirillum molischianum* by the group of Hartmut Michel at the Max-Planck Institute in Frankfurt, Germany. In addition, an 8.5-Å electron crystallography map of LH1 using two-dimensional crystals has given a broad outline of the structure of that complex. Strangely, plants seem to have evolved a totally different system of light-harvesting complexes. Werner Kühlbrandt at EMBL, Heidelberg, has determined the structure of a plant light-harvesting complex by electron crystallography and shown that the structure is quite different from those of the purple bacteria. We will here discuss only the bacterial system.

Chlorophyll molecules form circular rings in the light-harvesting complex LH2

The structure of the LH2 complex of *R. acidophila* is both simple and elegant (Figure 12.17). It is a ring of nine identical units, each containing an α and a β polypeptide of 53 and 41 residues, respectively, which both span the membrane once as α helices (Figure 12.18). The two polypeptides bind a total of three chlorophyll molecules and two carotenoids. The nine heterodimeric units form a hollow cylinder with the α chains forming the inner wall and the β chains the outer wall. The hole in the middle of the cylinder is empty, except for lipid molecules from the membrane.

Two of the chlorophyll molecules from each unit are in the space between the two walls and form a ring of 18 chlorophyll molecules near the periplasmic membrane surface (Figure 12.19a). The planar chlorophyll rings are oriented almost perpendicular to the plane of the membrane. Each magnesium atom of these chlorophyll molecules is bound to a histidine residue of either the α or the β chain. The third chlorophyll molecule is bound between the β chains forming part of the outer wall and oriented parallel to the plane of the membrane (Figure 12.19b). The magnesium atom is bound to an oxygen atom of the formylated N-terminus of the α chain and has a much more polar environment than the other two chlorophyll molecules. The third chlorophyll molecule therefore absorbs light at a considerably shorter wavelength than the other two molecules, with an absorption maximum at 800 nm compared with 850 nm. This arrangement allows the complex efficiently to capture a much broader energy band of the sunlight than would be possible if all the chlorophyll molecules had similar environments.

In contrast to bacteriorhodopsin or the reaction center, there is no direct contact within the membrane between the α helices in this complex. The helices are held together through contacts mediated by the pigments and by contacts at the ends of the polypeptide chains outside the membrane.

The LH2 complex of *R. molischianum* is very similar to that of *R. acidophila* except that the complex is built up into a ring of eight identical units instead of nine. In addition the third chlorophyll is coordinated to the O atom of an Asp residue of the α chain. Since the α and β chains of all bacterial light-harvesting complexes show some sequence similarity one can safely predict that they are all arranged in essentially the same way, except that the number of units forming the cylinder may differ. This prediction has been used to build a model of LH1 using the electron crystallographic map to 8.5 Å.

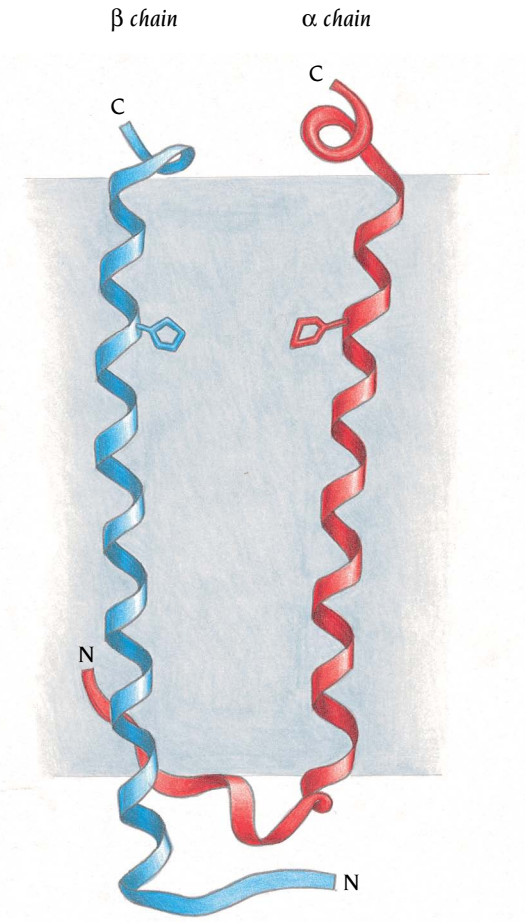


Figure 12.18 Ribbon diagram showing the α (red) and the β (blue) chains of the light-harvesting complex LH2. Each chain forms one transmembrane α helix, which contains a histidine residue that binds to the Mg atom of one bacteriochlorophyll molecule. (Adapted from G. McDermott et al., *Nature* 374: 517–521, 1995.)

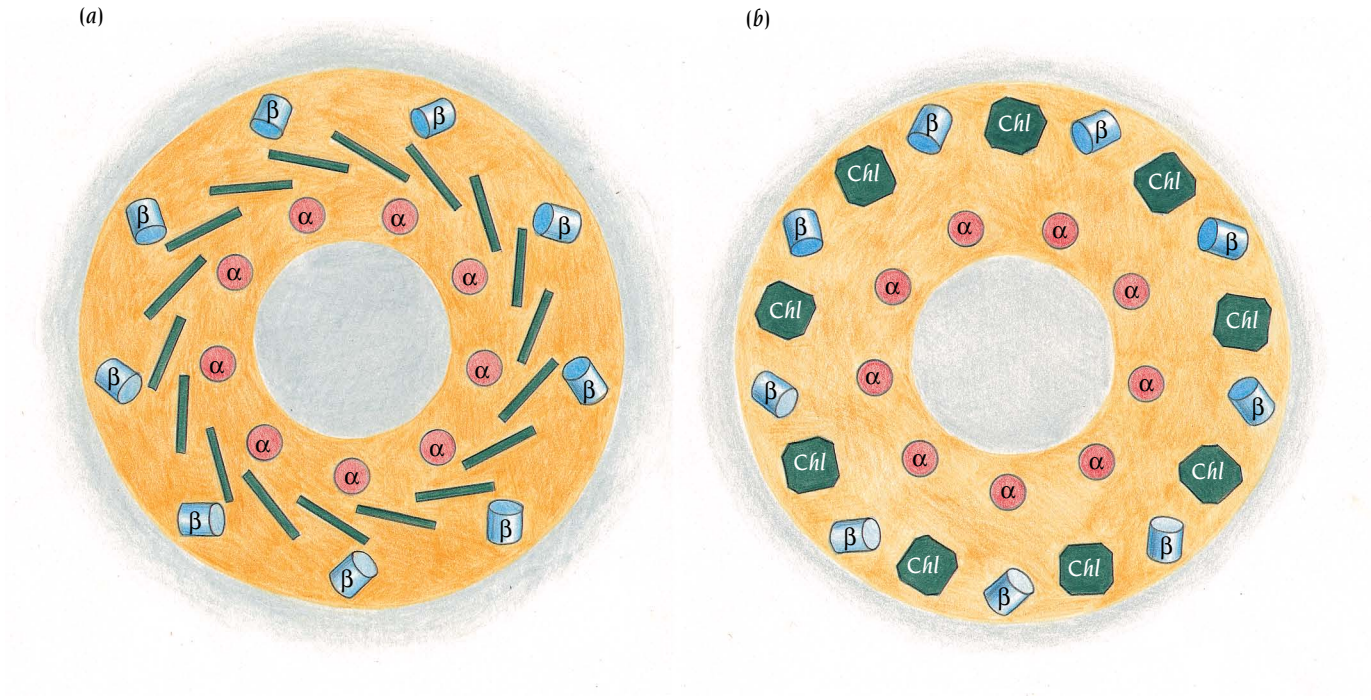


Figure 12.19 Schematic diagrams illustrating the arrangement of bacteriochlorophyll molecules in the light-harvesting complex LH2, viewed from the periplasmic space. (a) Eighteen bacteriochlorophyll molecules (green) are bound between the two rings of α (red) and β (blue) chains. The planes of these molecules are oriented perpendicular to the plane of the membrane and the molecules are bound close to the periplasmic space. (b) Nine bacteriochlorophyll molecules (green) are bound between the β chains (blue) with their planes oriented parallel to the plane of the membrane. These molecules are bound in the middle of the membrane.

The reaction center is surrounded by a ring of 16 antenna proteins of the light-harvesting complex LH1

The light-harvesting complex LH1 is directly associated with the reaction center in purple bacteria and is therefore referred to as the core or inner antenna, whereas LH2 is known as the peripheral antenna. Both are built up from hydrophobic α and β polypeptides of similar size and with low but significant sequence similarity. The two histidines that bind to chlorophyll with absorption maxima at 850 nm in the periplasmic ring of LH2 are also present in LH1, but the sequence involved in binding the third chlorophyll in LH2 is quite different in LH1. Not surprisingly, the chlorophyll molecules of the periplasmic ring are present in LH1 but the chlorophyll molecules with the 800 nm absorption maximum are absent.

The electron crystallographic map of LH1 at 8.5 Å resolution (Figure 12.20) shows a striking similarity in molecular design to LH2. The LH1 ring clearly consists of the same basic units as LH2. The α and β polypeptides form an inner and an outer wall with the pigment molecules in between. However, the LH1 ring contains 16 rather than 9 units and has a hole 68 Å in diameter in the middle. By analogy with the structure of LH2 it can be assumed that LH1 has a ring of 32 chlorophyll molecules in the region between these two walls at the same distance from the periplasmic side of the membrane as in LH2. These chlorophyll molecules presumably have an even more hydrophobic environment than those in LH2 since they have an absorption maximum at 875 nm.

On the basis of these structural results it is now possible to derive the following schematic picture (Figure 12.21) of the photosynthetic apparatus in purple bacteria, and to begin to understand the design of this highly efficient apparatus for capturing photons from the sun and funneling them to the reaction center with minimum loss of energy.

Spectroscopic measurements show that the reaction center and LH1 are tightly associated and therefore it is assumed that the ring of pigments in LH1 surrounds the reaction center. Careful model building indicates that the hole in the middle of LH1 is large enough to accommodate the whole reaction center molecule. We do not know exactly how the LH2 complexes are arranged in the membrane around the LH1–reaction center complex, but at least some of them should be in contact with the outer rim of LH1 for efficient

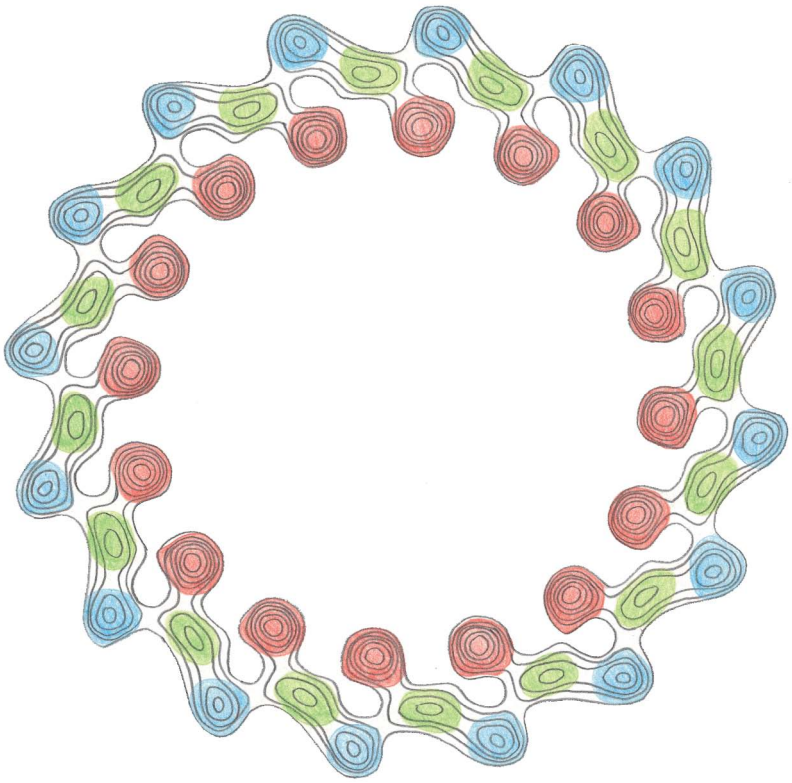


Figure 12.20 Electron density projection of the light-harvesting complex LH1 from *Rhodospirillum rubrum* determined by electron crystallography. On the basis of comparison with the LH2 complex the red regions can be interpreted as corresponding to the α chains, the blue regions to the β chains and each green region to two bacteriochlorophyll molecules bound between the α and the β chains. The ring of 16 units has a hole in the middle that is large enough to accommodate a complete reaction center molecule. (Adapted from Karrasch et al., *EMBO J.* 14: 631–638, 1995.)

energy transfer. There are 8 to 10 LH2 complexes for each reaction center and consequently around 300 energy-capturing chlorophyll molecules per reaction center.

In the photosynthetic membrane there is a downhill flow of energy from the light-harvesting proteins to the reaction centers. In purple bacteria, LH2 absorbs radiation at a shorter wavelength (higher energy) than LH1 and therefore delivers it to LH1, which in turn passes it on to the reaction center. A photon that is absorbed by the 800-nm chlorophylls in LH2 is rapidly transmitted to the energetically lower periplasmic ring of 850-nm chlorophyll in the same complex. Spectroscopic measurements have shown that the energy absorbed by a chlorophyll in the periplasmic rings spreads to the others, within 0.2 to 0.3 picoseconds. When a photon is absorbed by any of these chlorophylls it becomes in effect delocalized between the chlorophyll molecules of the ring. It can then easily jump to another chlorophyll in an adjacent complex, where it again becomes delocalized until it ends up at the reaction center, as schematically illustrated in Figure 12.22.

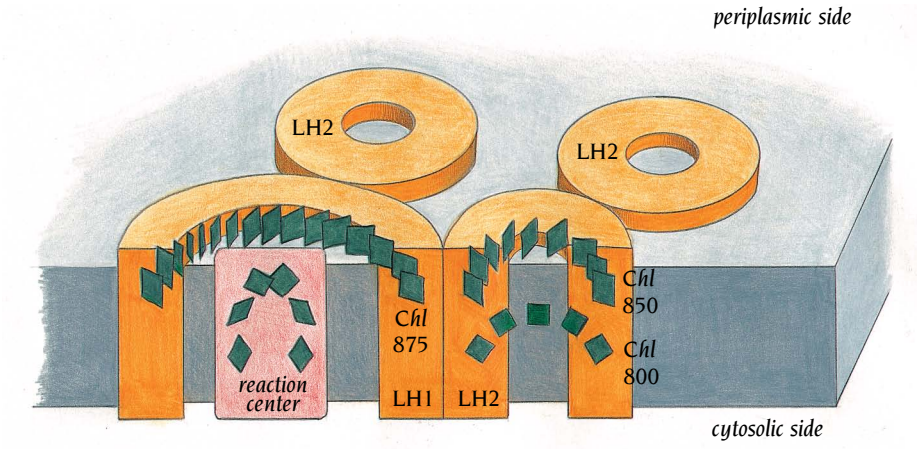


Figure 12.21 Schematic diagram of the relative positions of bacteriochlorophylls (green) in the photosynthetic membrane complexes LH1, LH2, and the reaction center. The special pair of bacteriochlorophyll molecules in the reaction center is located at the same level within the membrane as the periplasmic bacteriochlorophyll molecules Chl 875 in LH1 and the Chl 850 in LH2. (Adapted from W. Kühlbrandt, *Structure* 3: 521–525, 1995.)

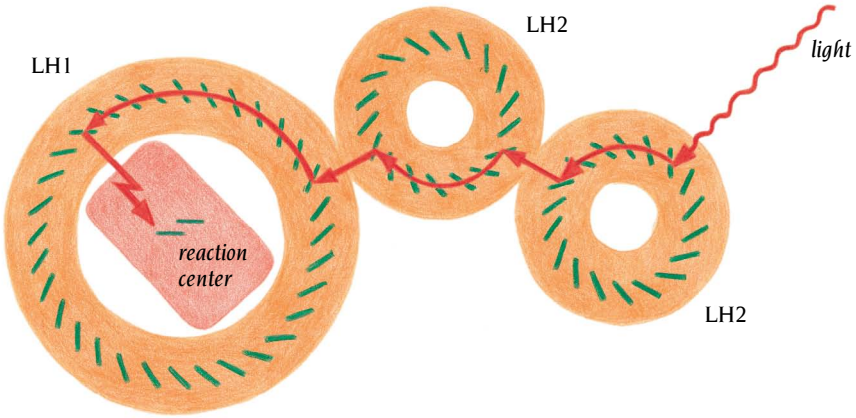


Figure 12.22 Schematic diagram showing the flow of excitation energy in the bacterial photosynthetic apparatus. The energy of a photon absorbed by LH2 spreads rapidly through the periplasmic ring of bacteriochlorophyll molecules (green). Where two complexes touch in the membrane , the energy can be transmitted to an adjacent LH2 ring. From there it passes by the same mechanism to LH1 and is finally transmitted to the special chlorophyll pair in the reaction center. (Adapted from W. Kühlbrandt, *Structure* 3: 521–525, 1995.)

Modeling of the reaction center inside the hole of LH1 shows that the primary photon acceptor—the special pair of chlorophyll molecules—is located at the same level in the membrane, about 10 Å from the periplasmic side, as the 850-nm chlorophyll molecules in LH2, and by analogy the 875-nm chlorophyll molecules of LH1. Furthermore, the orientation of these chlorophyll molecules is such that very rapid energy transfer can take place within a plane parallel to the membrane surface. The position and orientation of the chlorophyll molecules in these rings are thus optimal for efficient energy transfer to the reaction center.

Energy transfer and energy conversion in photosynthetic systems occur virtually without energy loss, whereas solid-state solar cells operate at efficiencies of around 20%. By learning some of the lessons implicit in the arrangement and the reactions of the pigment molecules in the bacterial photosynthetic membrane, and applying these principles to the design of new types of solar cells, it may one day become possible to devise a system of clean, environmentally compatible energy production that operates efficiently at low light intensities.

Transmembrane α helices can be predicted from amino acid sequences

We have seen in previous chapters that only short continuous regions of the polypeptide chains contribute to the hydrophobic interior of water-soluble globular proteins. In such proteins α helices are generally arranged so that one side of the helix is hydrophobic and faces the interior while the other side is hydrophilic and at the surface of the protein as discussed in Chapter 3. Beta strands are usually short, with the residues alternating between the hydrophobic interior and hydrophilic surface. Loop regions between these secondary structure elements are usually very hydrophilic. Therefore, in soluble globular proteins, regions of more than 10 consecutive hydrophobic amino acids in the sequence are rarely encountered.

In contrast, the transmembrane helices observed in the reaction center are embedded in a hydrophobic surrounding and are built up from continuous regions of predominantly hydrophobic amino acids. To span the lipid bilayer, a minimum of about 20 amino acids are required. In the photosynthetic reaction center these α helices each comprise about 25 to 30 residues, some of which extend outside the hydrophobic part of the membrane. From the amino acid sequences of the polypeptide chains, the regions that comprise the transmembrane helices can be predicted with reasonable confidence.

Naively, one might assume that it should be possible to scan the sequence and pick out regions with about 20 consecutive hydrophobic amino acids. However, no such regions occur in the reaction center proteins. Just as in soluble proteins there are hydrophobic side chains at the

Table 12.1 Hydrophobicity scales

Amino acid	Phe	Met	Ile	Leu	Val	Cys	Trp	Ala	Thr	Gly	Ser	Pro	Tyr	His	Gln	Asn	Glu	Lys	Asp	Arg
A	2.8	1.9	4.5	3.8	4.2	2.5	-0.9	1.8	-0.7	-0.4	-0.8	-1.6	-1.3	-3.2	-3.5	-3.5	-3.5	-3.9	-3.5	-4.5
B	3.7	3.4	3.1	2.8	2.6	2.0	1.9	1.6	1.2	1.0	0.6	-0.2	-0.7	-3.0	-4.1	-4.8	-8.2	-8.8	-9.2	-12.3

Row A is from J. Kyte and R.F. Doolittle; row B, from D.A. Engelman, T.A. Steitz, and A. Goldman.

hydrophilic surface of the molecule, in the transmembrane helices of the reaction center there are hydrophilic side chains (which are often important for function) among the hydrophobic. However, hydrophobic residues are in a clear majority in transmembrane helices, and such residues occur less frequently in other continuous regions of the polypeptide chain. Therefore, we need some method to measure the amount of hydrophobicity in a segment of the amino acid sequence in order to be able to predict whether or not it is likely to be a transmembrane helix.

Hydrophobicity scales measure the degree of hydrophobicity of different amino acid side chains

Each amino acid side chain within a transmembrane helix has a different hydrophobicity. It is easy to state that side chains such as Val, Met, and Leu are the most hydrophobic and that charged residues such as Arg and Asp are at the other end of the scale. However, to order all side chains according to hydrophobicity and to assign actual numbers that represent their degree of hydrophobicity is not trivial. Many such **hydrophobicity scales** have been developed over the past decade on the basis of solubility measurements of the amino acids in different solvents, vapor pressures of side-chain analogs, analysis of side-chain distributions within soluble proteins, and theoretical energy calculations. In Table 12.1 two of these hydrophobicity scales are listed. The most frequently used scale, which was introduced by J. Kyte and R.F. Doolittle at University of California, San Diego, is based on experimental data. A more refined scale was developed by D.A. Engelman, T.A. Steitz, and A. Goldman at Yale University. They used a semitheoretical approach to calculating the hydrophobicity, taking into account the fact that the side chains are attached to an α-helical framework.

Hydropathy plots identify transmembrane helices

These hydrophobicity scales are frequently used to identify those segments of the amino acid sequence of a protein that have hydrophobic properties consistent with a transmembrane helix. For each position in the sequence, a hydropathy index is calculated. The **hydropathy index** is the mean value of the hydrophobicity of the amino acids within a "window," usually 19 residues long, around each position. In transmembrane helices the hydropathy index is high for a number of consecutive positions in the sequence. Charged amino acids are usually absent in the middle region of transmembrane helices because it would cost too much energy to have a charged residue in the hydrophobic lipid environment. It might be possible, however, to have two residues of opposite charge close together inside the lipid membrane because they neutralize each other. Such charge neutralization has been observed in the hydrophobic interior of soluble globular proteins.

When the hydropathy indices are plotted against residue numbers, the resulting curves, called hydropathy plots, identify possible transmembrane helices as broad peaks with high positive values. Such hydropathy plots are shown in Figure 12.23 for the L and M chains of the reaction center.

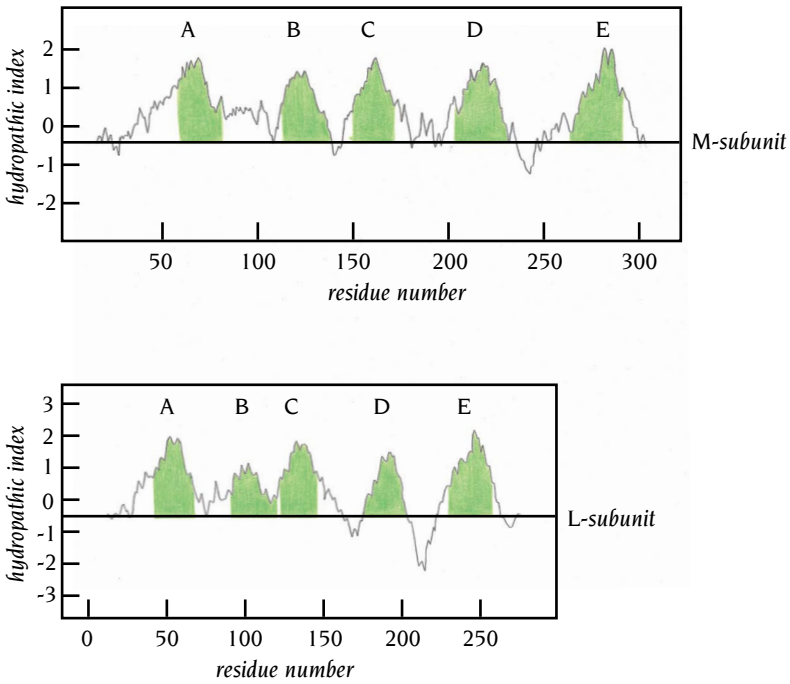


Figure 12.23 Hydropathy plots for the polypeptide chains L and M of the reaction center of *Rhodobacter sphaeroides*. A window of 19 amino acids was used with the hydrophobicity scales of Kyte and Doolittle. The hydropathy index is plotted against the tenth amino acid of the window. The positions of the transmembrane helices as found by subsequent x-ray analysis by the group of G. Feher, La Jolla, California, are indicated by the green regions.

Reaction center hydropathy plots agree with crystal structural data

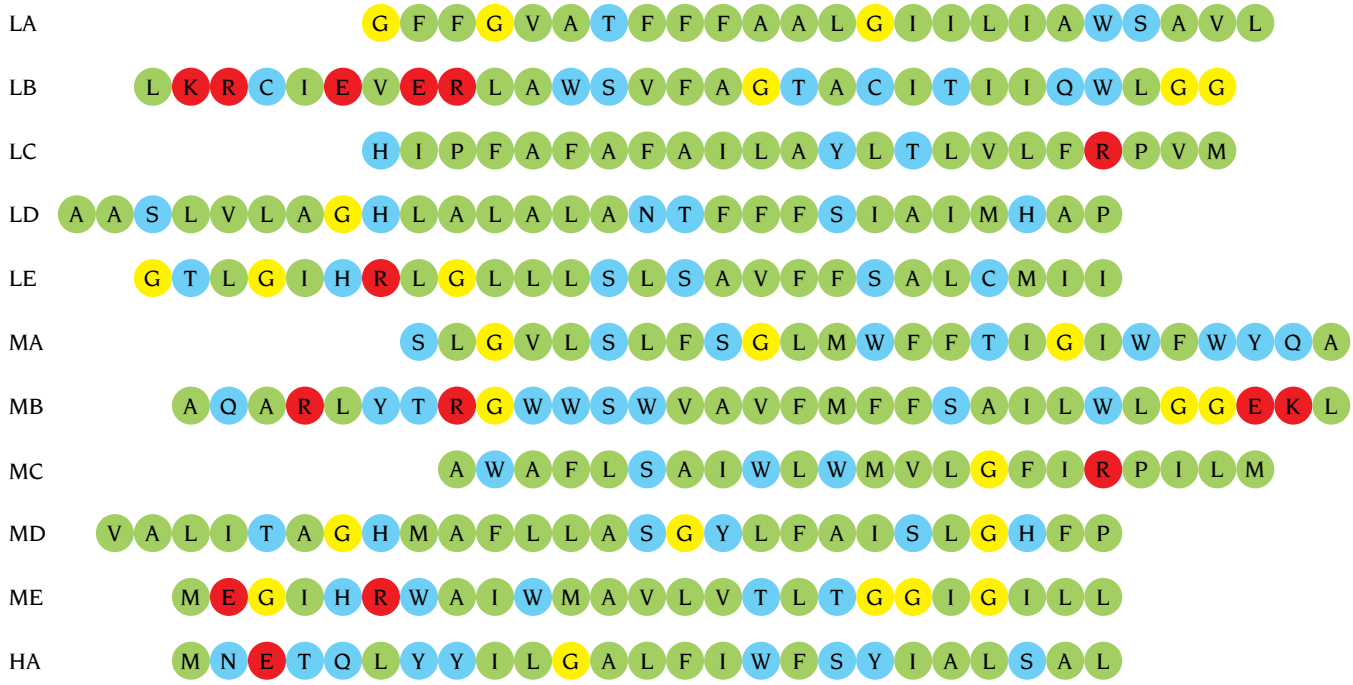
The hydropathy plots in Figure 12.23 were calculated and published several years before the x-ray structure of the reaction center was known. It is therefore of considerable interest to compare the predicted positions of the transmembrane-spanning helices with those actually observed in the x-ray structure. These observed positions are indicated in green in Figure 12.23.

It is immediately apparent that these plots correctly predict the number of transmembrane helices, five each in the L and M chains, and also their approximate positions in the polypeptide chain. This gives us confidence in the hydropathy plot method. Transmembrane helices are the only secondary structure elements that can be predicted from novel amino acid sequences with a high degree of confidence using current knowledge and methods. The exact ends of transmembrane helices, however, cannot be predicted, essentially because they are usually inserted within the polar head-groups of the membrane lipids and therefore contain charged and polar residues. The transmembrane helices in the reaction center, for example, contain a number of charged residues at their ends (Table 12.2), most of which are at the cytoplasmic side. All of the helices, however, have a segment of at least 19 consecutive amino acids that contain no charged side chains. The majority of residues in these segments are hydrophobic, but there are a number of polar residues, such as Ser, Thr, Tyr, and Trp, among them. The presence of histidine residues in the D and E helices of subunits L and M is accounted for by their special function in the reaction center; they are ligands to the magnesium atoms of chlorophyll molecules and to the Fe atom.

Membrane lipids have no specific interaction with protein transmembrane α helices

Comparison of the amino acid sequences of the L and M subunits of the reaction centers from three different bacterial species shows that about 50% of all residues in those two subunits are conserved in all three species. In the transmembrane helices, sequence conservation varies. Residues that are buried and have contacts either with pigments or with other transmembrane helices are about 60% conserved. In contrast, residues that are fully exposed to the membrane lipids are only 16% conserved. Clearly, fewer restrictions are

Table 12.2 Amino acid sequences of the transmembrane helices of the photosynthetic reaction center in *Rhodobacter sphaeroides*



The helices are aligned according to approximate positions within the membrane and with respect to the photosynthetic pigments. LA is the first helix of subunit L, ME is the last helix of subunit M, HA is the only transmembrane helix of subunit H. Charged residues are colored red, polar residues are blue, hydrophobic residues are green, and glycine is yellow. (From T.O. Yeates et al., *Proc. Natl. Acad. Sci. USA* 84: 6438–6442, 1987.)

placed on residues that are exposed to the membrane lipids than to residues having contact with polypeptide chains or pigments. This implies that there are relatively few specific interactions between these transmembrane helices and the fatty acid side chains of the membrane that require the presence of specific residues. This is consistent with the observation that membrane proteins can move within the plane of the membrane, by lateral diffusion, and are not at fixed positions.

Conclusion

The x-ray structure of a bacterial photosynthetic reaction center and the associated light-harvesting complexes has given insight into the mechanism for the primary reaction of photosynthesis: the capture and conversion of light energy to chemical energy by a stable separation of negatively and positively charged molecules. In addition, the structure has provided geometrical constraints for theoretical calculations on the electron flow through pigments across the membrane during this charge separation.

Important novel information has thus been obtained for the specific biological function of those molecules, but disappointingly few general lessons have been learned that are relevant for other membrane-bound proteins with different biological functions. In that respect the situation is similar to the failure of the structure of myoglobin to provide general principles for the construction of soluble protein molecules as described in Chapter 2.

The most important general lesson is that there are hydrophobic transmembrane helices, the positions of which within the amino acid sequence can be predicted with reasonable accuracy. This applies both to the single transmembrane-spanning helix within the H polypeptide chain of the reaction center and the five transmembrane helices of the L and M chains that

are connected by loop regions. This does not imply, however, that it is equally simple to predict the positions of transmembrane helices in different classes of proteins by hydropathy plots. The transmembrane helices of ion channels, for example, contain charged residues facing the channel. The latter would give quite different hydropathy plots from a single transmembrane helix in a receptor protein.

The structure of the reaction center also established that membrane-spanning helices can be tilted with respect to the plane of the membrane and that their relative positions within the membrane might be determined by the way they are anchored to the loop regions. Finally, several structures provide examples of how binding pockets for ligands are formed between such transmembrane-spanning helices.

The three-dimensional structure of the bacterial membrane protein, bacteriorhodopsin, was the first to be obtained from electron microscopy of two-dimensional crystals. This method is now being successfully applied to several other membrane-bound proteins.

Unlike the other membrane proteins discussed here, the porins in the outer membrane of Gram-negative bacteria have transmembrane regions that are β strands, not α helices. The β strands, 16 in some porins, 18 in others, are arranged into up and down antiparallel β barrels. These barrels, because they have so many strands in their walls, do not have a tightly packed hydrophobic core. Their center is a channel, which is partially blocked by loop regions between two of the strands, leaving an eyelet about 9 Å long and 8 Å in diameter. The complete functional porin molecule comprises three of these barrels; loop regions from the upper surface of the three barrels are in contact with the extracellular space and form a broad funnel leading to the barrels and via their eyelets to the periplasmic space. The porins in this way form size-restricted channels for the passive diffusion of molecules in and out of the periplasmic space.

Like the photosynthetic reaction center and bacteriorhodopsin, the bacterial K^+ ion channel also has tilted transmembrane helices, two in each of the subunits of the homotetrameric molecule that has fourfold symmetry. These transmembrane helices line the central and inner parts of the channel but do not contribute to the remarkable 10,000-fold selectivity for K^+ ions over Na^+ ions. This crucial property of the channel is achieved through the narrow selectivity filter that is formed by loop regions from the four subunits and lined by main-chain carbonyl oxygen atoms, to which dehydrated K^+ ions bind.

Selected readings

General

Barber, J., Andersson, B. Revealing the blueprint of photosynthesis. *Nature* 370: 31–34, 1994.

Cowan, S.W. Bacterial porins: lessons from three high-resolution structures. *Curr. Opin. Struct. Biol.* 3: 501–507, 1993.

Cowan, S.W., Rosenbusch, J.P. Folding pattern diversity of integral membrane proteins. *Science* 264: 914–916, 1994.

Deisenhofer, J., Michael, H. Nobel lecture. The photosynthetic reaction center from the purple bacterium *Rhodopseudomonas viridis*. *EMBO J.* 8: 2149–2169, 1989.

Engelman, D.M., Steitz, T.A., Goldman, A. Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. *Annu. Rev. Biophys. Biophys. Chem.* 15: 321–353, 1986.

Fasman, G.D., Gilbert, W.A. The prediction of transmembrane protein sequences and their conformation: an evaluation. *Trends Biochem. Sci.* 15: 89–95, 1990.

Garavito, R.M., Rosenbusch, J.P. Isolation and crystallization of bacterial porin. *Methods Enzymol.* 125: 309–328, 1986

Gust, D., Moore, T.A. Mimicking photosynthesis. *Science* 244: 35–41, 1989.

Huber, R. Nobel lecture. A structural basis of light energy and electron transfer in biology. *EMBO J.* 8: 2125–2147, 1989.

Kovari, L.C., Momany, C., Rossmann, M.C. The use of antibody fragments for crystallization and structure determinations. *Structure* 3: 1291–1293, 1995.

Kühlbrandt, W. Structure and function of bacterial light-harvesting complexes. *Structure* 3: 521–525, 1995.

Landau, E.M., Rosenbuch, J.P. Lipid cubic phases: a concept for the crystallization of membrane proteins. *Proc. Natl. Acad. Sci. USA* 93: 14532–14535, 1996.

Lanyi, J.K. Bacteriorhodopsin as a model for proton pumps. *Nature* 375: 461–464, 1995.

Michel, H. Crystallization of membrane proteins. *Trends Biochem. Sci.* 8: 56–59, 1983.

Michel, H., Deisenhofer, J. Relevance of the photosynthetic reaction center from purple bacteria to the structure of photosystem II. *Biochemistry* 27: 1–7, 1988.

Nikaido, H. Porins and specific diffusion channels in bacterial outer membranes. *J. Biol. Chem.* 269: 3905–3908, 1994.

Norris, J.R., Schiffer, M. Photosynthetic reaction centers in bacteria. *Chem. Eng. News* 68(31): 22–37, 1990.

Rees, D.C., et al. The bacterial photosynthetic reaction center as a model for membrane proteins. *Annu. Rev. Biochem.* 58: 607–633, 1989.

Unwin, N., Henderson, R. The structure of proteins in biological membranes. *Sci. Am.* 250(2): 78–95, 1984.

Youvan, D.C., Marrs, B.L. Molecular mechanisms of photosynthesis. *Sci. Am.* 256(6): 42–49, 1987

Specific structures

Allen, J.P., et al. Structure of the reaction center from *Rhodobacter sphaeroides* R-26: the cofactors. *Proc. Natl. Acad. Sci. USA* 84: 5730–5734, 1987.

Allen, J.P., et al. Structure of the reaction center from *Rhodobacter sphaeroides* R-26: the protein subunits. *Proc. Natl. Acad. Sci. USA* 84: 6162–6166, 1987.

Brisson, A., Unwin, P.N.T. Quaternary structure of the acetylcholine receptor. *Nature* 315: 474–477, 1985.

Cowan, S.W., et al. Crystal structures explain functional properties of two *E. coli* porins. *Nature* 358: 727–733, 1992.

Deisenhofer, J., et al. Structure of the protein subunits in the photosynthetic reaction center of *Rhodopseudomonas viridis* at 3 Å resolution. *Nature* 318: 618–624, 1985.

Deisenhofer, J., et al. X-ray structure analysis of a membrane protein complex. Electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction center from *Rhodopseudomonas viridis*. *J. Mol. Biol.* 180: 385–398, 1984.

Doyle, D.A., et al. The structure of the potassium channel: molecular basis of K^+ conduction and selectivity. *Science* 280: 69–77, 1998.

Dutzler, R., et al. Crystal structures of various maltooligosaccharides bound to maltoporin reveal a specific sugar translocation pathway. *Structure* 4: 127–134, 1996.

Eisenberg, D., Weiss, R.M., Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. USA* 82: 140–144, 1984.

Garavito, R.M., Rosenbusch, J.P. Three-dimensional crystals of an integral membrane protein: an initial x-ray analysis. *J. Cell. Biol.* 86: 327–329, 1980.

Henderson, R., et al. Model for the structure of bacteriorhodopsin based on high-resolution electron cryomicroscopy. *J. Mol. Biol.* 213: 899–929, 1990.

Henderson, R., Unwin, P.N.T. Three-dimensional model of purple membrane obtained by electron microscopy. *Nature* 257: 28–32, 1975.

Iwata, S., Ostermeier, C., Ludwig, B., Michel, H. Structure at 2.8 Å resolution of cytochrome c oxidase from *Paracoccus denitrificans*. *Nature* 376: 660–669, 1995.

Karrasch, S., Bullough, P.A., Ghosh, R. 8.5-Å projection map of the light-harvesting complex I from *Rhodospirillum rubrum* reveals a ring composed of 16 subunits. *EMBO J.* 14: 631–638, 1995.

Koepke, J., et al. The crystal structure of the light-harvesting complex II (B800–850) from *Rhodospirillum molischianum*. *Structure* 4: 581–597, 1996.

Kühlbrandt, W., Wang, D.A., Fujiyoshi, Y. Atomic model of the plant light-harvesting complex. *Nature* 367: 614–621, 1994.

Kyte, J., Doolittle, R.F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* 157: 105–132, 1982.

Leifer, D., Henderson, R. Three-dimensional structure of orthorhombic purple membrane at 6.5 Å resolution. *J. Mol. Biol.* 163: 451–466, 1983.

MacKinnon, R., et al. Structural conservation in prokaryotic and eukaryotic potassium channels. *Science* 280: 106–109, 1998.

McDermott, G., et al. Crystal structure of an integral membrane light-harvesting complex from photosynthetic bacteria. *Nature* 374: 517–521, 1995.

Michel, H. Three-dimensional crystals of a membrane protein complex. The photosynthetic reaction center from *Rhodopseudomonas viridis*. *J. Mol. Biol.* 158: 567–572, 1982.

Michel, H., et al. The "heavy" subunit of the photosynthetic reaction center from *Rhodopseudomonas viridis*: isolation of the gene, nucleotide and amino acid sequence. *EMBO J.* 4: 1667–1672, 1985.

Michel, H., et al. The "light" and "medium" subunits of the photosynthetic reaction center from *Rhodopseudomonas viridis*: isolation of the genes, nucleotide and amino acid sequence. *EMBO J.* 5: 1149–1158, 1986.

Michel, H., Epp, O., Deisenhofer, J. Pigment–protein interactions in the photosynthetic reaction center from *Rhodopseudomonas viridis*. *EMBO J.* 5: 2445–2451, 1986.

Picot, D., Loll, P.J., Garavito, R.M. The x-ray crystal structure of the membrane protein prostaglandin H2 synthase-1. *Nature* 367: 243–249, 1994.

Rees, D.C., DeAntonio, L., Eisenberg, D. Hydrophobic organization of membrane proteins. *Science* 245: 510–513, 1989.

Schirmer, T., Keller, T.A., Wang, Y.-F., Rosenbusch, J.P. Structural basis for sugar translocation through maltoporin channels at 3.1 Å resolution. *Science* 267: 512–514, 1995.

Tsukihara, T., et al. The whole structure of the 13-subunit oxidized cytochrome c oxidase at 2.8 Å resolution. *Science* 272: 1136–1144, 1996.

Unwin, N. Acetylcholine receptor channel imaged in the open state. *Nature* 373: 37–43, 1995.

Valpuesta, J.M., Henderson, R., Frey, T.G. Electron cryomicroscopic analysis of crystalline cytochrome oxidase. *J. Mol. Biol.* 214: 237–251, 1990.

Weiss, M.S., et al. Molecular architecture and electrostatic properties of a bacterial porin. *Science* 254: 1627–1630, 1991.

Yeates, T.O., et al. Structure of the reaction center from *Rhodobacter sphaeroides* R-26: membrane–protein interactions. *Proc. Natl. Acad. Sci. USA* 84: 6438–6442, 1987.

Prediction, Engineering, and Design of Protein Structures

17

Over a period of more than 3 billion years, a large variety of protein molecules has evolved to become the complex machinery of present-day cells and organisms. These molecules have evolved by random changes of genes by point mutations, exon shuffling, recombination and gene transfer between species, in combination with natural selection for those gene products that have conferred some functional advantage contributing to the survival of individual organisms.

Long before Darwin and Wallace proposed the theory of evolution and Mendel discovered the laws of genetics, plant and animal breeders had begun to interfere with the process of evolution in the species that gave rise to domesticated animals and cultivated plants. Considering their total lack of knowledge of both evolutionary theory and genetics, their achievements, brought about by forcing the pace of and subverting natural selection, were impressive albeit very gradual. With the advent of molecular genetics and in particular techniques for gene manipulation, we have now entered an era of genetic exploitation of organisms undreamed of only 50 years ago. We can now design genes to produce, in host organisms, novel gene products for the benefit of human beings; we are no longer restricted to selecting useful genes that arise by mutation. We are, however, only at the beginning of this new era, and so far we have only scratched the surface of the knowledge that is required for true engineering and design of protein molecules. We distinguish **protein engineering**, by which we mean mutating the gene of an existing protein in an attempt to alter its function in a predictable way, from **protein design**, which has the more ambitious goal of designing *de novo* a protein to fulfill a desired function.

Genome projects have now provided us with a description of the complete sequences of all the genes in more than a dozen organisms, and they will provide many more complete genome sequences within the next decade, including that of the human genome. These databases provide great opportunities for the analysis and exploitation of genes and their corresponding proteins. Central to reaping the intellectual and commercial benefits of this genetic information is the ability to find out the function of individual gene products. Almost all functional assignments to date have been based on sequence similarity to proteins of known function.

Knowledge of a protein's tertiary structure is a prerequisite for the proper understanding and engineering of its function. Unfortunately, in spite of recent significant technological advances, the experimental determination

of tertiary structure is still slow compared with the rate of accumulation of amino acid sequence data. This makes the **folding problem**, the successful prediction of a protein's tertiary structure from its amino acid sequence, central to rapid progress in post-genomic biology. We will, therefore, in this chapter first briefly describe implications of protein homology and methods for the prediction of secondary and tertiary structure before giving some examples of protein engineering and protein design.

Homologous proteins have similar structure and function

The term **homology** as used in a biological context is defined as similarity of structure, physiology, development and evolution of organisms based upon common genetic factors. The statement that two proteins are homologous therefore implies that their genes have evolved from a common ancestral gene.

Homologous proteins are mostly recognized by statistically significant similarities in their amino acid sequences. Usually, they also have similar functions although there are some known exceptions, where genes for ancient enzymes have been recruited at a later stage in evolution to produce proteins with quite different functions. An example is provided by one of the structural components in the eye lens that is homologous to the ancient glycolytic enzyme lactate dehydrogenase. Once a novel gene has been cloned and sequenced, a search for amino acid sequence similarity between the corresponding protein and other known protein sequences should be made. Usually, this is done by comparison with databases of known protein sequences using one of the standard sequence alignment computer programs.

Two proteins are considered to be homologous when they have identical amino acid residues in a significant number of sequential positions along the polypeptide chains. Using statistical methods based on comparisons of computer-generated random sequences, it is relatively straightforward to assess how many positions need to be identical for a statistically significant identity between two sequences. However, it is frequently found that two proteins with sequence identity below the level of statistical significance have similar functions and similar three-dimensional structures. In these cases, functionally important residues are identical and usually such residues form sequence patterns or motifs that can be used to identify other proteins that belong to the same functional family. Frequently, members of such families are also considered to be homologous, even though the identities are not statistically significant, only functionally significant. Databases for such families, based on identical or similar sequence motifs, are available on the World Wide Web (see pp. 393–394) and they are very useful for assigning function to a novel protein.

If significant amino acid sequence identity is found with a protein of known crystal structure, a three-dimensional model of the novel protein can be constructed, using computer modeling, on the basis of the sequence alignment and the known three-dimensional structure. This model can then serve as an excellent basis for identifying amino acid residues involved in the active site or in antigenic epitopes, and the model can be used for protein engineering, drug design, or immunological studies.

Since the sequence databases are large and growing exponentially, currently comprising more than 500,000 known protein sequences, the standard sequence alignment programs have been designed to provide a compromise between the speed and the accuracy of the search. As a result, they work well only when there is a reasonably high degree of sequence identity, usually of the order of 30% or more. Much more sensitive programs have been written that search for both identity and conserved structural properties and also for relatedness in different physical properties, but these inevitably require far more computing time. Carefully used, such programs can identify structural and functional similarity where the standard programs fail to do so.

Homologous proteins have conserved structural cores and variable loop regions

Homologous proteins always contain a core region where the general folds of the polypeptide chains are very similar. This core region contains mainly the secondary structure elements that build up the interior of the protein: in other words, the scaffolds of homologous proteins have similar three-dimensional structures. Even distantly related proteins with low sequence identity have similar scaffold structures, although minor adjustments occur in the positions of the secondary structure elements to accommodate differences in the arrangements of the hydrophobic side chains in the interior of the protein. The greater the sequence identity, the more closely related are the scaffold structures (Figure 17.1). This has important implications for model building of homologous proteins; the more distantly related two proteins are, the more the scaffold must be adjusted to model the new structure.

Loop regions that connect the building blocks of scaffolds can vary considerably both in length and in structure. The problem of predicting the three-dimensional structure of a protein that is homologous to a protein of known three-dimensional structure is therefore mainly a question of predicting the structure of loop regions and side-chain conformations, after the scaffold has been adjusted. As mentioned in Chapter 2, loop regions do not have random structures, and their main-chain conformations cluster in sets of similar structures. The conformation of each set depends more on the number of amino acids in the loop and the type of secondary structure elements that it connects, whether they are α-α, β-β, α-β, or β-α connections, than on the actual amino acid sequences. Therefore it is possible to use a database of loop regions from proteins of known structure to obtain a preliminary model of the loops of an unknown structure. To model a protein structure, suitable main-chain loop conformations from this database are attached to the scaffold modeled to have a structure similar to that of the known homologous protein. Finally, the conformations of the side chains are predicted by energy refinement of the model, which minimizes the free energy of the protein by maximizing the interaction energies of the amino acids. Analysis of structures determined to high resolution has shown that only a few side-chain conformations frequently occur. These are called rotamers and model building of side chains employs databases of such rotamers.

An instructive example of the use of such procedures has been in modeling antigen-binding sites in immunoglobulins. These binding sites are built up from three hypervariable loop regions, CDR1–CDR3, from the variable domains of both the light and the heavy chains of immunoglobulins as described in Chapter 15. There is usually high sequence identity within the scaffolds of the variable domains in different immunoglobulin molecules. Consequently, the scaffold of variable domains of known three-dimensional structures can be used in modeling a new monoclonal antibody with a known amino acid sequence. However, the CDR regions of a new antibody are usually very different in sequence from those of any other known antibody, and their three-dimensional structures must be predicted. By comparing

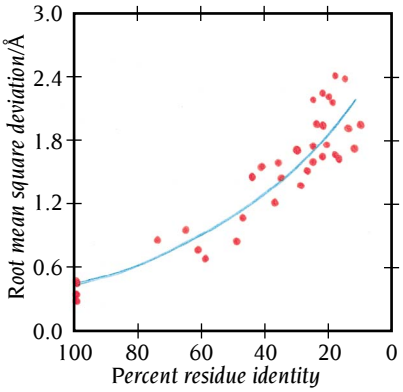
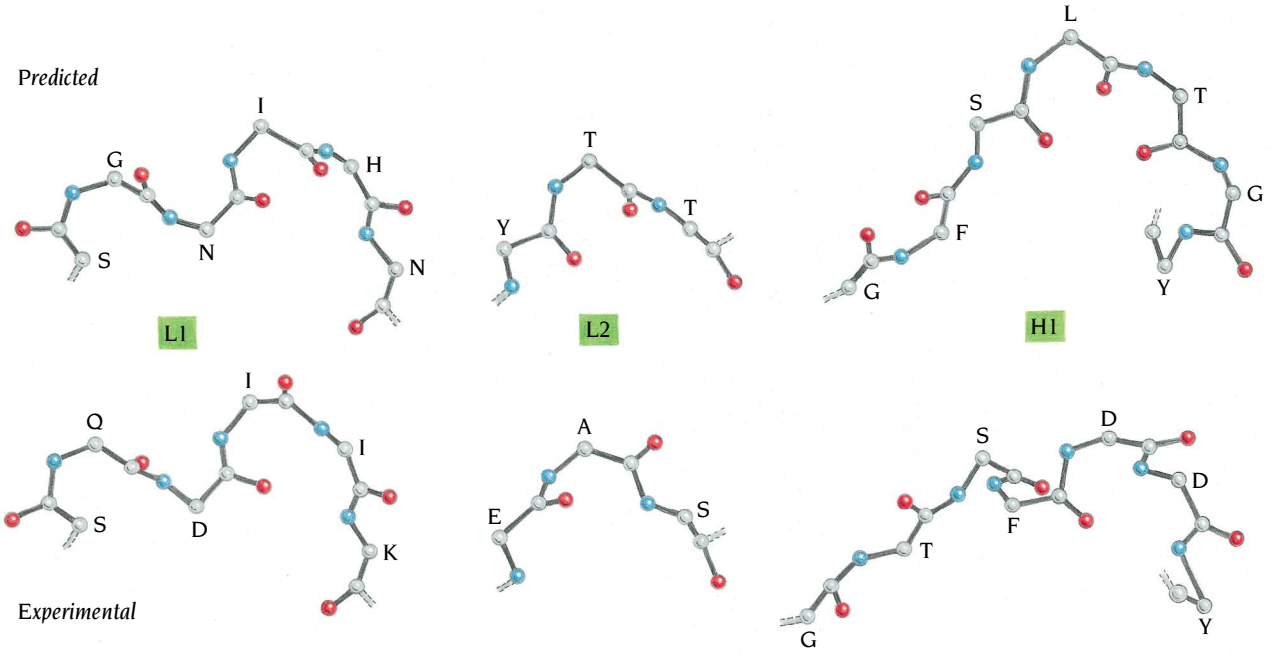


Figure 17.1 The relation between the divergence of amino acid sequence and three-dimensional structure of the core region of homologous proteins. Known structures of 32 pairs of homologous proteins such as globins, serine proteinases, and immunoglobulin domains have been compared. The root mean square deviation of the main-chain atoms of the core regions is plotted as a function of amino acid homology (red dots). The curve represents the best fit of the dots to an exponential function. Pairs with high sequence homology are almost identical in three-dimensional structure, whereas deviations in atomic positions for pairs of low homology are of the order of 2 Å. (From C. Chothia and A. Lesk, *EMBO J.* 5: 823–826, 1986.)



known antibody structures and sequences, it has been shown that there is only a small repertoire of main-chain conformations for at least five of the six CDR regions and that the particular conformation adopted is determined by a few key conserved residues for each loop conformation. For example, three different conformations were found for the CDR3 regions of the light chains in nine known x-ray structures. More than 90% of the known sequences of light-chain CDR3 regions obey the sequence constraints of one or other of these three conformations. By using this repertoire of loop conformations, considerable success has been achieved in correctly predicting the structure of antigen-binding surfaces. An example of such a prediction compared with the actual structure, subsequently determined, is given in Figure 17.2.

Figure 17.2 An example of prediction of the conformations of three CDR regions of a monoclonal antibody (*top row*) compared with the unrefined x-ray structure (*bottom row*). L1 and L2 are CDR regions of the light chain, and H1 is from the heavy chain. The amino acid sequences of the loop regions were modeled by comparison with the sequences of loop regions selected from a database of known antibody structures. The three-dimensional structure of two of the loop regions, L1 and L2, were in good agreement with the preliminary x-ray structure, whereas H1 was not. However, during later refinement of the x-ray structure errors were found in the conformations of H1, and in the refined x-ray structure this loop was found to agree with the predicted conformations. In fact, all six loop conformations were correctly predicted in this case. (From C. Chothia et al., *Science* 233: 755–758, 1986.)

Knowledge of secondary structure is necessary for prediction of tertiary structure

What can be done by predictive methods if the sequence search fails to reveal any homology with a protein of known tertiary structure? Is it possible to model a tertiary structure from the amino acid sequence alone? There are no methods available today to do this and obtain a model detailed enough to be of any use, for example, in drug design and protein engineering. This is, however, a very active area of research and quite promising results are being obtained; in some cases it is possible to predict correctly the type of protein, α, β, or α/β, and even to derive approximations to the correct fold.

Today's predictive methods rely on prediction of secondary structure: in other words, which amino acid residues are α-helical and which are in β strands. We have emphasized in Chapter 12 that secondary structure cannot in general be predicted with a high degree of confidence with the possible exceptions of transmembrane helices and α-helical coiled coils. This imposes a basic limitation on the prediction of tertiary structure. Once the correct secondary structure is known, we know enough about the rules for packing elements of secondary structure against each other (see Chapter 2 for helix packing) to derive a very limited number of possible stable globular folds. Consequently, secondary structure prediction lies at the heart of the prediction of tertiary structure from the amino acid sequence.

Unfortunately for predictive methods, secondary and tertiary structures are closely linked in the sense that global tertiary structure imposes local secondary structure at least in some regions of the polypeptide chain. The ability of a specific short sequence of amino acids to form an α helix, a β strand, or a loop region is dependent not only on the sequence of that region but also on its environment in the three-dimensional structure. For example, by analyzing all the known tertiary structures, it has been shown that peptide regions of up to five residues long with identical amino acid sequences are α-helical in one structure and a β strand or a loop in other structures. While this interdependence of secondary and tertiary structure complicates secondary structure predictions, it can, sometimes, be used to improve such predictions, by an iterative scheme in which a preliminary assignment of secondary structure is used to predict the type of domain structure, for example, a four-helix bundle or an α/β barrel. The structure type of the domain imposes additional constraints on possible secondary structure, which can be used to refine the secondary structure prediction.

Prediction methods for secondary structure benefit from multiple alignment of homologous proteins

Over 20 different methods have been proposed for predictions of secondary structure; they can be categorized in two broad classes. The empirical statistical methods use parameters obtained from analyses of known sequences and tertiary structures. All such methods are based on the assumption that the local sequence in a short region of the polypeptide chain determines local structure; as we have seen, this is not a universally valid assumption. The second group of methods is based on stereochemical criteria, such as compactness of form with a tightly packed hydrophobic core and a polar surface. Three frequently used methods are the empirical approaches of P.Y. Chou and G.D. Fasman and of J. Garnier, D.J. Osguthorpe and B. Robson (the GOR method), and third, the stereochemical method of V.I. Lim.

Although these three methods use quite different approaches to the problem, the accuracy of their secondary structure prediction is about the same. All three methods can be used to assign one of three states to each residue: α helix, β strand, or loop. Random assignment of these three states to residues in a polypeptide chain will give an average score of 33% correctly predicted states. The methods have been assessed in an analysis of single sequences of a large number of known x-ray structures comprising more than 10,000 residues. For the three-state definition of secondary structure, the overall accuracy of prediction was about 55%. Other objective assessments have given similar results.

However, when these predictive methods are used on a set of homologous proteins the predictive power is considerably higher. The underlying assumption is that secondary and tertiary structure has been more conserved during evolution than amino acid sequence; in other words only such changes have been retained during evolution that conserve the structure. Consequently, the pattern of residue changes within homologous proteins contains specific information about the structure. Conserved hydrophobic residues are usually in the interior of the protein with a high probability of belonging to helices or sheet strands. Insertions and deletions almost always occur in loop regions and not in the scaffold built up from helices and strands.

Several programs are now available that use multiple alignment of homologous proteins for prediction of secondary structure. One such program, called PHD, which was developed by Chris Sander and coworkers, EMBL, Heidelberg, has reached a mean accuracy of prediction of 72% for new structures.

A large fraction of the remaining errors occur at the ends of α helices and β strands and, in addition, some errors occur because of occasional difficulties

in distinguishing between α helices and β strands. These latter errors can be corrected if the structural class, α, β, or α/β, can be deduced from a combination of physical studies, for example, circular dichroism spectra, and the general features of the secondary structure prediction. For example, if the prediction scheme assigns one or two short α helices among many β strands in a protein of the β class, there is a high probability that the regions of secondary structures are essentially correctly predicted but that they should all be β strands.

These predictive methods are very useful in many contexts; for example, in the design of novel polypeptides for the identification of possible antigenic epitopes, in the analysis of common motifs in sequences that direct proteins into specific organelles (for instance, mitochondria), and to provide starting models for tertiary structure predictions.

Many different amino acid sequences give similar three-dimensional structures

How many completely different amino acid sequences might give a similar three-dimensional structure for an average-sized domain of 150 amino acid residues? Simple combinatorial calculations show that there are a total of 20^{150} or roughly 10^{200} possible amino acid sequences for such a domain, given the 20 different amino acids in natural proteins. This number is much larger than the number of atoms in the known universe. A more laborious calculation shows that out of these 10^{200} possible combinations we can extract about 10^{38} members that have less than 20% amino acid sequence identity with each other and that therefore can be considered to have different sequences. In other words, there are 10^{38} different ways of constructing a domain of 150 amino acids using the 20 standard amino acids as building blocks. We do not know how many of these can form a stable three-dimensional structure but, assuming say that one out of a billion (10^9) can, we are left with 10^{29} folded possible proteins. In the previous chapters we have seen that simple structural motifs arrange themselves into a limited number of topologically different domain structures. It has been estimated on reasonable grounds that there are about 1000 topologically different domain structures. Since there are 10^{29} possible different sequences that might fold into 10^3 different structures, it follows that there are of the order of 10^{26} different side chain arrangements with less than 20% amino acid sequence identity that can give similar polypeptide folds. Only a small fraction of these possible proteins will be found in nature.

For each of the 500 or so different domain structures that have so far been observed, we might at best know about a dozen of these different possible sequences. It is not trivial to recognize the general sequence patterns that are common to specific domain structures from such a limited knowledge base.

Prediction of protein structure from sequence is an unsolved problem

How to predict the three-dimensional structure of a protein from its amino acid sequence is the major unsolved problem in structural molecular biology. We would like to have a computer program that could simulate the action of the processes that operate in a test tube or a living cell when a polypeptide chain with a specific amino acid sequence folds into a precise three-dimensional structure. Why is this prediction of protein folding so difficult? The answer is usually formulated in terms of the complexity of the task of searching through all the possible conformations of a polypeptide chain to find those with low energy. It requires enormous amounts of computing time, in addition to the complication discussed in Chapter 6 that the energy difference between a stable folded molecule and its unfolded state is a small number containing large errors.

With the realization that there are only a limited number of stable folds and many unrelated sequences that have the same fold, biologically oriented computer scientists started to address what is called the **inverse folding problem**; namely, which sequence patterns are compatible with a specific fold? If this question can be answered, such patterns could be used to search through the genome sequence databases and extract those sequences that have a specific fold, such as the α/β barrel or the immunoglobulin fold.

However, given the large number of possible unrelated sequences for each fold and the limited number of known sequences, a variation of this problem has recently been addressed by a large number of groups; namely, which of the known folds, if any, is most compatible with a specific sequence? The methodology used is called **threading** because it involves threading a specific sequence through all known folds and, for each fold, estimating the probability that the sequence can have that fold. Considerable progress has recently been made in threading, and in blind tests several structures have been correctly predicted by different groups.

Threading methods can assign amino acid sequences to known three-dimensional folds

Threading methods, which are also called protein fold assignments or fold recognition, are a promising and rapidly evolving field of computational structural biology. The goal is to assign to each genome-derived protein sequence the protein fold to which it most closely corresponds, or to determine whether there is no known fold to which the sequence belongs. A further goal is to align the new sequence properly to the three-dimensional structure of the fold to which it belongs to provide a low-resolution model. In order to test different methods of threading, blind tests are arranged, called Critical Assessment of Structure Prediction (CASP), in which the participants are given sequences and invited to predict the fold and make an alignment before the structure is determined experimentally. We will briefly describe here the methods used by one of the more successful participants in these tests, the group of David Eisenberg at University of California, Los Angeles.

The first requirement for threading is to have a database of all the known different protein folds. Eisenberg has used his own library of about 800 folds, which represents a minimally redundant set of the more than 6000 structures deposited at the Protein Data Bank. Other groups use databases available on the World Wide Web, where the folds are hierarchically ordered according to structural and functional similarities, such as SCOP, designed by Alexey Murzin and Cyrus Chothia in Cambridge, UK.

For each fold one searches for the best alignment of the target sequence that would be compatible with the fold; the core should comprise hydrophobic residues and polar residues should be on the outside, predicted helical and strand regions should be aligned to corresponding secondary structure elements in the fold, and so on. In order to match a sequence alignment to a fold, Eisenberg developed a rapid method called the 3D profile method. The environment of each residue position in the known 3D structure is characterized on the basis of three properties: (1) the area of the side chain that is buried by other protein atoms, (2) the fraction of side chain area that is covered by polar atoms, and (3) the secondary structure, which is classified in three states: helix, sheet, and coil. The residue positions are rather arbitrarily divided into six classes by properties 1 and 2, which in combination with property 3 yields 18 environmental classes. This classification of environments enables a protein structure to be coded by a sequence in an 18-letter alphabet, in which each letter represents the environmental class of a residue position.

Each of the 20 different amino acids has different preferences for each of the 18 environmental classes; for instance a Leu has a high preference for being in a helical class with a high fraction of buried side chain area, whereas

an Asp has a very low preference for that position. Numerical values for these preferences, called 3D-1D scores, were derived from a set of well-refined high-resolution protein structures, together with sets of sequences similar to the sequences of the 3D structures. This produced a scoring table in which for each environmental class a numerical value of preference is associated with each of the 20 amino acids. This table is used to set up a 3D profile table of a protein structure, in which each residue position is assigned an environmental class with corresponding numerical values for preference for each type of amino acid. The essence of this method is that the three-dimensional structure is reduced to a one-dimensional array, which facilitates matching to a one-dimensional sequence.

A target amino acid sequence is aligned against this structure profile in such a way that the best possible match—the highest total score—is obtained, allowing gaps and insertions. Such an alignment is conceptually similar to alignment of two sequences and similar methods have been used. The match of a sequence to a 3D structure profile for a specific fold is expressed quantitatively by a value called the Z-score, which is the number of standard deviations above the mean alignment score for other sequences of similar length. A high Z-score means there is a high probability that the sequence has the corresponding fold.

The methods described here have subsequently been improved and extended by Eisenberg, but the principle remains essentially the same. Other groups use different methods to screen the sequence–structure alignments and different criteria to assess the matches. Manfred Sippl at the University of Salzburg, Austria, has developed a set of potentials to screen and assess the alignments, the essence of which is to maximize the number of hydrophobic interactions and to minimize the number of buried polar atoms that do not participate in hydrogen bonds. These and similar potentials are now used by many groups in their threading programs. Correct folds can be predicted with a reasonably high probability for small and medium-sized proteins. Correct alignment of the sequence to the selected fold is, however, less accurate.

Proteins can be made more stable by engineering

Protein engineering, via site-directed mutagenesis of DNA, can be used to answer very specific questions about protein stability, and the results of these studies are now being used to increase the stability of industrially important enzymes. To illustrate some of the factors of importance for protein stability that have been revealed by protein engineering studies, we have chosen the extensive work on the enzyme lysozyme from bacteriophage T4 that has been done by the group of Brian Mathews, University of Oregon, Eugene.

Lysozyme from bacteriophage T4 is a 164 amino acid polypeptide chain that folds into two domains (Figure 17.3) There are no disulfide bridges; the two cysteine residues in the amino acid sequence, Cys 54 and Cys 97, are far apart in the folded structure. The stability of both the wild-type and mutant proteins is expressed as the melting temperature, Tm, which is the temperature at which 50% of the enzyme is inactivated during reversible heat denaturation. For the wild-type T4 lysozyme the Tm is 41.9 °C.

We will discuss three different approaches to engineer a more thermostable protein than wild-type T4 lysozyme, namely (1) reducing the difference in entropy between folded and unfolded protein, which in practice means reducing the number of conformations in the unfolded state, (2) stabilizing the α helices, and (3) increasing the number of hydrophobic interactions in the interior core.

Disulfide bridges increase protein stability

The greater the number of unfolded conformations of a protein, the higher the entropic cost of folding that protein into its single native state (see Chapter 6). Reducing the number of unfolded conformations therefore increases

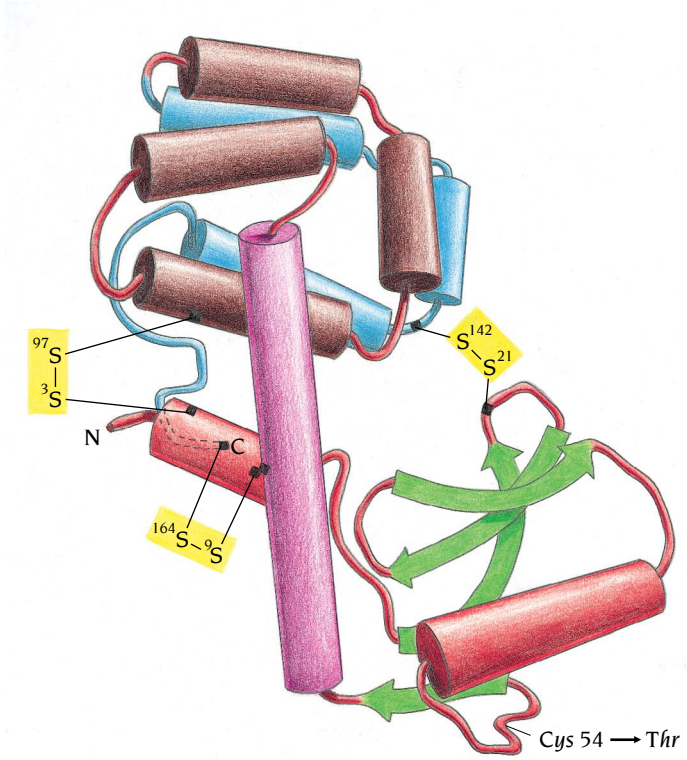


Figure 17.3 The polypeptide chain of lysozyme from bacteriophage T4 folds into two domains. The N-terminal domain is of the α + β type, built up from two α helices (red) and a four-stranded antiparallel β sheet (green). The C-terminal domain comprises seven short α helices (brown and blue) in a rather irregular arrangement. (The last half of this domain is colored blue for clarity.) One long α helix connects the two domains (purple). Thermostable mutants of this protein were constructed by introducing disulfide bridges at three different places (yellow). The position of Cys 54, which was mutated to Thr, is also shown. (Adapted from M. Matsumura et al., *Nature* 342: 291–293, 1989.)

the stability of the native state. The most obvious way to decrease the number of unfolded conformations is to introduce a novel disulfide bond based on knowledge of the tertiary structure of the folded protein. The longer the loop between the cysteine residues, the more restricted is the unfolded polypeptide chain, giving more stabilization of the folded structure. To design such bridges is, however, not a simple task, since the geometry of an unstrained -CH_2-S-S-CH_2- bridge in proteins is confined to rather narrow conformational limits, and deviations from this geometry will introduce strains into the folded structure and hence reduce rather than increase its stability. It is, therefore, not sufficient to choose at random two residues close together in space to make such a bridge, rather the protein engineer must carefully select pairs of residues with main-chain conformations that fulfill the conditions needed for an unstrained disulfide bridge.

Mathews made a very careful comparison between the geometry of the 295 disulfide bridges in known x-ray structures and all possible pairs of amino acid residues close enough to each other in the refined T4 lysozyme structure to accommodate a disulfide bridge. This was followed by energy minimization of the most likely candidate disulfide bridges and an analysis of stabilizing interactions present in the wild-type structure that would be lost by mutation to a Cys residue. Such losses should be minimized. Three candidate disulfide bridges remained after this filtering, one of which, Cys 3–Cys 97, contained one of the cysteine residues (Cys 97) that is present in the wild type. The five amino acid residues—Ile 3, Ile 9, Thr 21, Thr 142, and Leu 164 (see Figure 17.3)—were mutated to Cys residues in separate experiments so that all single (3–97, 9–164, and 21–142) as well as combinations of double and triple disulfide bonds could be formed. In addition, the second Cys residue of the wild-type enzyme, Cys 54, was mutated to Thr to avoid the formation of incorrect disulfide bonds during folding.

The results of this careful design of novel disulfide bridges were very encouraging (Figure 17.4). All the mutants were more stable in their oxidized forms than wild-type protein. The longer the loop between the cysteine

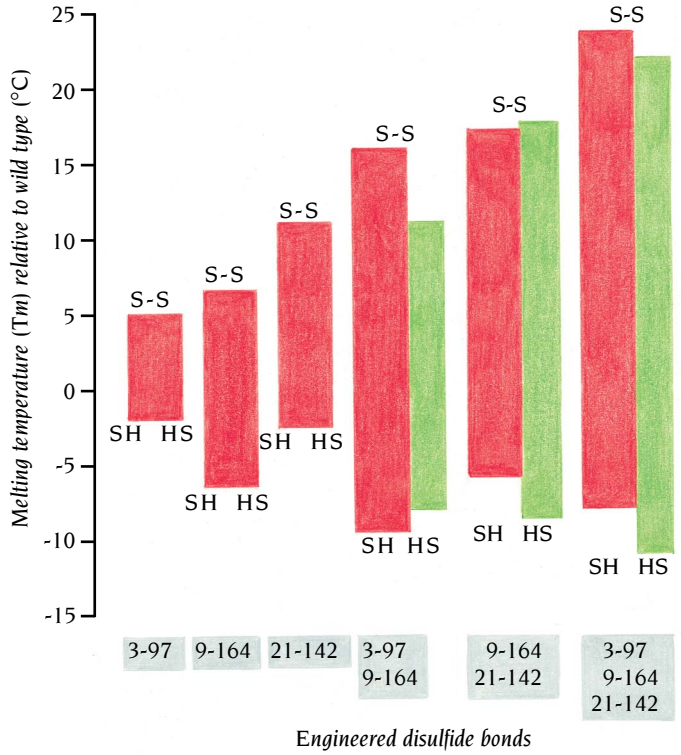


Figure 17.4 Melting temperatures, Tm, of engineered single-, double-, and triple-disulfide-containing mutants of T4 lysozyme relative to wild-type lysozyme. The red bars show the differences in Tm values of the oxidized and reduced forms of the mutant lysozymes. The green bars for the multiple-bridged proteins correspond to the sum of the differences in Tm values for the constituent single-bridged lysozymes. (Adapted from M. Matsumura et al., *Nature* 342: 291–293, 1989.)

residues of the mutants with single disulfide bonds, the larger was the effect on stability. Furthermore, the effects were additive so that the increase in Tm of 23 °C for the mutant with three disulfide bonds was approximately equal to that of the sum of the increases in Tm values for the three mutants with single disulfide bonds (4.8 °C + 6.4 °C + 11.0 °C ⊕22 °C). The effect on the stability of the protein from reducing the number of possible unfolded structures through introduction of disulfide bridges, the entropic effect, is even larger than these values show because the reduced forms of the mutants had a lower Tm than wild type, which indicates that favorable contacts in the folded structure had been lost by the mutations. These experiments show that engineered disulfide bridges can be combined together to enhance stability dramatically. Needless to say, knowledge of the three-dimensional structure of the protein is a prerequisite to engineer increased stability in this way.

Glycine and proline have opposite effects on stability

Glycine residues have more conformational freedom than any other amino acid, as discussed in Chapter 1. A glycine residue at a specific position in a protein has usually only one conformation in a folded structure but can have many different conformations in different unfolded structures of the same protein and thereby contribute to the diversity of unfolded conformations. Proline residues, on the other hand, have less conformational freedom in unfolded structures than any other residue since the proline side chain is fixed by an extra covalent bond to the main chain. Another way to decrease the number of possible unfolded structures of a protein, and hence stabilize the native structure, is, therefore, to mutate glycine residues to any other residue and to increase the number of proline residues. Such mutations can only be made at positions that neither change the conformation of the main chain in the folded structure nor introduce unfavorable, or cause the loss of favorable, contacts with neighboring side chains.

Both types of mutations have been made in T4 lysozyme. The chosen mutations were Gly 77–Ala, which caused an increase in Tm of 1 °C, and Ala 82–Pro, which increased Tm by 2 °C. The three-dimensional structures of these mutant enzymes were also determined: the Ala 82–Pro mutant had a structure essentially identical to the wild type except for the side chain of residue 82; this strongly indicates that the effect on Tm of Ala 82–Pro is indeed due to entropy changes. Such effects are expected to be additive, so even though each mutation makes only a small contribution to increased stability, the combined effect of a number of such mutations should significantly increase a protein's stability.

Stabilizing the dipoles of α helices increases stability

In Chapter 2 we described the α helix as a dipole with a positive charge at its N-terminus and a negative charge at the C-terminus. Negative ions, such as phosphate groups in coenzymes or substrates, are usually bound to the positive ends of such helical dipoles. The α helices that are not part of a binding site frequently have a negatively charged side chain at the N-terminus or a positively charged residue at the C-terminus that interacts with the dipole of the helix. Such dipole-compensating residues stabilize the helical forms of small synthetic peptides in solution. Do these helix-stabilizing residues also contribute to the overall stability of globular proteins? Of the 11 α helices of T4 lysozyme, 7 helices have negatively charged residues close to their N-termini; two of the remaining four α helices were therefore chosen for engineering studies to answer this question (Figure 17.5).

Two different mutant proteins with single substitutions at the N-terminus of each of these helices, Ser 38–Asp and Asn 144–Asp, were made as well as the corresponding double mutant. The single mutants both showed an increase in Tm of about 2 °C; the effects are additive since the double mutant had a Tm about 4 °C higher than wild type. This corresponds to 1.6 kcal/mol of stabilization energy. From the x-ray structures of these mutants it is apparent that the stabilization is due to electrostatic interactions and not to specific hydrogen bonding between the substituted amino acid and the end of the helix. Alan Fersht in Cambridge, UK has shown, using a different system, the small bacterial ribonuclease, barnase, that a histidine residue at

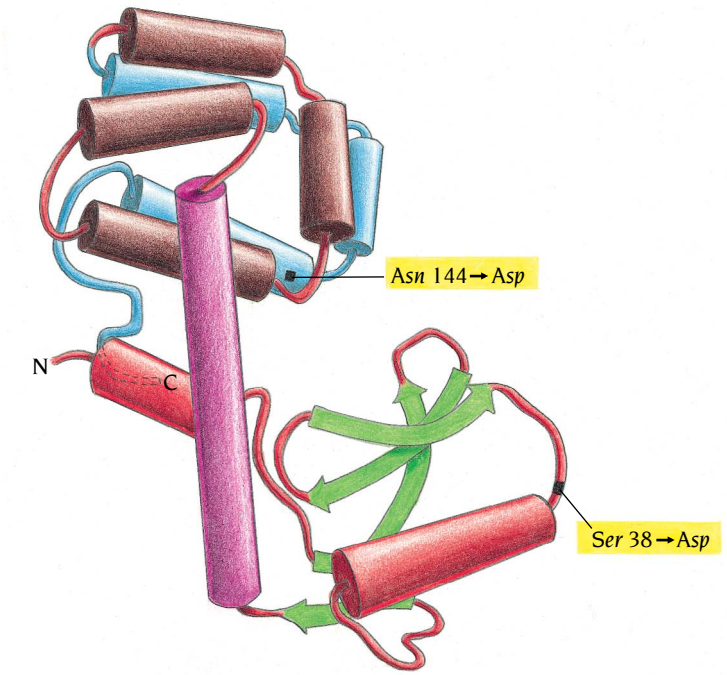


Figure 17.5 Diagram of the T4 lysozyme structure showing the locations of two mutations that stabilize the protein structure by providing electrostatic interactions with the dipoles of α helices. (Adapted from H. Nicholson et al., *Nature* 336: 651–656, 1988.)

the C-terminus of a helix stabilizes the barnase structure by about 2.1 kcal/mol. Significant stabilization of α-helical structures might, therefore, be obtained by combining several such helix-stabilizing mutations.

Mutants that fill cavities in hydrophobic cores do not stabilize T4 lysozyme

We emphasized in Chapter 2 that burying the hydrophobic side chains in the interior of the molecule, thereby shielding them from contact with solvent, is a major determinant in the folding of proteins. The surface that is buried inside a folded protein contributes directly to the stabilization energy of the molecule. Studies of destabilizing mutants in barnase, where **cavities** have been engineered into the hydrophobic core of the wild-type enzyme by mutations such as Ile to Val or Phe to Leu show that the introduction of a cavity the size of one -CH_2- group destabilizes the enzyme by about 1 kcal/mol. By analogy it should be possible to stabilize a wild-type protein by making mutations that fill existing cavities in its hydrophobic core. Even though proteins have the atoms of their hydrophobic cores packed approximately as tight as atoms are packed in crystals of simple organic molecules, there are cavities in the cores of almost every protein.

T4 lysozyme has two such cavities in the hydrophobic core of its α helical domain. From a careful analysis of the side chains that form the walls of the cavities and from building models of different possible mutations, it was found that the best mutations to make would be Leu 133–Phe for one cavity and Ala 129–Val for the other. These specific mutants were chosen because the new side chains were hydrophobic and large enough to fill the cavities without making too close contacts with surrounding atoms.

The two single mutants were constructed, purified, analyzed for stability, and crystallized. They were both less stable than wild type by 0.5 to 1.0 kcal/mol. The x-ray structures of the mutants provide a rational explanation for this disappointing result. It turns out that in order to fill the cavities, the new side chains in the mutants adopt energetically unfavorable conformations. This introduces strain in the structure, which obviously costs more energy than is gained by the new hydrophobic interactions. Even careful model building is obviously not sufficient to predict detailed structural and energetic effects of mutations in the hydrophobic core of proteins. Apparently, the observed core structure in T4 lysozyme, and probably in most proteins, reflects a compromise between the hydrophobic effect, which will tend to maximize the core-packing density, and the strain energy that would be incurred in eliminating all packing defects. Therefore, mutations designed to fill existing cavities may be effective in some cases, but they are not likely to provide a general route to substantial improvement in protein stability.

Proteins can be engineered by combinatorial methods

The ultimate goal of protein engineering is to design proteins to carry out predicted functions. However, we do not yet completely understand the rules governing protein folding and molecular recognition, making design of proteins difficult. Protein engineers have therefore invented **combinatorial methods**, in which **libraries** of related proteins are analyzed simultaneously. By sorting these libraries to select for a particular function, the small number of active proteins can be separated from millions of inactive variants. Combinatorial libraries have been used to increase the activity of enzymes, to improve the binding affinity and specificity of proteins, and even to identify novel peptide ligands. Additionally, researchers hope to use the structural and functional data obtained through library selection to improve their ability precisely to engineer molecular interactions.

Combinatorial methods are often referred to as *in vitro* or directed evolution techniques. In nature, the random DNA mutations that lead to changes in protein sequences occur rarely and so evolution is usually a slow

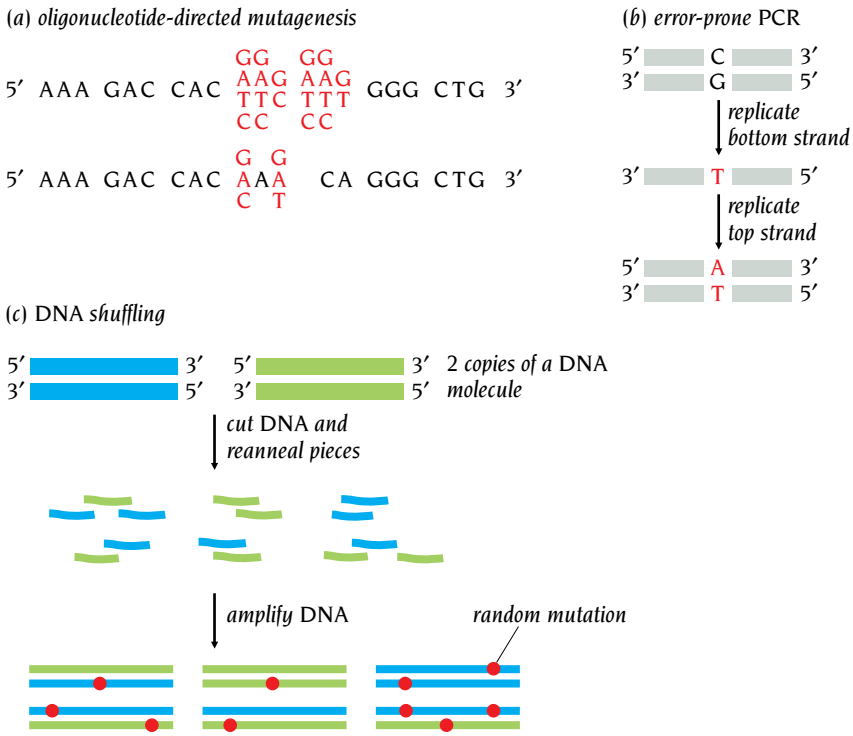


Figure 17.6 Methods of random mutagenesis. Several techniques are available for generating DNA libraries. (a) Oligonucleotides (short molecules of DNA) can be synthesized to contain mixtures of nucleotides at specific codons. An NNS or NNK codon, containing all bases at the first two positions and only two bases at the third position, each allows 20 possible amino acids. Alternatively, a restricted set of bases gives a more focused library; the lower example shown would permit only hydrophilic amino acids (E, K, Q, D, N, H). Oligonucleotide libraries are then incorporated into the gene of interest. (b) Error-prone polymerase chain reaction (PCR) uses a DNA polymerase to replicate the target gene. Conditions are chosen to decrease the fidelity of replication, leading to single base pair errors throughout the gene. (c) In DNA shuffling, a gene is first cut into pieces and then regenerated using a DNA polymerase. The polymerase introduces mutations similar to error-prone PCR; additionally, the pieces of DNA get mixed, so that mutations from separate copies of the gene can be combined.

process. Combinatorial methods accelerate evolution by controlling both the level and location of genetic mutation. Usually, a large number of mutations are concentrated in a single gene through **random mutagenesis**. Because mutagenesis techniques differ in the number and dispersion of mutations introduced to a gene, the appropriate method of mutagenesis depends on what questions the protein engineer seeks to address; the most widely used mutagenesis strategies are outlined in Figure 17.6. The mutated genes are then selected *in vivo*, by conferring a function to cells, or *in vitro*, by binding to an immobilized target. The most common method for *in vitro* selection, **phage display**, is discussed in the following section. The optimal strategies for generating and sorting a library depend on the affinity and specificity of the library for the target. The following examples will therefore illustrate some important combinatorial methods as well as the information that has been gained by using these techniques.

Phage display links the protein library to DNA

In designing a selection strategy, we must consider how to isolate and characterize the functional proteins in a library. Classically, molecular biologists have screened for altered protein function *in vivo*, by measuring effects on whole cells. There are, however, several possible advantages to using *in vitro* selections. For example, we no longer require a selection which modifies the growth of a host organism and can instead focus on an isolated function, such as a ligand binding to its receptor. However, *in vitro* selection does require that we connect each member of the protein library to its gene so that we can readily amplify and identify selectants. Bacteriophage (phage) display provides a simple mechanism to link the protein to its DNA.

Phage display typically utilizes bacteriophage M13. This filamentous phage contains single-stranded DNA encased in a protein coat. In contrast to the spherical viruses discussed in Chapter 16, M13 is long (1–2 μm) and narrow (7 nm) and contains five coat proteins, including approximately 2700 copies of the major coat protein gVIIIp (gene VIII protein) and five copies of the infectivity protein gIIIp. In phage display, the gene encoding the peptide

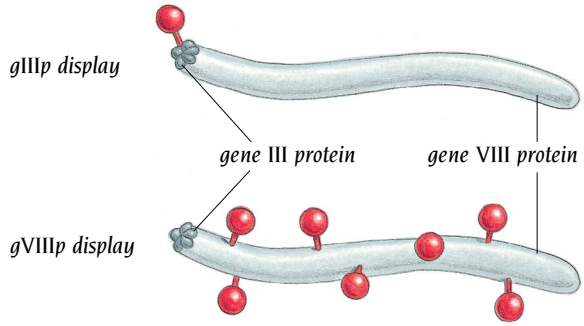


Figure 17.7 Proteins displayed on filamentous phage. In phage display, proteins are usually fused to the major coat protein, gVIIIp (2700 copies per phage), or to the infectivity protein, gIIIp (5 copies per phage). During assembly of the virus in bacteria, capsid fusion proteins are incorporated into the virus and are displayed on the surface of the phage. When gVIIIp fusions are produced, there can be many copies of protein per phage particle, leading to multivalent display. By contrast, gIIIp fusions typically give only one copy per phage (monovalent display).

or protein of interest is usually fused to one of these genes (Figure 17.7). When phage particles are produced in bacterial cells, the capsid fusion protein is incorporated into the viral particle and the phage DNA, containing the gene fusion, is packaged into the phage. The protein phenotype on the phage surface is thereby linked to the DNA genotype within the virus. Since progeny phage will not usually assemble from only capsid fusion proteins, wild-type capsid proteins are also expressed in the bacteria using a so-called helper phage that specifies wild-type capsid proteins but which is deficient in phage packaging. Several copies of the gVIIIp fusion protein molecules can be incorporated into each phage, giving multivalent display, but usually only one gIIIp fusion protein molecule is incorporated, giving monovalent display. A phage-display library results from each bacterium producing phage with a different capsid fusion protein. The typical phage display experiment includes three steps, shown in Figure 17.8. First, a library containing around 10^8 phage is screened for binding to an immobilized target. The selected phage are then propagated in bacteria and the phage DNA is characterized to determine the sequence of the gene corresponding to the mutated binding

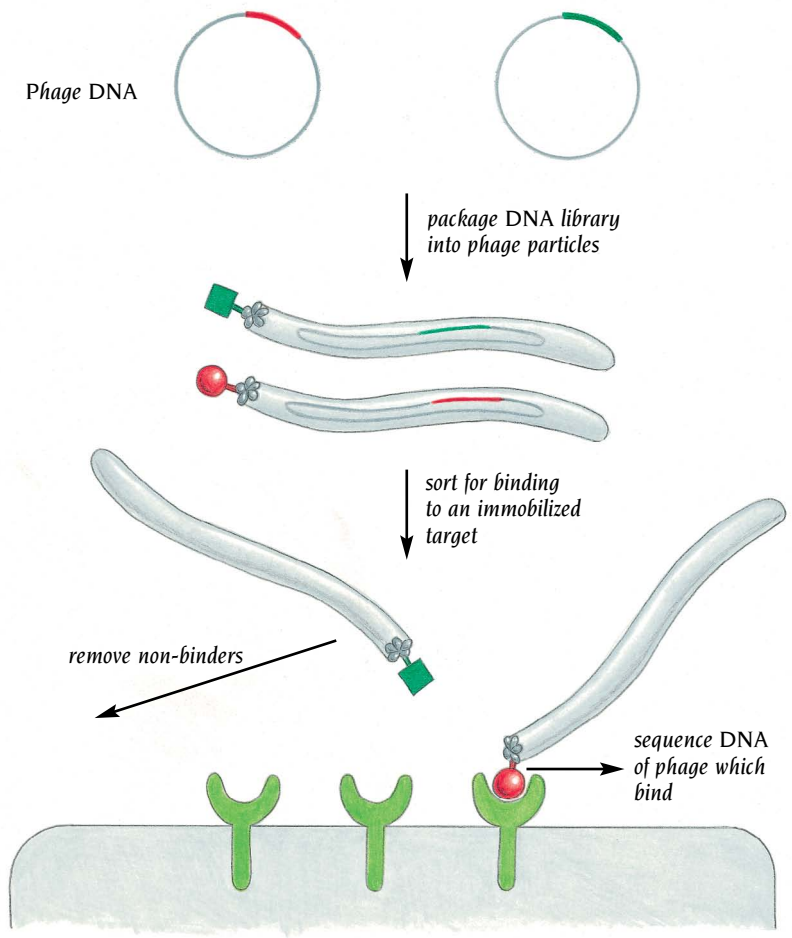


Figure 17.8 Sorting phage display libraries. Each phage particle in a library contains one protein–phage fusion and its corresponding DNA code. This phage library is added to an immobilized target protein; if a fusion protein does not bind the target, the phage displaying that protein is washed away. If a fusion protein does bind, its phage is eluted from the target, propagated in bacteria, and resorted under more stringent binding conditions. After several rounds of this sorting procedure, the DNA from phage are sequenced in order to determine the amino acid sequence of the selected protein.

protein. These three steps can then be repeated after mutagenesis of the selected genes to improve further the properties of the selected proteins.

In 1985, George Smith at the University of Missouri first demonstrated that peptide–phage fusions could be selected through the binding of the peptide to an antibody immobilized on a plate. Since that time, phage display has been used to improve the affinity and specificity of both antibodies and antigens, hormones and receptors; researchers have also studied the interactions of proteins with small molecules or nucleic acids. In the sections below, we focus on the use of phage display to characterize the interactions between proteins, including examples in which a protein scaffold is mutated to change the ligand specificity, a truncated protein is mutated to construct a minimized binding domain, and a random peptide library is sorted to identify a novel receptor agonist.

Affinity and specificity of proteinase inhibitors can be optimized by phage display

The blood coagulation cascade involves several trypsin-like serine proteinases (see Chapter 11) including plasmin, kallikrein, factor XIa, and the tissue factor–factor VIIa (TF–FVIIa) complex. While there would be important clinical benefits to engineering specific protease inhibitors, the active sites of these enzymes are highly conserved, showing as high as 81% identity, making the design of specific inhibitors difficult. **Kunitz domains** make up one family of protein inhibitors of the trypsin-like proteinases (Figure 17.9). These protein domains of approximately 60 residues, stabilized by three disulfide bonds, maintain a highly conserved structure with a sequence identity as low as 33%. Each Kunitz domain recognizes one or more proteinases through a set of 10–14 residues, most of which are in the "binding loop" (residues 11–19, green in Figure 17.9).

As described in Chapter 11, the active site cleft of a serine protease forms a row of subsite pockets, named S5 through S4′, which fit the substrate residues, numbered P5 through P4′. Table 17.1 gives the correlation between subsite and residue number for APPI (Alzheimer's amyloid β-protein precursor inhibitor) and LACI-D1 (lipoprotein-associated coagulation inhibitor D1) Kunitz domains. The major binding determinant of inhibitors for all trypsin-like proteinases is the presence of Lys 15 or Arg 15 at the P1 site; the specificity of Kunitz domains for different enzymes is then determined by the other subsite residues. Phage display experiments have identified specific Kunitz domain inhibitors with mutations in the subsite residues of the primary binding loop. From the sequences of these specific inhibitors, we can learn the rules governing the recognition between Kunitz domains and serine proteinases.

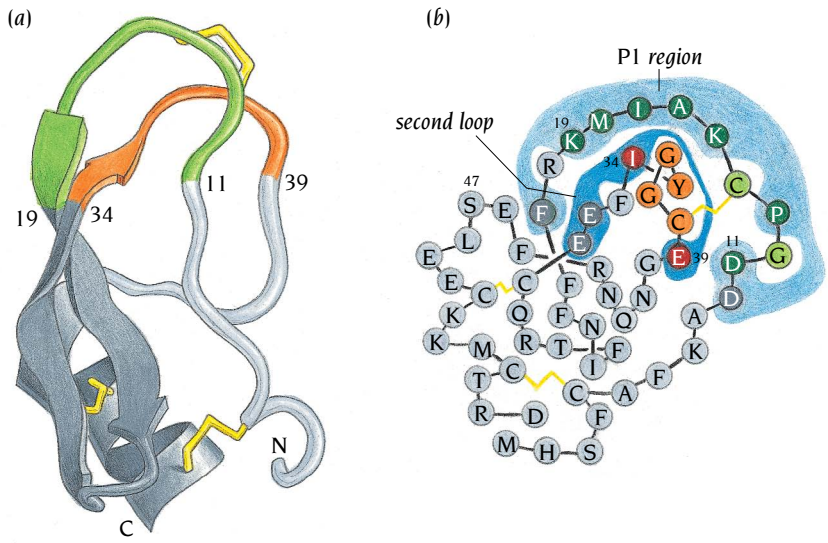


Figure 17.9 Structure and protease-binding properties of Kunitz domains. (a) Structure of APPI determined by x-ray crystallography. This 58 residue protein domain is characteristic of Kunitz domains. The three disulfide bonds are colored yellow. Residues 11–19, colored green, constitute the primary loop involved in binding to trypsin-like serine proteinases. Residues 34–39, colored orange, are also thought to be involved in binding to the proteinase or structuring the binding loop. (b) Amino acid sequence of Kunitz domain LACI-D1. Residues in Kunitz domains making the principal contacts with trypsin-like proteinases are shown in dark colors, colored according to (a). While most important proteinase interactions are with the primary binding loop (11–19), the second loop (34–39) contains some residues which contact either the proteinase or the primary binding loop. [(b) Adapted from W. Markland et al., *Biochemistry* 35: 8045–8057, 1996.]

Table 17.1 Phage-optimized sequences of Kunitz domain libraries

wt sequence and target	*primary binding loop*								*secondary loop*		K_i *for target* (M)
	P_5	P_4	P_3	P_1	P_1'	P_2'	P_3'	P_4'			
residue number	11	12	13	15	16	17	18	19	34	39	
LACI-D1	D	G	P	K	A	I	M	K	I	E	
kallikrein	D	G	P	R	A	A	H	P	S	G	40×10^{-12}
APPI	T	G	P	R	A	M	I	S	F	G	
kallikrein	D	G	H	R	A	A	H	P	Y	G	15×10^{-12}
TF-FVII*a* + *kallikrein*	P	G	P	R	A	L	I	L	F	Y	$\sim 2 \times 10^{-9}$
TF-VII*a*	P	G	P	K/R	A	L/M	M	K	I	Y/H	$\sim 10 \times 10^{-9}$

The sequences of LACI-D1 and APPI Kunitz domain protease-binding regions are shown, with the sequences of phage-optimized kallikrein and TF-FVIIa-binding variants given below. Variants of both LACI-D1 and APPI were selected for binding to kallikrein, and additionally APPI variants that bound TF-FVIIa and kallikrein were further selected for TF-VIIa preferential binding.

Two groups have selected phage-displayed Kunitz domains for binding to kallikrein. Mark Dennis and Robert Lazarus at Genentech, US, used the human Kunitz domain APPI as a scaffold. They made three libraries, each containing four or five randomized codons, and combined the selected mutations from all libraries to form a consensus sequence. The consensus sequence, given in Table 17.1, showed an inhibition constant (K_i) for kallikrein of 15 pM, which was lower than the K_i of any of the individual libraries. This result is consistent with the additivity principle which states that the effects of noninteracting mutations tend to be independent. Robert Ladner and coworkers at Protein Engineering Corporation, US, used a different human Kunitz domain, LACI-D1 as a scaffold for their kallikrein-binding libraries. These researchers designed DNA libraries using a restricted set of codons (see Figure 17.6) based on the residues commonly found in Kunitz domains at each position. This strategy reduces diversity, thus permitting more residues to be randomized in each library. However, important interactions might be missed when only a subset of amino acids are available. Nevertheless, the phage selectants identified by this method had very similar sequences and K_i values (40 pM) to the proteins selected at Genentech. The fact that these two phage display libraries arrived at very similar sequences, having started from different scaffolds and library designs, can be described as convergent *in vitro* evolution.

Dennis and Lazarus also sorted their APPI libraries against TF–FVIIa. They found that the tightest inhibitors, (K_i approximately 2 nM), also inhibited factor XIa and kallikrein. To identify inhibitors specific for TF–FVIIa, they employed a competitive, or subtractive, sorting strategy in which soluble competitor (factor XIa) was added during sorting. Kunitz domains that had a high affinity for soluble factor XIa thus remained in solution and were not selected for binding to immobilized TF–FVIIa. Selectants from this competitive sorting maintained nanomolar affinity for TF–FVIIa but inhibited both factor XIa and kallikrein ~1000-fold more weakly. Subtractive sorting thus shifted *in vitro* selection towards mutations that were tolerated only by the desired target.

How do the mutations identified by phage display improve binding specificity? There is as yet no direct structural information on the phage-selected inhibitors; however they can be modeled using data from the crystal structures of other Kunitz domains bound to serine proteinases. These studies lead to the conclusion that the mutations identified by phage display improve binding specificity by maximizing complementarity between the

primary binding loop and the proteinase active site. Analysis of the models of these new Kunitz domains with altered specificity has given novel insights into the mechanisms of Kunitz domain–protease specificity.

Structural scaffolds can be reduced in size while function is retained

In most proteins, only a small number of residues are directly involved in ligand or receptor binding; the rest of the protein provides the three-dimensional structure to position correctly the functional parts of the molecule. James Wells and colleagues at Genentech have used phage display to ask whether a binding epitope can be transferred to a smaller scaffold. This is an interesting question because protein scaffolds often seem larger than they need to be, and also the production of useful proteins is most efficient if the proteins are small. One such minimization involved the Z domain of bacterial protein A, shown in Figure 17.10. The 59-residue Z domain forms an antiparallel three-helix bundle that binds to the Fc portion of IgG (see Chapter 15) with a dissociation constant (K_d) of 20 nM. The Fc-binding epitope is discontinuous, involving residues from both helix 1 and helix 2. The third helix does not contact Fc, but is required to maintain the structure of the binding domain. Minimizing the Z domain thus poses a difficult design problem: a smaller version must maintain the correct three-dimensional placement of functional groups that are not adjacent in sequence. Furthermore, protein structures smaller than 50 residues are relatively rare, and only recently have 30 residue peptides been designed as stable domains (see below).

Andrew Braisted and J.A. Wells prepared phage containing Z domain helices 1 and 2 and restored Fc binding of this 38 residue minidomain in three iterative stages (see Figure 17.10). The truncated peptide was first randomized at four hydrophobic residues which contact helix 3 in the complete Z domain. The consensus sequence from this library maintained the wild-type residues Ile 17 and Leu 23 while the hydrophobic residues Leu 20 and

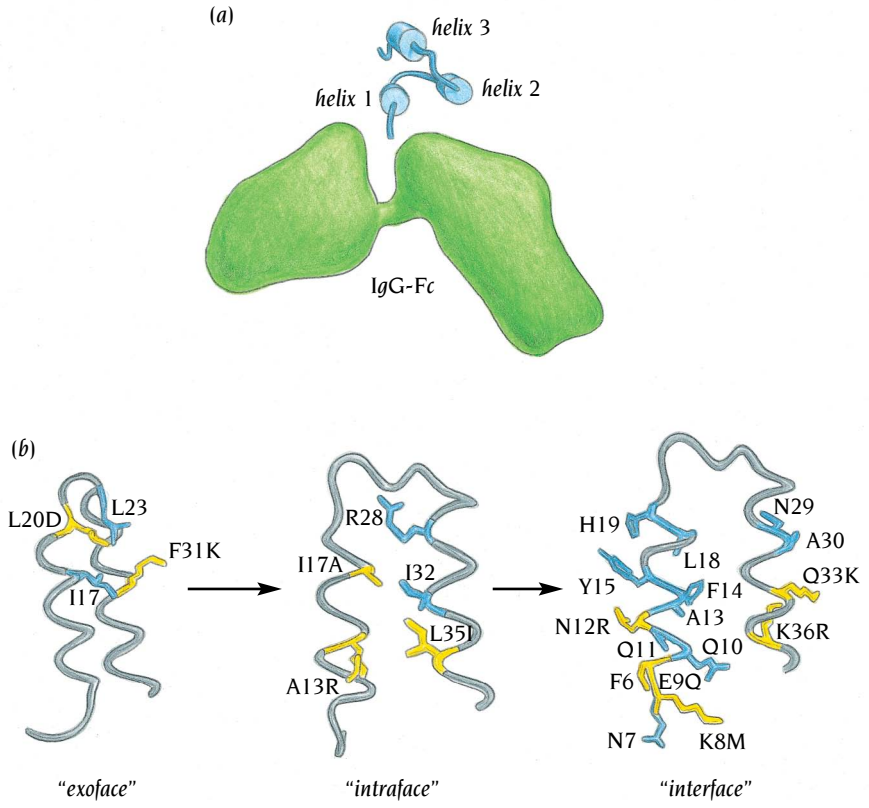


Figure 17.10 Construction of a two helix truncated Z domain. (a) Diagram of the three-helix bundle Z domain of protein A (blue) bound to the Fc fragment of IgG (green). The third helix stabilizes the two Fc-binding helices. (b) Three phage-display libraries of the truncated Z-domain peptide were selected for binding to the Fc. First, four residues at the former helix 3 interface ("exoface") were sorted; the consensus sequence from this library was used as the template for an "intraface" library, in which residues between helices 1 and 2 were randomized. The most active sequence from this library was used as a template for five libraries in which residues on the Fc-binding face ("interface") were randomized. Colored residues were randomized; blue residues were conserved as the wild-type amino acid while yellow residues reached a nonwild-type consensus. [(b) Adapted from A.C. Braisted and J.A. Wells, *Proc. Natl. Acad. Sci. USA* 93: 5688–5692, 1996.]

Phe 31 were mutated to the charged residues Asp and Lys, respectively. This double mutant bound Fc with a K_d of 3.4 μM, a greater than a hundredfold improvement over the unmutated fragment, and had a large increase in α-helical content as shown by circular dichroism spectroscopy. This face of the two-helix peptide was thus converted from a protein core to a protein surface. The second library fixed Asp 20 and Lys 31 and randomized five positions at the "intraface" of helix 1 and helix 2. From this library, three new residues were selected at the open end of the intrahelix interface, and the K_d for this peptide was around 300 nM, a tenfold improvement over the first library. Finally, five libraries were randomized at the Fc-binding face of the two-helix bundle and the consensus mutations were combined. Some libraries yielded mutations that improved binding 2–3 fold, and all together the two-helix bundle contained 12 mutations from the Z domain sequence. A final truncation of the five N-terminal residues yielded a 33 residue peptide with a K_d of 40 nM, very close to the wild-type value of 20 nM. X-ray crystallography and NMR spectroscopy indicated that the two helix bundle had the same three-dimensional structure and the same Fc-binding epitope as the Z domain. Thus, iterative cycles of phage display were used to create a more stable protein surface and to repack the protein core. The resulting scaffold was half the size of the native domain but maintained the three-dimensional arrangement of Fc-binding residues at the interface.

Phage display of random peptide libraries identified agonists of erythropoietin receptor

Many biological processes are activated by hormone–receptor interactions, and a great deal of research has been devoted to identifying peptides or small molecules which either inhibit or simulate hormone function. The idea that a small molecule can mimic the function of a large protein is based on the "hot-spot" principle that a small number of residues at a binding interface contribute most of the binding energy. Nicholas Wrighton and coworkers at Affymax, US, in collaboration with scientists at Johnson & Johnson and at the Scripps Institute, have used phage display methods to isolate a 20 residue peptide, called EMP1, which mimics the activity of erythropoietin (EPO) by promoting dimerization and activation of erythropoietin receptor (EPOR; the extracellular domain of which is called EBP). Erythropoietin is a cytokine hormone which stimulates formation of red blood cells (see Chapter 13).

EMP1 was selected by two cycles of phage display. First, random peptide libraries of the sequence CX_8C, where X is any residue and the cysteines form an intramolecular disulfide bond, were displayed as gVIIIp fusions (see Figure 17.7). Multivalent gVIII protein fusion display permitted selection of weak binders by avidity (multiple) binding to immobilized dimers of EBP. These weakly binding peptides were expanded to the form $X_5CX_8CX_3$ and partially randomized in the X_8 residues. Fusion of the new library to gIIIp yielded a lower valency of display and allowed selection of tight binders. EMP1, isolated from this library, bound to EBP with a K_d of 0.2 μM and stimulated EPOR activity *in vivo*. The dimerization of EMP1, critical for its agonist activity, was probably selected through the interactions of the multivalent peptides with the immobilized EBP dimer.

The crystal structure of EMP1 bound to EBP shows the remarkable structural economy of the EMP1 dimer (Figure 17.11). Each peptide monomer forms a β hairpin, stabilized by an intramolecular disulfide bond. The peptide dimer forms a four-stranded β sheet maintained by four main-chain hydrogen bonds and by the packing of hydrophobic residues. Each monomer makes hydrogen bonds and hydrophobic contacts to both EBP receptors, forming a total of 20 interactions between the peptide dimer and the two EBP proteins; most of the peptide is directly involved in binding (Figure 17.12). Furthermore, EMP1 seems to mimic the binding interactions used by EPO itself. Although there is no structural information on the EPO–EBP complex, the structure of another cytokine hormone, human growth hormone, bound

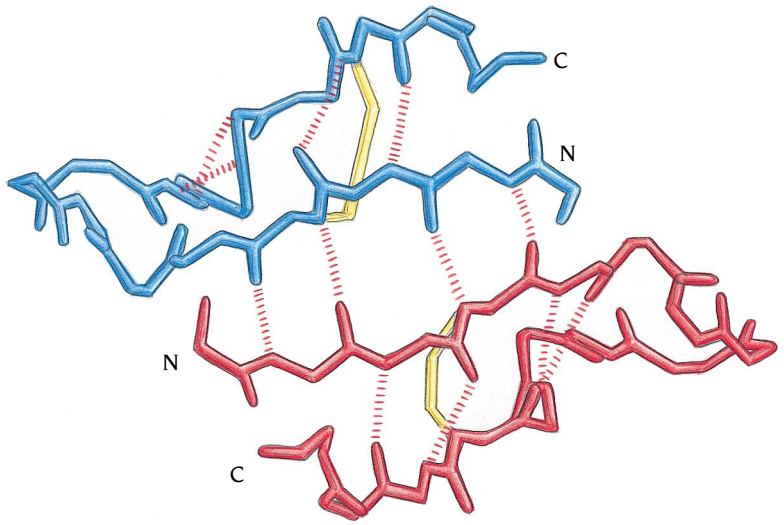


Figure 17.11 Structure of EMP1 dimer from x-ray crystallography. In the presence of EBP, the EMP1 peptide forms a dimer. Each monomer (shown in red and blue) forms a β hairpin structure stabilized by hydrogen bonds (red dashes) and a disulfide bond (yellow). The two peptides form a symmetrical dimer stabilized by four hydrogen bonds (red dashes) and hydrophobic contacts. The two monomers form a four-stranded, anti-parallel pleated sheet.

to its receptor (see Chapter 13) was used as a model. Like EPO, EMP1 contacts four of the six loops on EBP that have sequence similarities to loops of the growth hormone receptor involved in hormone binding. In particular, the three residues of the growth hormone receptor most critical for binding of growth hormone, which correspond to residues Phe 93, Phe 205, and Met 150 in EBP, are well-buried in the EMP1–EBP crystal structure.

EMP1, selected by phage display from random peptide libraries, demonstrates that a dimer of a 20-residue peptide can mimic the function of a monomeric 166-residue protein. In contrast to the minimized Z domain, this selected peptide shares neither the sequence nor the structure of the natural hormone. Thus, there can be a number of ways to solve a molecular recognition problem, and combinatorial methods such as phage display allow us to sort through a multitude of structural scaffolds to discover novel solutions.

DNA shuffling allows accelerated evolution of genes

Natural selection works through the complementary processes of mutation and genetic reassortment by **recombination**. The oligonucleotide-directed mutagenesis methods used in the foregoing examples do not allow for recombination; instead, mutations are combined manually to optimize a protein sequence. Willem Stemmer at Maxygen invented a method of directed evolution that uses both mutation and recombination. This method, called

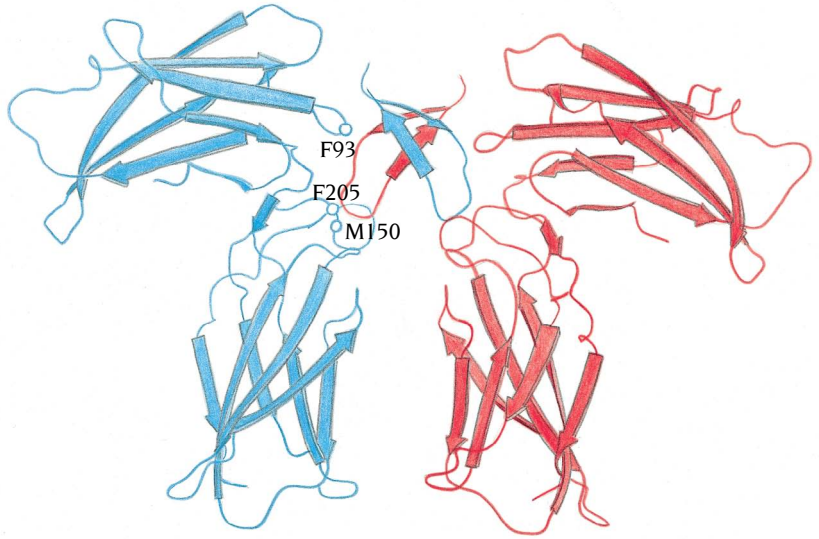
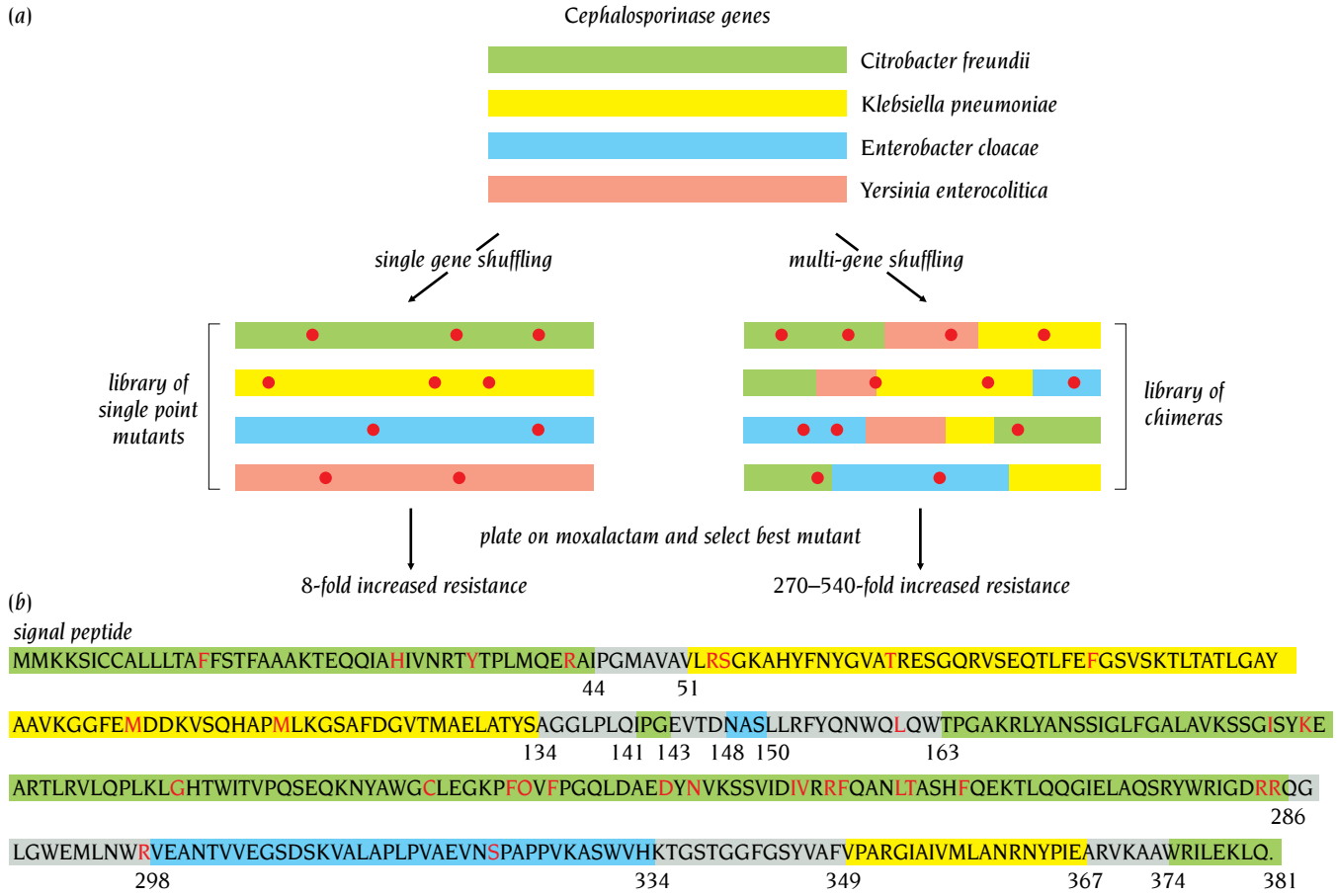


Figure 17.12 Ribbon diagram of EMP1 bound to the extracellular domain of the erythropoietin receptor (EBP). Binding of EMP1 causes dimerization of erythropoietin receptor. The x-ray crystal structure of the EMP1–EBP complex shows a nearly symmetrical dimer complex in which both peptide monomers interact with both copies of EBP. Recognition between the EMP1 peptides and EBP utilizes more than 60% of the EMP1 surface and four of six loops in the erythropoietin-binding pocket of EBP. In particular, three residues thought to be critical for binding erythropoietin (Phe 93, Met 150, and Phe 205) are fully buried in the structure of the peptide–receptor complex. (From J.A. Wells, *Science* 273: 449–450, 1996.)



DNA shuffling (see Figure 17.6), begins by randomly cutting a gene into fragments about 100–300 base pairs in length. This DNA pool is then reassembled with a DNA polymerase, which combines fragments from different copies of the original gene. This process also incorporates point mutations at a defined rate. When point mutations originally in different copies of a gene end up in the same piece of DNA, *in vitro* recombination has occurred. These shuffled DNA libraries are then sorted *in vitro* by phage display or selected *in vivo* through bacterial selection, as described below.

Stemmer and coworkers have extended DNA shuffling to include homologous genes from different organisms. In their first attempt, they mixed the genes encoding class C cephalosporinases from four species (the DNA being 58–82% identical). Their goals were to select for bacterial resistance to the antibiotic moxalactam and to compare the evolution of individual genes with a shuffled gene family. As shown in Figure 17.13, one cycle of evolution using all four genes resulted in recombination of DNA segments as well as incorporation of random point mutations. By contrast, one cycle of DNA shuffling of a single gene only introduced point mutations. The results of recombination of different genes was dramatic: the single genes yielded eightfold increases in antibiotic resistance, while the four genes together gave 270 to 540-fold improvements. The best clone contained eight discrete DNA segments from three of the four genes as well as 33 point mutations. Figure 17.14 shows a three-dimensional model based on the crystal structure of one of the native enzymes. The different colors identify the origin of each protein segment; interestingly, these segments form units of secondary structure. This example demonstrates the utility of recombination and mutation in directed evolution. Additionally, this work underscores the power of combinatorial techniques to construct enzymes whose design would be impossible given our current understanding of the factors governing protein structure and function.

Figure 17.13 DNA shuffling of cephalosporinase genes. (a) Four homologous genes from bacteria were subjected to one cycle of cleavage and recombination, resulting in mutant genes containing point mutations (shown in red circles) and large pieces of each wild-type sequence. These mutant genes were transformed into bacteria and screened for resistance to the antibiotic moxalactam; the most active mutants increased resistance by 540-fold over the wild-type genes. By contrast, DNA shuffling of each individual gene yielded only point mutations; the most active of these mutants only increased resistance to moxalactam eightfold. (b) Amino acid sequence of the most active mutant, colored as in (a) to show the origin of regions of the protein, with point mutations shown in red. The gray regions cannot be unambiguously assigned to one of the original cephalosporinase genes. (From A. Crameri et al., *Nature* 391: 288–291, 1998.)

Figure 17.14 Model of evolved mutant from cephalosphorinase shuffling. The sequence of the most active cephalosporinase mutant was modeled using the crystal structure of the class C cephalosporinase from *Enterobacter cloacae*. The mutant and wild-type proteins were 63% identical. This chimeric protein contained portions from three of the starting genes, including *Enterobacter* (blue), *Klebsiella* (yellow), and *Citrobacter* (green), as well as 33 point mutations (red). (Courtesy of A. Crameri.)

Protein structures can be designed from first principles

The ultimate goal of protein engineering is to design an amino acid sequence that will fold into a protein with a predetermined structure and function. Paradoxically, this goal may be easier to achieve than its inverse, the solution of the folding problem. It seems to be simpler to start with a three-dimensional structure and find one of the numerous amino acid sequences that will fold into that structure than to start from an amino acid sequence and predict its three-dimensional structure. We will illustrate this by the design of a stable zinc finger domain that does not require stabilization by zinc.

The classic zinc fingers, the DNA-binding properties of which are discussed in Chapter 10, are small compact domains of about 30 residues that fold into an antiparallel β hairpin followed by an α helix. All known classic zinc fingers have a zinc atom bound to two cysteines in the hairpin and two histidines in the helix, creating a sequence motif common to all zinc finger genes. In the absence of zinc the structure is unfolded.

Stephen Mayo and Bassil Dahiyat at the California Institute of Technology asked the question: Is it possible to design from first principles a sequence whose main chain obtains this zinc finger fold without a zinc atom to stabilize the structure? They chose the second zinc finger of Zif 268 (see Chapter 10) with 28 residues as their target fold and applied their recently developed computer algorithm to the problem. Briefly, this algorithm searches through a very large number of possible sequences using a fast selection procedure for those sequences that stabilize a given fold.

On the basis of the template fold, side chain positions are divided into three categories: core, surface, and boundary positions. Allowed residues at the core positions are Ala, Val, Leu, Ile, Phe, Tyr and Trp, and at the surface positions Ala, Ser, Thr, His, Asp, Asn, Glu, Gln, Lys and Arg. The combined

Table 17.2 Amino acid sequences of the second zinc finger of Zif 268 and the designed peptide FSD-1

	1									11										21								28
FSD-1	Q	Q	Y	T	A	K	I	K	G	R	T	F	R	N	E	K	E	L	R	D	F	I	E	K	F	K	G	R
Zif 268	K	P	F	Q	C	R	I	C	M	R	N	F	S	R	S	D	H	L	T	T	H	I	R	T	H	T	G	E

Residues in the hydrophobic core of FSD-1 are green.

core and surface sets (16 residues since Ala belongs to both) are allowed at the boundary positions. They are thus using a restricted set of residues where Pro, Cys, and Met are absent and the only hydrophobic residue allowed at the surface is Ala. Gly is used in special positions to minimize backbone strain due to the conformation of the template.

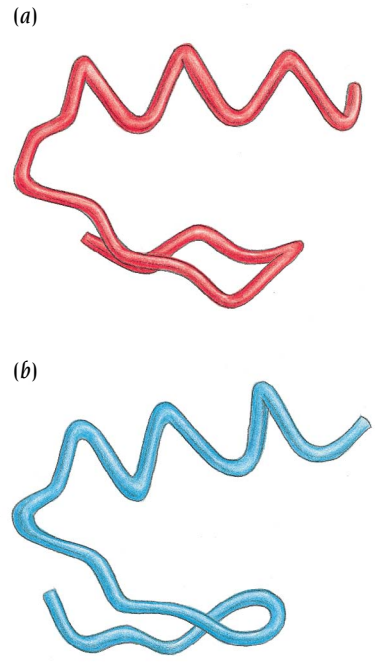
The selection algorithm used involves energy functions for van der Waals interactions, hydrogen bonding and solvation combined with a secondary structure propensity potential. The total number of amino acid sequences that must be screened is the product of the number of possible amino acids at each residue position. For Zif 268 there is one core position with seven possible amino acids, 20 surface positions each with 10 possible amino acids and seven boundary positions each with 16 possible amino acids, giving a total of about 10^{27} possible sequences. This virtual library size is 15 orders of magnitude larger than an experimental combinatorial library using current methods. A corresponding peptide library consisting of only a single molecule for each 28-residue sequence would have a mass of 11,600 kilograms. The actual virtual search space is, however, even larger since different conformations, rotamers, must be considered for every possible side chain.

The optimal sequence obtained, called FSD-1 for full sequence design, is shown in Table 17.2 and compared with the sequence of the template Zif 268. A search of the FSD-1 sequence against protein databases did not reveal a statistically significant similarity with any other protein, including zinc finger proteins.

In order to examine whether this sequence gave a fold similar to the template, the corresponding peptide was synthesized and its structure experimentally determined by NMR methods. The result is shown in Figure 17.15 and compared to the design target whose main chain conformation is identical to that of the Zif 268 template. The folds are remarkably similar even though there are some differences in the loop region between the two β strands. The core of the molecule, which comprises seven hydrophobic side chains, is well-ordered whereas the termini are disordered. The root mean square deviation of the main chain atoms are 2.0 Å for residues 3 to 26 and 1.0 Å for residues 8 to 26.

In addition to being a remarkable demonstration of the power of computer-based combinatorial design of a protein fold, this designed peptide is the shortest known peptide consisting entirely of naturally occurring amino acids that folds into a well-ordered structure without metal binding, oligomerization or disulfide bond formation.

Figure 17.15 Schematic diagrams of the main-chain conformations of the second zinc finger domain of Zif 268 (red) and the designed peptide FSD-1 (blue). The zinc finger domain is stabilized by a zinc atom whereas FSD-1 is stabilized by hydrophobic interactions between the β strands and the α helix. (Adapted from B.I. Dahiyat and S.L. Mayo, *Science* 278: 82–87, 1997.)



A β structure has been converted to an α structure by changing only half of the sequence

In 1994 a prize of $1000 called the Paracelsus challenge was offered to any designer of protein structures who could convert one protein fold into another while retaining 50% of the original sequence. The spirit of this challenge was to take one important step towards solving the folding problem by assessing the fraction of a protein's amino acid sequence that is sufficient to specify its structure. On the basis of our knowledge of naturally occurring proteins, this challenge did not look like a feasible project: all proteins of known structure with more than 30% sequence identity had been shown to have identical folds. Yet, it did not take more than three years before the prize

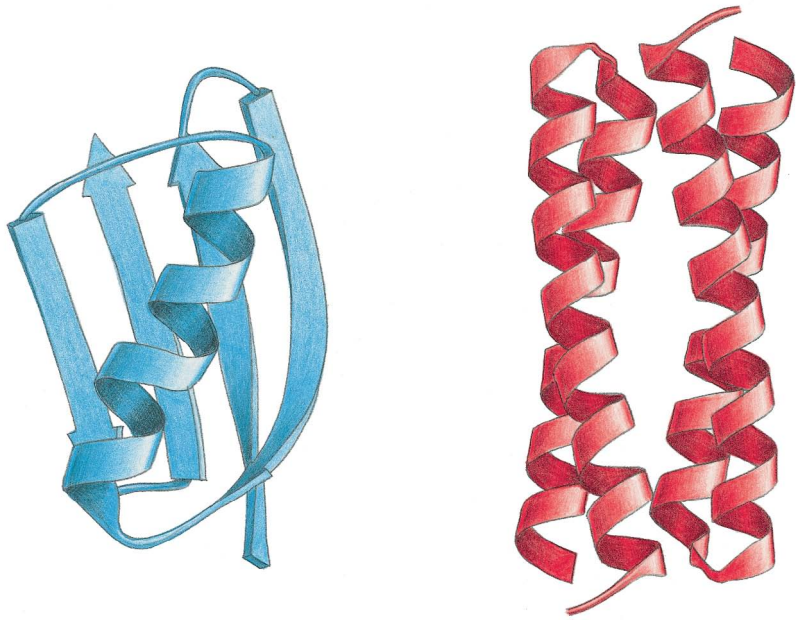


Figure 17.16 Ribbon diagram representations of the structures of domain B1 from protein G (blue) and the dimer of Rop (red). The fold of B1 has been converted to an α-helical protein like Rop by changing 50% of its amino acids sequence. (Adapted from S. Dalal et al., *Nature Struct. Biol.* 4: 548–552, 1997.)

was won. In 1997 the group of Lynne Regan at Yale University converted a protein that folded into a mainly β-sheet structure into an α-helical structure by changing 50% of its sequence.

They started from the sequence of a domain, B1, from an IgG-binding protein called Protein G. This domain of 56 amino acid residues folds into a four-stranded β sheet and one α helix (Figure 17.16). Their aim was to convert this structure into an all α-helical structure similar to that of Rop (see Chapter 3). Each subunit of Rop is 63 amino acids long and folds into two α helices connected by a short loop. The last seven residues are unstructured and were not considered in the design procedure. Two subunits of Rop form a four-helix bundle (Figure 17.16).

Aligning the two sequences from their amino ends produces only three identical residues in the 56 aligned positions (Table 17.3). In order to retain 50% of the original sequence, 28 residues could be changed in the B1 domain for the fold to switch to a Rop-like structure. The rationale for these changes was to use current knowledge about helix formation and stability to identify and change the subset of residues that are key determinants of the fold, considering both local and long-range interactions. Residues in the B1 domain

Table 17.3 Amino acid sequences of domain B1, the designed protein Janus, and Rop

	1										11										
B1	M	T	Y	K	L	I	L	N	G	K	T	L	K	G	E	T	T	T	E		
Janus	M	T	K	K	A	I	L	A	L	N	T	A	K	F	L	R	T	Q	A		
Rop	G	T	K	Q	E	K	T	A	L	N	M	A	R	F	I	R	S	Q	T		
		21												31							
B1	A	V	D	A	A	T	A	E	K	V	F	K	Q	Y	A	N	D	N	G		
Janus	A	V	L	A	A	K	L	E	K	L	G	A	Q	E	A	N	D	N	A		
Rop	L	T	L	L	E	K	L	N	E	L	D	A	D	E	Q	A	D	I	C		
		41												51							
B1	V	D	G	E	W	T	Y	D	D	A	T	K	T	F	T	V	T	E			
Janus	V	D	L	E	D	T	A	D	D	L	Y	K	T	L	L	V	L	A			
Rop	E	S	L	H	D	H	A	D	E	L	Y	R	S	C	L	A	R	F			

Red residues have been changed to the same type as present in Rop, blue residues have been changed to a different type, and black are unchanged.

with high preference for β sheet formation were replaced with residues with high preference for α helix formation, such as Tyr to Lys and Leu to Ala, in regions that were required to switch conformation. Since Rop is a coiled-coil protein, hydrophobic residues were incorporated in appropriate *a* and *d* positions of the heptad repeat (see Chapter 3). Among other changes was the introduction of an intramonomer salt bridge between Arg 16 and Asp 46 that is present between the two helices of Rop.

During this process of designing sequence changes, models were built and assessed to ensure that there were no obvious steric clashes and that the hydrophobic core was well packed. Furthermore, secondary structure prediction was also used to monitor the progress of change and to choose among different possible substitutions. The final sequence (see Table 17.3) contains 28 changes; it had 50% identity to B1 and the similarity to Rop had increased from 5.4% identity to 41%.

A gene encoding this sequence was synthesized and the corresponding protein, called Janus, was expressed, purified, and characterized. The atomic structure of this protein has not been determined at the time of writing but circular dichroic and NMR spectra show very clear differences from B1 and equally clear similarities to Rop. The protein is a dimer in solution like Rop and thermodynamic data indicate that it is a stably folded protein and not a molten globule fold like several other designed proteins.

These results indicate that is it possible to change the fold of a protein by changing a restricted set of residues. They also confirm the validity of the rules for stability of helical folds that have been obtained by analysis of experimentally determined protein structures. One obvious impliction of this work is that it might be possible, by just changing a few residues in Janus, to design a mutant that flip-flops between α helical and β sheet structures. Such a polypeptide would be a very interesting model system for prions and other amyloid proteins.

Conclusion

Homologous proteins have similar three-dimensional structures. They contain a core region, a scaffold of secondary structure elements, where the folds of the polypeptide chains are very similar. Loop regions that connect the building blocks of the scaffolds can vary considerably both in length and in structure. From a database of known immunoglobulin structures it has, nevertheless, been possible to predict successfully the conformation of hypervariable loop regions of antibodies of known amino acid sequence.

Methods for the prediction of the secondary structure of a set of homologous proteins can reach an accuracy of about 75%, most of the errors occur at the ends of α helices or β strands. The central regions of these secondary structure elements are often correctly predicted but the methods do not always correctly distinguish between α helices and β strands.

Prediction of tertiary structure from the amino acid sequence is the major unsolved problem in structural molecular biology. The inverse problem, to predict which amino acid sequences can have a given fold seem to be easier to solve. Significant progress has been made in recent years in threading methods, which assign a known fold to a given sequence by threading the sequence through all known folds.

Protein engineering is now routinely used to modify protein molecules either via site-directed mutagenesis or by combinatorial methods. Factors that are important for the stability of proteins have been studied, such as stabilization of α helices and reducing the number of conformations in the unfolded state. Combinatorial methods produce a large number of random mutants from which those with the desired properties are selected *in vitro* using phage display. Specific enzyme inhibitors, increased enzymatic activity and agonists of receptor molecules are examples of successful use of this method.

Small protein molecules with a predetermined fold can be designed *in silico* by energy calculations of all possible combinations of a restricted set of amino acid residues. A designed zinc finger fold that is stable in the absence of zinc showed no significant sequence similarity to any known protein sequence. Important progress has been made in assessing the fraction of a protein's amino acid sequence that is sufficient to specify its structure. A protein that folds into a mainly β-sheet structure was converted into an α-helical structure by changing only 50% of its sequence.

Selected readings

General

Alber, T. Mutational effects on protein stability. *Annu. Rev. Biochem.* 58: 765–798, 1989.

Barton, G.J. Protein secondary structure prediction. *Curr. Opin. Struct. Biol.* 5: 372–376, 1995.

Blundell, T.L., et al. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* 326: 347–352, 1987.

Bowie, J.U., Eisenberg, D. Inverted protein structure prediction. *Curr. Opin. Struct. Biol.* 3: 437–444, 1993.

DeGrado, W.F., Wasserman, Z.R., Lear, J.D. Protein design, a minimalist approach. *Science* 243: 622–628, 1989.

Fasman, G.D. Protein conformational prediction. *Trends Biochem. Sci.* 14: 295–299, 1989.

Fersht, A.R. Protein engineering. *Protein Eng.* 1: 7–16, 1986.

Fersht, A.R. The hydrogen bond in molecular recognition. *Trends Biochem. Sci.* 12: 301–304, 1987.

Finkelstein, A.V. Protein structure: what is possible to predict now? *Curr. Opin. Struct. Biol.* **7**: 60-71, 1997.

Jones, D.T., Thornton, J. Potential energy functions for threading. *Curr. Opin. Struct. Biol.* 6: 210–216, 1996.

Kay, B.K, Winter, J., McCafferty, J. eds. *Phage Display of Peptides and Proteins: a Laboratory Manual*. San Diego: Academic Press, 1996.

Moult, J., et al. A large-scale experiment to assess protein–structure prediction methods. *Proteins* 23: ii–iv, 1995.

Murzin, A., et al. Scop: a structural classification of proteins. Database for the investigation of sequences and structures. *J. Mol. Biol.* 247: 536–540, 1995.

Orengo, C.A., et al. CATH—a hierarchic classification of protein domain structures. *Structure* 5: 1093–1108, 1997.

Richardson, J.S., Richardson, D.C. The *de novo* design of protein structures. *Trends Biochem. Sci.* 14: 304–309, 1989.

Sippl, M.J. Knowledge-based potentials for proteins. *Curr. Opin. Struct. Biol.* 5: 229–235, 1995.

Sonnhammer, E.L., Eddy, S.R., Durbin, R. Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* 28: 405–420, 1997.

Thornton, J.M., Gardner, S.P. Protein motifs and data-base searching. *Trends Biochem. Sci.* 14: 300–304, 1989.

von Heijne, G. *Sequence Analysis in Molecular Biology: Treasure Trove or Trivial Pursuit*. San Diego: Academic Press, 1987.

Wells, J.A. Additivity of mutational effects in proteins. *Biochemistry* 29: 8509–8517, 1990.

Specific structures

Aszodi, A., Taylor, W.R. Homology modelling by distance geometry. *Fold. Des.* 1: 325–334, 1996.

Bowie, J.U. Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science* 247: 1306–1310, 1990.

Bowie, J.U., Luthy, R., Eisenberg, D. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* 253: 164–170, 1991.

Braisted, A.C., Wells, J.A. Minimizing a binding domain from protein A. *Proc. Natl. Acad. Sci. USA* 93: 5688–5692, 1996.

Bryant, S.H. Evaluation of threading specificity and accuracy. *Proteins* 26: 172–185, 1996.

Chothia, C., et al. Conformations of immunoglobulin hypervariable regions. *Nature* 342: 877–883, 1989.

Chothia, C., et al. The predicted structure of immunoglobulin D 1.3 and its comparison with the crystal structure. *Science* 233: 755–758, 1986.

Chothia, C., Lesk, A. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5: 823–826, 1986.

Chou, P.Y., Fasman, G.D. Prediction of protein conformation. *Biochemistry* 13: 222–245, 1974.

Cohen, C., Parry, D.A.D. α-helical coiled coils and bundles: how to design an α-helical protein. *Proteins: Struct. Funct. Gen.* 7: 1–15, 1990.

Crameri, A., Raillard, S.-A., Bermudez, E., Stemmer, W.P.C. DNA shuffling of a family of genes from diverse species accelerates directed evolution. *Nature* 391: 288–291, 1998.

Dahiyat, B.I., Mayo, S.L. De novo protein design: fully automated sequence selection. *Science* 278: 82–87, 1997.

Dalal, S., Balasubramanian, S., Regan, L. Protein alchemy: changing β sheet into α helix. *Nature Struct. Biol.* 4: 548–552, 1997.

DeGrado, W.F., Regan, L., Ho, S.P. The design of a four-helix bundle protein. *Cold Spring Harbor Symp. Quant. Biol.* 52: 521–526, 1987.

Dennis, M.S., Herzka, A., Lazarus, R.A. Potent and selective Kunitz domain inhibitors of plasma kallikrein designed by phage display. *J. Biol. Chem.* 270: 25411–25417, 1995.

Dennis, M.S., Lazarus, R.A. Kunitz domain inhibitors of tissue factor-factor VIIa. I. Potent inhibitors selected from libraries by phage display. *J. Biol. Chem.* 269: 22129–22136, 1994.

Dennis, M.S., Lazarus, R.A. Kunitz domain inhibitors of tissue factor-factor VIIa. II. Potent and specific inhibitors by competitive phage selection. *J. Biol. Chem.* 269: 22137–22144, 1994.

Faber, H.R., Matthews, B.W. A mutant T4 lysozyme displays five different crystal conformations. *Nature* 348: 263–266, 1990.

Fersht, A.R., et al. Hydrogen bonding and biological specificity analyzed by protein engineering. *Nature* 314: 235–238, 1985.

Garnier, J., Osguthorpe, D.J., Robson, B. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* 120: 97–120, 1978.

Gribskov, M., McLaschlan, A.D., Eisenberg, D.E. Profile analysis: detection of distantly related proteins. *Proc. Natl. Acad. Sci. USA* 84: 4355–4358, 1987.

Jones, D.T., Taylor, W.R., Thornton, J.M. A new approach to protein fold recognition. *Nature* 358: 86–89, 1992.

Karpusas, M., et al. Hydrophobic packing in T4 lysozyme probed by cavity-filling mutants. *Proc. Natl. Acad. Sci. USA* 86: 8237–8241, 1989.

Kellis, J.T., et al. Contribution of hydrophobic interactions to protein stability. *Nature* 333: 784–786, 1988.

Lim, V.I. Algorithms for prediction of α-helical and β-structural regions in globular proteins. *J. Mol. Biol.* 88: 873–894, 1974.

Livnah, O., Stura, E.A., Johnson, D.L., Middleton, S.A., Mulcahy, L.S., Wrighton, N.C., Dower, W.J., Jolliffe, L.K., Wilson, I.A. Functional mimicry of a protein hormone by a peptide agonist: the EPO receptor complex at 2.8 Å. *Science* 273: 464–471, 1996.

Matsumura, M., Signor, G., Matthews, B.W. Substantial increase of protein stability by multiple disulfide bonds. *Nature* 342: 291–293, 1989.

Matthews, B.W., Nicholson, H., Becktel, W.J. Enhanced protein thermostability from site-directed mutations that decrease the entropy of unfolding. *Proc. Natl. Acad. Sci. USA* 84: 6663–6667, 1987.

Nicholson, H., Becktel, W.J., Matthews, B.W. Enhanced protein thermostability from designed mutations that interact with α-helix dipoles. *Nature* 336: 651–656, 1988.

Regan, L., DeGrado, W.F. Characterization of a helical protein designed from first principles. *Science* 241: 976–978, 1988.

Rice, D.W., et al. A 3D-1D substitution matrix for protein fold recognition that includes predicted secondary structure of the sequence. *J. Mol. Biol.* 267: 1026–1038, 1997.

Rice, D.W., et al. Fold assignments for amino acid sequences of the CASP2 experiment. *Proteins* (Suppl. 1) 113–122, 1997.

Richardson, J.S., Richardson, D.C. Amino acid preferences for specific locations at the ends of α helices. *Science* 240: 1648–1652, 1988.

Rost, B., Sander, C. Combining evolutionary information and neural networks to predict secondary structure. *Proteins* 19: 55–72, 1994.

Rost, B., Sander, C., Schneider, R. Redefining the goals of protein secondary structure prediction. *J. Mol. Biol.* 235: 13–26, 1994.

Smith, G.P. Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science* 228: 1315–1317, 1985.

Starovasnik, M.A., Braisted, A.C., Wells, J.A. Structural mimicry of a native protein by a minimized binding domain. J. A. *Proc. Natl. Acad. Sci. USA* 94: 10080–10085, 1997.

Stemmer, W.P.C. DNA shuffling by random fragmentation and reassembly: *In vitro* recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* 91: 10747–10751, 1994.

Weaver, L.H., Matthews, B.W. Structure of bacteriophage T4 lysozyme refined at 1.7 Å resolution. *J. Mol. Biol.* 193: 189–199, 1987.

Wells, T.N.C., Fersht, A.R. Hydrogen bonding in enzymatic catalysis analyzed by protein engineering. *Nature* 316: 656–657, 1985.

Wetzel, R. Harnessing disulfide bonds using protein engineering. *Trends Biochem. Sci.* 12: 478–482, 1987.

Wrighton, N.C., et al. Small peptides as potent mimetics of the protein hormone erythropoietin. *Science* 273: 458–463, 1996.

Zvelebil, M.J., et al. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* 195: 957–961, 1987.

Determination of Protein Structures

18

The structures described in this book have been determined by physical methods: most of them by x-ray crystallography, some of the smaller ones by nuclear magnetic resonance (NMR). We conclude the book with a short description of these techniques. It is not our aim to convert biologists into x-ray crystallographers and NMR spectroscopists; a complete explanation of the physical basis of these techniques and of the methods as currently practiced would fill more than one textbook. Our purpose is rather to convey the essence of the principles and procedures involved, so as to provide a general understanding of what is entailed in solving protein structures by these means. We will see how deriving a three-dimensional protein structure from x-ray or NMR data depends not only on the quality of the data themselves, but also on biochemical and sometimes genetic information that are essential to their interpretation.

Several different techniques are used to study the structure of protein molecules

Different techniques give different and complementary information about protein structure. The primary structure is obtained by biochemical methods, either by direct determination of the amino acid sequence from the protein or indirectly, but more rapidly, from the nucleotide sequence of the

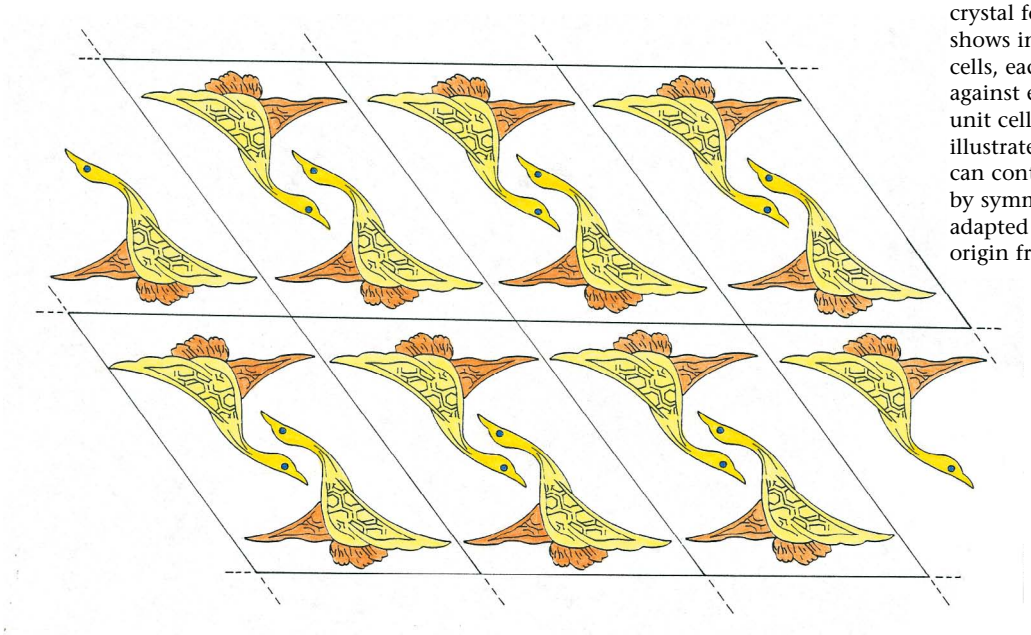
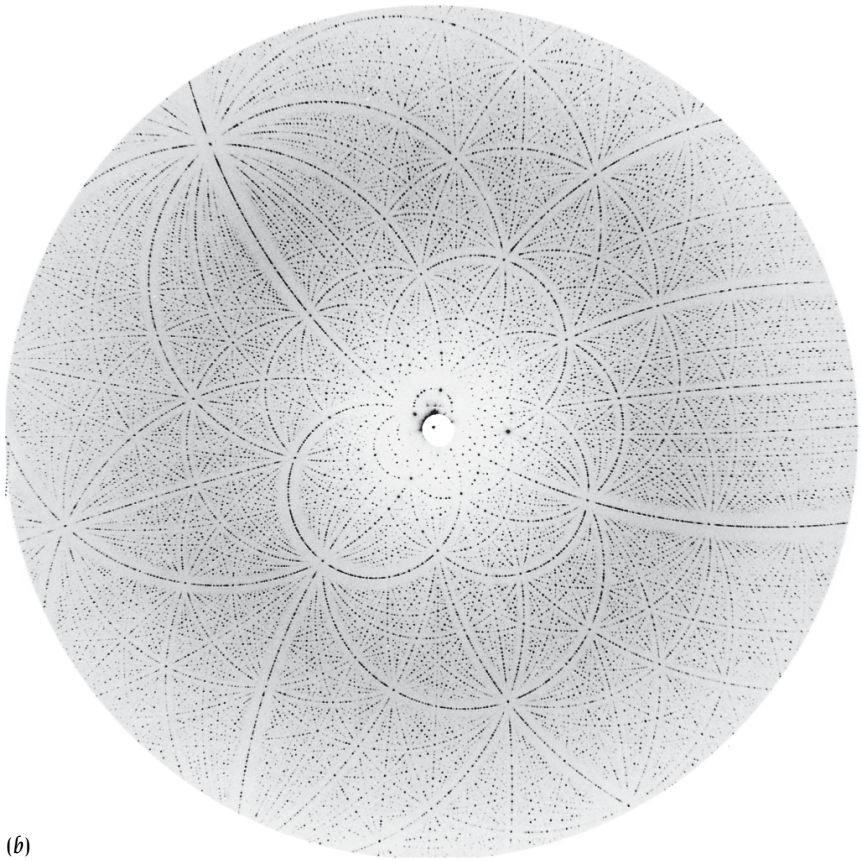


Figure 18.1 A crystal is built up from many billions of small identical units, or unit cells. These unit cells are packed against each other in three dimensions much as identical boxes are packed and stored in a warehouse. The unit cell may contain one or more than one molecule. Although the number of molecules per unit cell is always the same for all the unit cells of a single crystal, it may vary between different crystal forms of the same protein. The diagram shows in two dimensions several identical unit cells, each containing two objects packed against each other. The two objects within each unit cell are related by twofold symmetry to illustrate that each unit cell in a protein crystal can contain several molecules that are related by symmetry to each other. (The pattern is adapted from a Japanese stencil of unknown origin from the nineteenth century.)

(*a*)



(*b*)

corresponding gene or cDNA. The quaternary structure of large proteins or aggregates such as virus particles, ribosomes, or gap junctions can be determined by **electron microscopy**. In general, this method gives structural information at very low resolution, with no atomic details, although, if one can obtain ordered two-dimensional arrays of the object, the noise in the electron microscopic image can be reduced enough to reveal the shape of individual subunits or, in rare cases, even determine the path of the polypeptide chain within a protein molecule.

To obtain the secondary and tertiary structure, which requires detailed information about the arrangement of atoms within a protein, the main method so far has been x-ray crystallography. In recent years NMR methods have been developed to obtain three-dimensional models of small protein molecules, and NMR is becoming increasingly useful as it is further developed.

Figure 18.2 Well-ordered protein crystals diffract x-rays and produce diffraction patterns that can be recorded on film. The crystal shown in (a) is of the enzyme RuBisCo from spinach and the photograph in (b) is a recording (Laue photograph) of the diffraction pattern of a similar crystal of the same enzyme. The diffraction pattern was obtained using polychromatic radiation from a synchrotron source in the wavelength region 0.5 to 2.0 Å. More than 100,000 diffracted beams have been recorded on this film during an exposure of the crystal to x-rays for less than one second. (The Laue photograph was recorded by Janos Hajdu, Oxford, and Inger Andersson, Uppsala, at the synchrotron radiation source in Daresbury, England.)

Protein crystals are difficult to grow

The first prerequisite for solving the three-dimensional structure of a protein by **x-ray crystallography** is a well-ordered crystal that will diffract x-rays strongly. The crystallographic method depends, as we will see, upon directing a beam of x-rays onto a regular, repeating array of many identical molecules (Figure 18.1) so that the x-rays are diffracted from it in a pattern, a **diffraction pattern**, from which the structure of an individual molecule can be retrieved. The repeating unit forming the crystal is called the **unit cell**, and each unit cell may contain one or more molecules. Well-ordered crystals (Figure 18.2) are difficult to grow because globular protein molecules are large, spherical, or ellipsoidal objects with irregular surfaces, and it is impossible to pack them into a crystal without forming large holes or channels between the individual molecules. These channels, which usually occupy

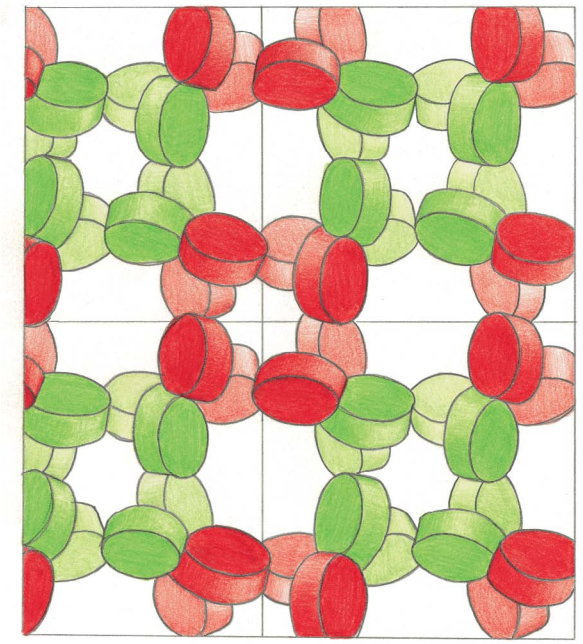


Figure 18.3 Protein crystals contain large channels and holes filled with solvent molecules, as shown in this diagram of the molecular packing in crystals of the enzyme glycolate oxidase. The subunits (colored disks) form octamers of molecular weight around 300 kDa, with a hole in the middle of each of about 15 Å diameter. Between the molecules there are channels (white) of around 70 Å diameter through the crystal. (Courtesy of Ylva Lindqvist, who determined the structure of this enzyme to 2.0 Å resolution in the laboratory of Carl Branden, Uppsala.)

more than half the volume of the crystal, are filled with disordered solvent molecules (Figure 18.3). The protein molecules are in contact with each other at only a few small regions, and even in these regions many interactions are indirect, through one or several layers of solvent molecules, usually water. This is one reason why structures of proteins determined by x-ray crystallography are the same as those for the proteins in solution.

Crystallization is usually quite difficult to achieve, and crystal growth can be slow; in some cases it may require months for sufficiently large crystals (~0.5 mm) to grow from microcrystals. The formation of crystals is also critically dependent on a number of different parameters, including pH, temperature, protein concentration, the nature of the solvent and precipitant as well as the presence of added ions or ligands to the protein. Many crystallization experiments are therefore required to screen all these parameters for the few combinations that might give crystals suitable for x-ray diffraction analysis. Crystallization robots and commercially available crystallization kits automate and speed up the tedious work of reproducibly setting up large numbers of crystallization experiments.

A pure and homogeneous protein sample is crucial for successful crystallization, and recombinant DNA techniques have been a major breakthrough in this regard. Proteins obtained from cloned genes in efficient expression vectors can be purified quickly to homogeneity in large quantities in a few purification steps. As a rule of thumb, a protein to be crystallized should ideally be more than 97% pure according to standard criteria of homogeneity. Crystals form when molecules are precipitated very slowly from **supersaturated solutions**. The most frequently used procedure for making protein crystals is the **hanging-drop** method (Figure 18.4), in which a drop of protein solution is brought very gradually to supersaturation by loss of water from the droplet to the larger reservoir that contains salt or polyethylene glycol solution.

Since there are so few direct packing interactions between protein molecules in a crystal, small changes in, for example, the pH of the solution can cause the molecules to pack in different ways to produce different crystal forms. The structures of some protein molecules such as lysozyme and myoglobin have been determined in different crystal forms and found to be essentially similar, except for a few side chains involved in packing interactions. Because they are so few, these interactions between protein molecules in a crystal do not change the overall structure of the protein. However,

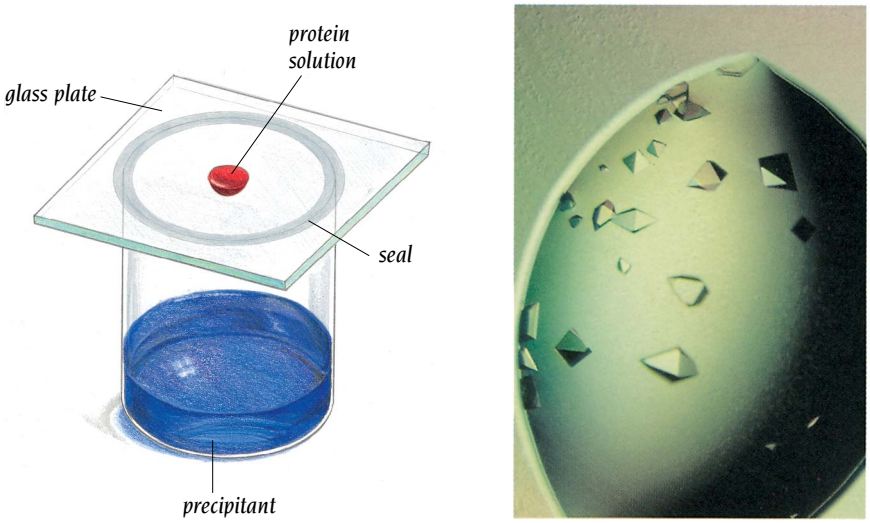


Figure 18.4 The hanging-drop method of protein crystallization. (a) About 10 µl of a 10 mg/ml protein solution in a buffer with added precipitant—such as ammonium sulfate, at a concentration below that at which it causes the protein to precipitate—is put on a thin glass plate that is sealed upside down on the top of a small container. In the container there is about 1 ml of concentrated precipitant solution. Equilibrium between the drop and the container is slowly reached through vapor diffusion, the precipitant concentration in the drop is increased by loss of water to the reservoir, and once the saturation point is reached the protein slowly comes out of solution. If other conditions such as pH and temperature are right, protein crystals will occur in the drop. (b) Crystals of recombinant enzyme RuBisCo from *Anacystis nidulans* formed by the hanging-drop method. (Courtesy of Janet Newman, Uppsala, who produced these crystals.)

different crystal forms can be more or less well ordered and hence give diffraction patterns of different quality. As a general rule, the more closely the protein molecules pack, and consequently the less water the crystals contain, the better is the diffraction pattern because the molecules are better ordered in the crystal.

X-ray sources are either monochromatic or polychromatic

X-rays are electromagnetic radiation at short wavelengths, emitted when electrons jump from a higher to a lower energy state. X-rays can be produced by high-voltage tubes in which a metal plate, the anode, is bombarded with accelerating electrons and thereby caused to emit x-rays of a specific wavelength, so-called **monochromatic x-rays**. The high voltage rapidly heats up the metal plate, which therefore has to be cooled. Efficient cooling is achieved by so-called rotating anode x-ray generators, where the metal plate revolves during the experiment so that different parts are heated up. Rotating anode x-ray generators are the conventional equipment used in most protein crystallography laboratories.

More powerful x-ray beams can be produced in synchrotron storage rings where electrons (or positrons) travel close to the speed of light. These particles emit very strong radiation at all wavelengths from short gamma rays to visible light. When used as an x-ray source, only radiation within a window of suitable wavelengths is channeled from the storage ring. **Polychromatic x-ray beams** are produced by having a broad window that allows through x-ray radiation with wavelengths of 0.2–2.0 Å. Such beams were used to record the Laue diffraction picture shown in Figure 18.2b. These very intense beams allow extremely short exposure times in diffraction experiments and can be used to collect data in experiments designed to observe changes in protein structure over very short periods of time; for example, electron transfer occurring in nanoseconds. Such studies are called **time-resolved crystallography.**

A very narrow window produces monochromatic radiation that is still several orders of magnitude more intense than the beam from conventional rotating anode x-ray sources. Such beams allow crystallographers to record diffraction patterns from very small crystals of the order of 50 micrometers or smaller. In addition, the diffraction pattern extends to higher resolution and consequently more accurate structural details are obtained as described later in this chapter. The availability and use of such beams have increased enormously in recent years and have greatly facilitated the x-ray determination of protein structures.

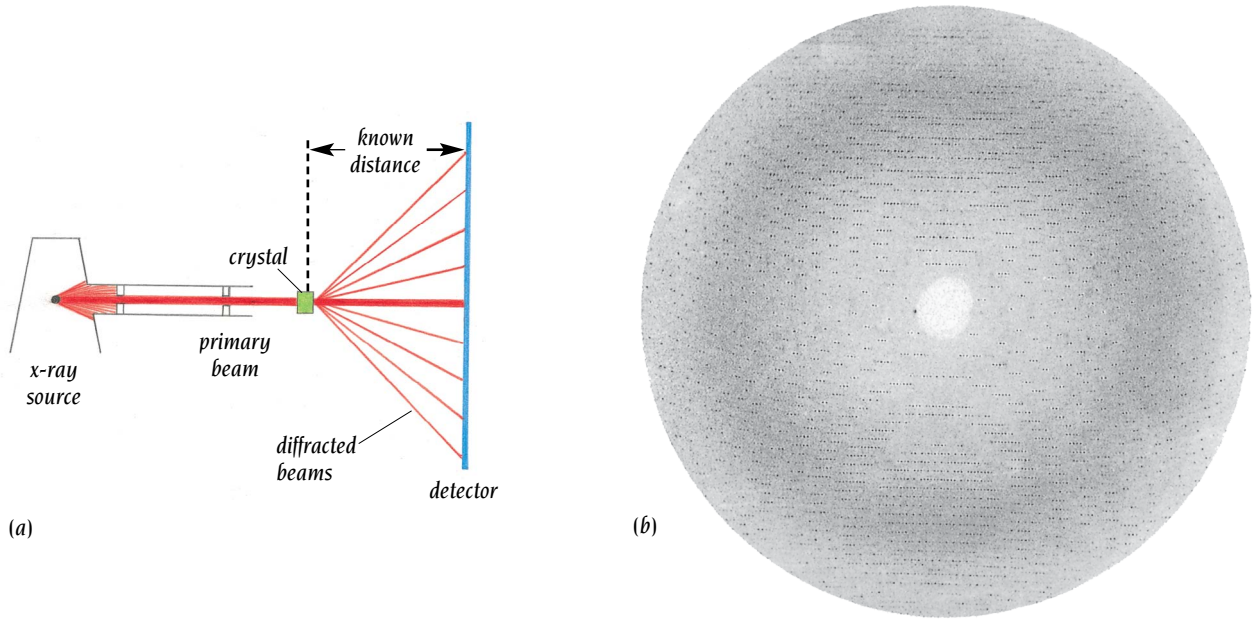


Figure 18.5 Schematic view of a diffraction experiment. (a) A narrow beam of x-rays (red) is taken out from the x-ray source through a collimating device. When the primary beam hits the crystal, most of it passes straight through, but some is diffracted by the crystal. These diffracted beams, which leave the crystal in many different directions, are recorded on a detector, either a piece of x-ray film or an area detector. (b) A diffraction pattern from a crystal of the enzyme RuBisCo using monochromatic radiation (compare with Figure 18.2b, the pattern using polychromatic radiation). The crystal was rotated one degree while this pattern was recorded.

In diffraction experiments a narrow and parallel beam of x-rays is taken out from the x-ray source and directed onto the crystal to produce diffracted beams (Figure 18.5a). The primary beam must strike the crystal from many different directions to produce all possible diffraction spots; and so the crystal is rotated in the beam during the experiment. Rotating the crystal is much easier than rotating the x-ray source, especially when it is a synchrotron.

The incident primary beam causes damage to both protein and solvent molecules. This produces free radicals that in turn damage other molecules in the crystal. In addition, heat is generated, especially from synchrotron radiation, and eventually the primary beam burns through the crystal. To minimize this damage the crystal in the x-ray beam can be cooled to about –150 °C Such cooling does not prevent the formation of free radicals but greatly reduces the rate at which they can diffuse in the crystal, and greatly prolongs the life of the crystal in the x-ray beam. To prevent ice forming and destroying the crystal, it is essential to replace some of the water in the crystal with cryoprotectants. The crystal is suspended in the cryoprotectant, then chilled in liquid nitrogen and transferred to the x-ray beam where it is kept in a fine jet of nitrogen gas from boiling liquid nitrogen. Cryocooling has proved to be one very important technical innovation and during recent years has been adopted by most crystallographic laboratories.

X-ray data are recorded either on image plates or by electronic detectors

Today the diffracted spots are usually recorded on an image plate rather than on x-ray film, the classical method (see Figure 18.5b), or by an electronic detector. The **image plate** is in effect a reusable film. The diffraction pattern recorded on the plate is scanned and stored in a computer. The image plate is then erased and ready for reuse. Electronic **area detectors** feed the signals they detect directly in a digitized form into a computer, and can therefore be regarded as an electronic film. They significantly reduce the time required to collect and measure diffraction data. To determine the structure of a protein, as we will see, it is necessary to compare x-ray data from native crystals of the protein with those from crystals in which different atoms of the protein are complexed with heavy metals. Moreover, to elucidate a protein's function x-ray data must also be collected from complexes with different types of bound ligands. In total, therefore, several hundred thousand diffraction spots are usually collected and measured for each protein.

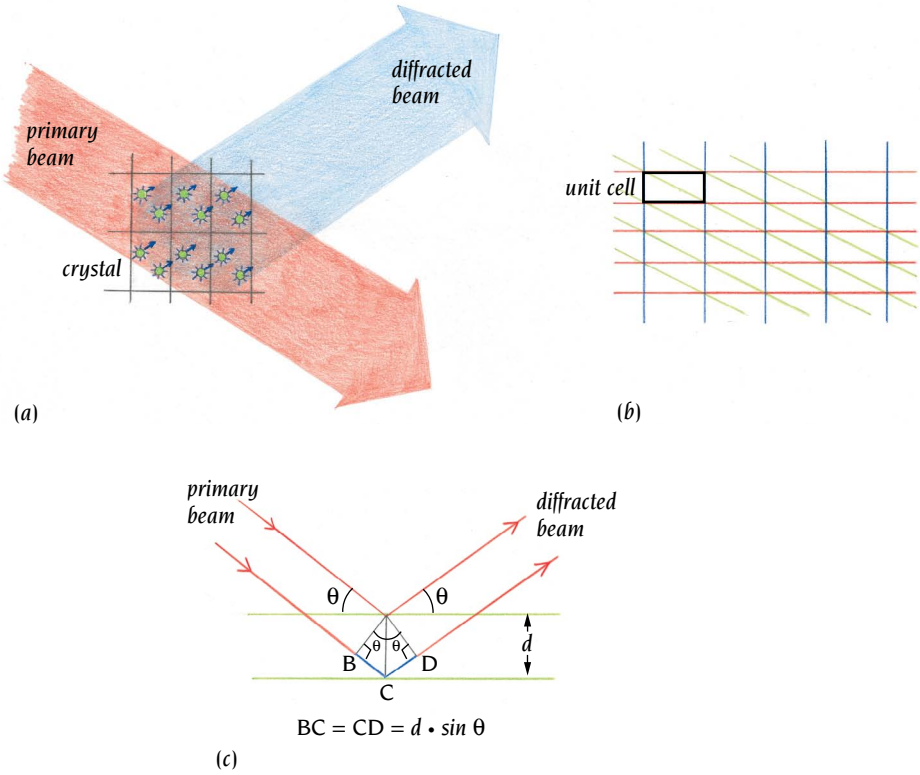


Figure 18.6 Diffraction of x-rays by a crystal. (a) When a beam of x-rays (red) shines on a crystal all atoms (green) in the crystal scatter x-rays in all directions. Most of these scattered x-rays cancel out, but in certain directions (blue arrow) they reinforce each other and add up to a diffracted beam. (b) Different sets of parallel planes can be arranged through the crystal so that each corner of all unit cells is on one of the planes of the set. The diagram shows in two dimensions three simple sets of parallel lines: red, blue, and green. A similar effect is seen when driving past a plantation of regularly spaced trees. One sees the trees arranged in different sets of parallel rows. (c) X-ray diffraction can be regarded as reflection of the primary beam from sets of parallel planes in the crystal. Two such planes are shown (green), separated by a distance *d*. The primary beam strikes the planes at an angle θ and the reflected beam leaves at the same angle, the reflection angle. X-rays (red) that are reflected from the lower plane have traveled farther than those from the upper plane by a distance BC + CD, which is equal to $2d \cdot \sin\theta$. Reflection can only occur when this distance is equal to the wavelength λ of the x-ray beam and Bragg's law—$2d \cdot \sin\theta = \lambda$—gives the conditions for diffraction. To determine the size of the unit cell, the crystal is oriented in the beam so that reflection is obtained from the specific set of planes in which any two adjacent planes are separated by the length of one of the unit cell axes. This distance, *d*, is then equal to $\lambda/(2 \cdot \sin\theta)$. The wavelength, λ, of the beam is known since we use monochromatic radiation. The reflection angle, θ, can be calculated from the position of the diffracted spot on the film, using the relation derived in Figure 18.7, where the crystal to film distance can be easily measured. The crystal is then reoriented, and the procedure is repeated for the other two axes of the unit cell.

The rules for diffraction are given by Bragg's law

When the primary beam from an x-ray source strikes the crystal, most of the x-rays travel straight through it. Some, however, interact with the electrons on each atom and cause them to oscillate. The oscillating electrons serve as a new source of x-rays, which are emitted in almost all directions. We refer to this rather loosely as **scattering**. When atoms and hence their electrons are arranged in a regular three-dimensional array, as in a crystal, the x-rays emitted from the oscillating electrons interfere with one another. In most cases, these x-rays, colliding from different directions, cancel each other out; those from certain directions, however, will add together to produce diffracted beams of radiation that can be recorded as a pattern on a photographic plate or detector (Figure 18.6a).

How is the diffraction pattern obtained in an x-ray experiment such as that shown in Figure 18.5b related to the crystal that caused the diffraction? This question was addressed in the early days of x-ray crystallography by Sir Lawrence Bragg of Cambridge University, who showed that diffraction by a crystal can be regarded as the reflection of the primary beam by sets of parallel planes, rather like a set of mirrors, through the unit cells of the crystal (see Figure 18.6b and c).

X-rays that are reflected from adjacent planes travel different distances (see Figure 18.6c), and Bragg showed that diffraction only occurs when the difference in distance is equal to the wavelength of the x-ray beam. This distance is dependent on the reflection angle, which is equal to the angle between the primary beam and the planes (see Figure 18.6c).

The relationship between the reflection angle, θ, the distance between the planes, *d*, and the wavelength, λ, is given by **Bragg's law**: $2d \cdot \sin\theta = \lambda$. This relation can be used to determine the size of the unit cell (see legend to Figure 18.6c and Figure 18.7). Briefly, the position on the film of the diffraction data relates each spot to a specific set of planes through the crystal. By using Bragg's law, these positions can be used to determine the size of the unit cell.

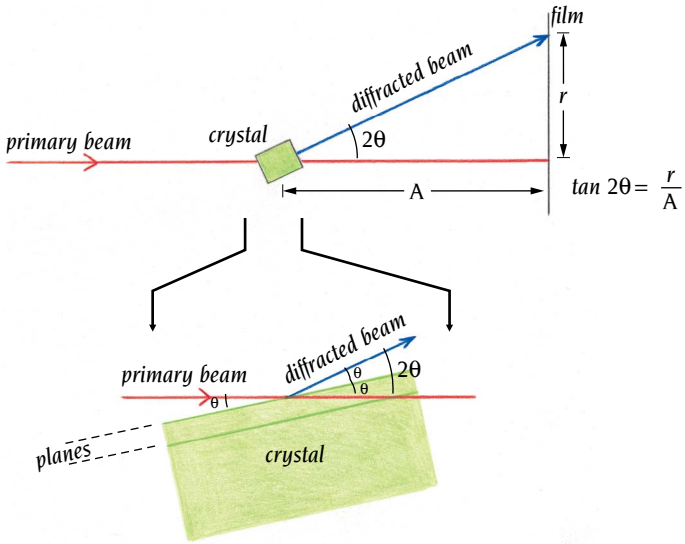


Figure 18.7 The reflection angle, θ, for a diffracted beam can be calculated from the distance (r) between the diffracted spot on a film and the position where the primary beam hits the film. From the geometry shown in the diagram the tangent of the angle $2\theta = r/A$. A is the distance between crystal and film that can be measured on the experimental equipment, while r can be measured on the film. Hence θ can be calculated. The angle between the primary beam and the diffracted beam is 2θ as can be seen on the enlarged insert at the bottom. It shows that this angle is equal to the angle between the primary beam and the reflecting plane plus the reflection angle, both of which are equal to θ.

Phase determination is the major crystallographic problem

Each atom in a crystal scatters x-rays in all directions, and only those that positively interfere with one another, according to Bragg's law, give rise to diffracted beams (see Figure 18.6a) that can be recorded as a distinct **diffraction spot** above background. Each diffraction spot is the result of interference of all x-rays with the same diffraction angle emerging from all atoms. For a typical protein crystal, myoglobin, each of the about 20,000 diffracted beams that have been measured contains scattered x-rays from each of the around 1500 atoms in the molecule. To extract information about individual atoms from such a system requires considerable computation. The mathematical tool that is used to handle such problems is called the **Fourier transform**, invented by the French mathematician Jean Baptiste Joseph Fourier while he served as a bureaucrat in the government of Napoleon Bonaparte.

Each diffracted beam, which is recorded as a spot on the film, is defined by three properties: the **amplitude**, which we can measure from the intensity of the spot; the **wavelength**, which is set by the x-ray source; and the **phase**, which is lost in x-ray experiments (Figure 18.8). We need to know all three properties for all of the diffracted beams to determine the position of the atoms giving rise to the diffracted beams. How do we find the phases of the diffracted beams? This is the so-called phase problem in x-ray crystallography.

In small-molecule crystallography the phase problem was solved by so-called direct methods (recognized by the award of a Nobel Prize in chemistry to Jerome Karle, US Naval Research Laboratory, Washington, DC, and Herbert Hauptman, the Medical Foundation, Buffalo). For larger molecules, protein crystallographers have stayed at the laboratory bench using a method pioneered by Max Perutz and John Kendrew and their co-workers to circumvent the phase problem. This method, called **multiple isomorphous replacement**

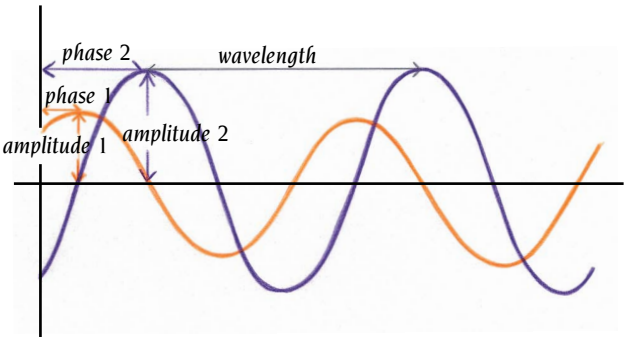


Figure 18.8 Two diffracted beams (purple and orange), each of which is defined by three properties: amplitude, which is a measure of the strength of the beam and which is proportional to the intensity of the recorded spot; phase, which is related to its interference, positive or negative, with other beams; and wavelength, which is set by the x-ray source for monochromatic radiation.

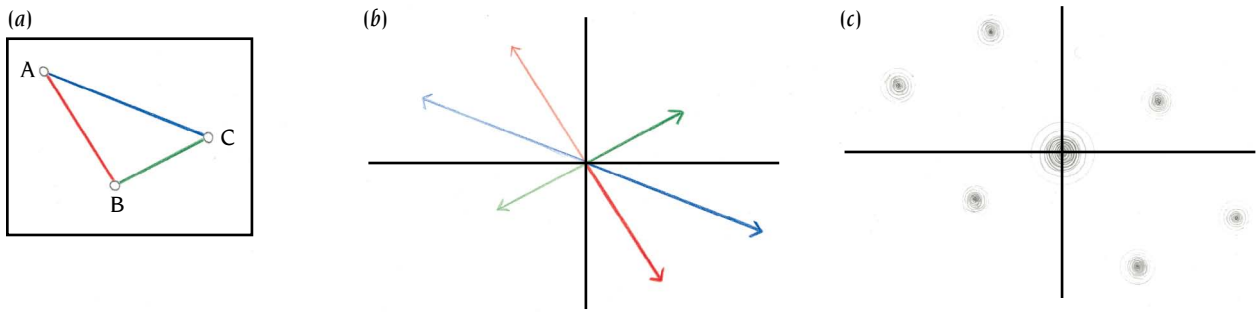


Figure 18.9 Fourier summations of the intensity differences between diffracted spots from crystals of the protein alone and protein plus heavy metals give vector maps between the heavy atoms. Three atoms—A, B, and C—are at specific positions in the unit cell in (a). They give vectors A–B, A–C, and B–C, which are drawn from a common origin in (b) in dark colors. They also give the same vectors in the opposite directions as shown in light colors. The experimentally observed vector map is shown in (c) with a large peak at the origin corresponding to zero vectors between an atom and itself. It is straightforward to deduce the map in (c) from the atomic arrangement in (a). It is more difficult to do the reverse, to deduce the atomic arrangement in (a) from the vector map in (c), especially if there are many atoms in the unit cell that give rise to a large number of peaks in the vector map. For example, with 10 atoms in the unit cell there are 90 different vectors between the atoms.

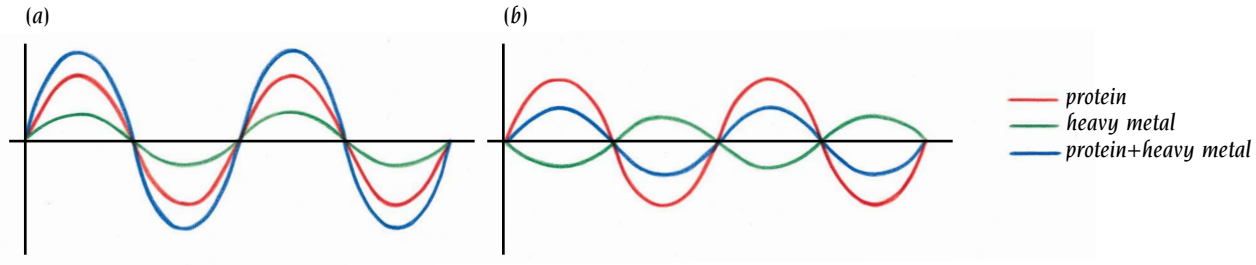
(**MIR**), requires the introduction of new x-ray scatterers into the unit cell of the crystal. These additions should be heavy atoms (so that they make a significant contribution to the diffraction pattern); there should not be too many of them (so that their positions can be located); and they should not change the structure of the molecule or of the crystal cell—in other words, the crystals should be isomorphous. In practice, isomorphous replacement is usually done by diffusing different heavy-metal complexes into the channels of preformed protein crystals. With luck the protein molecules expose side chains in these solvent channels, such as SH groups, that are able to bind heavy metals. It is also possible to replace endogenous light metals in metalloproteins with heavier ones, e.g., zinc by mercury or calcium by samarium.

Since such heavy metals contain many more electrons than the light atoms, H, N, C, O, and S, of the protein, they scatter x-rays more strongly. All diffracted beams would therefore increase in intensity after heavy-metal substitution if all interference were positive. In fact, however, some interference is negative; consequently, following heavy-metal substitution, some spots measurably increase in intensity, others decrease, and many show no detectable difference.

How do we find phase differences between diffracted spots from intensity changes following heavy-metal substitution? We first use the intensity differences to deduce the positions of the heavy atoms in the crystal unit cell. Fourier summations of these intensity differences give maps of the vectors between the heavy atoms, the so-called **Patterson maps** (Figure 18.9). From these vector maps it is relatively easy to deduce the atomic arrangement of the heavy atoms, so long as there are not too many of them. From the positions of the heavy metals in the unit cell, one can calculate the amplitudes and phases of their contribution to the diffracted beams of the protein crystals containing heavy metals.

How is that knowledge used to find the phase of the contribution from the protein in the absence of the heavy-metal atoms? We know the phase and amplitude of the heavy metals and the amplitude of the protein alone. In addition, we know the amplitude of protein plus heavy metals (i.e., protein heavy-metal complex); thus we know one phase and three amplitudes. From this we can calculate whether the interference of the x-rays scattered by the heavy metals and protein is constructive or destructive (Figure 18.10). The extent of positive or negative interference plus knowledge of the phase of the heavy metal together give an estimate of the phase of the protein.

Figure 18.10 The diffracted waves from the protein part (red) and from the heavy metals (green) interfere with each other in crystals of a heavy-atom derivative. If this interference is positive as illustrated in (a), the intensity of the spot from the heavy-atom derivative (blue) crystal will be stronger than that of the protein (red) alone (larger amplitude). If the interference is negative as in (b), the reverse is true (smaller amplitude).



Unfortunately, the problem is underdetermined so that two different phase angles are equally good solutions. To distinguish between these two possible solutions, a second heavy-metal complex must be used, which also gives two possible phase angles. Only one of these will have the same value as one of the two previous phase angles; it therefore represents the correct phase angle. In practice, more than two different heavy-metal complexes are needed to give a reasonably good phase determination for all reflections. Each individual phase estimate contains experimental errors arising from errors in the measured amplitudes; furthermore, for many reflections, the intensity differences are too small to measure after one particular isomorphous replacement, and others must be tried.

Phase information can also be obtained by Multiwavelength Anomalous Diffraction experiments

For certain x-ray wavelengths, the interaction between the x-rays and the electrons of an atom causes the electrons to absorb the energy of the x-ray. This causes a change in the x-ray scattering of the atom, called **anomalous scattering**, that depends both on the type of atom and on the wavelength of the x-rays. The size of this change is small and negligible for light atoms such as hydrogen, carbon, nitrogen and oxygen but is measurable for heavier atoms such as selenium, iron, zinc, and mercury. Due to this effect, small changes of the wavelength of the incident x-ray beam around the absorption edge of the heavy atom produce measurable intensity differences in the diffraction pattern. It is sufficient to have only one such heavy atom bound to each protein molecule of medium size (about 200 amino acid residues) for the effect to be measurable.

The intensity differences obtained in the diffraction pattern by illuminating such a crystal by x-rays of different wavelengths can be used in a way similar to the method of multiple isomorphous replacement to obtain the phases of the diffracted beams. This method of phase determination which is called **Multiwavelength Anomalous Diffraction**, **MAD**, and which was pioneered by Wayne Hendrickson at Columbia University, US, is now increasingly used by protein cystallographers.

The MAD method requires access to synchrotron radiation since different wavelengths are used, and it also requires that the crystal contains heavy atoms. Some protein molecules, such as metalloenzymes, contain intrinsic metal atoms but most proteins do not. However, using recombinant DNA technology it is possible to incorporate selenomethionine instead of methionine into recombinant proteins, thereby fulfilling the requirements for using the MAD method. Proteins with selenomethionine have very similar structures to the methionine-containing proteins. The structure of a number of selenomethionine-containing proteins, including several described in earlier chapters, has been determined using data collected at experimental stations specifically designed for MAD experiments at several synchrotron sources. Alternatively, proteins can be soaked in heavy metal solutions as for conventional x-ray structure determination (as discussed above).

Building a model involves subjective interpretation of the data

The amplitudes and the phases of the diffraction data from the protein crystals are used to calculate an **electron-density map** of the repeating unit of the crystal. This map then has to be interpreted as a polypeptide chain with a particular amino acid sequence. The interpretation of the electron-density map is complicated by several limitations of the data. First of all, the map itself contains errors, mainly due to errors in the phase angles. In addition, the quality of the map depends on the **resolution** of the diffraction data, which in turn depends on how well-ordered the crystals are. This directly influences the image that can be produced. The resolution is measured in Å

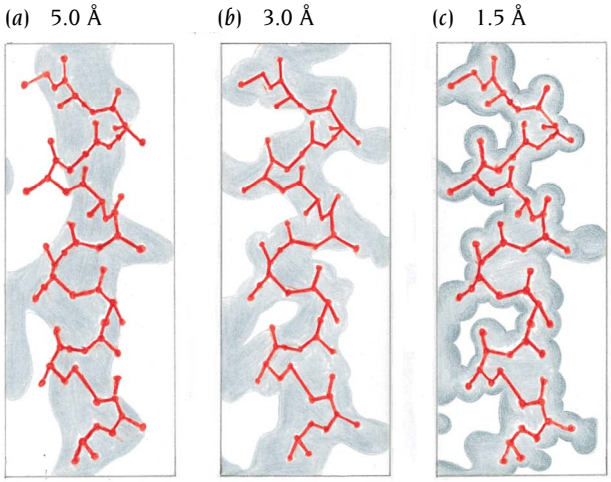


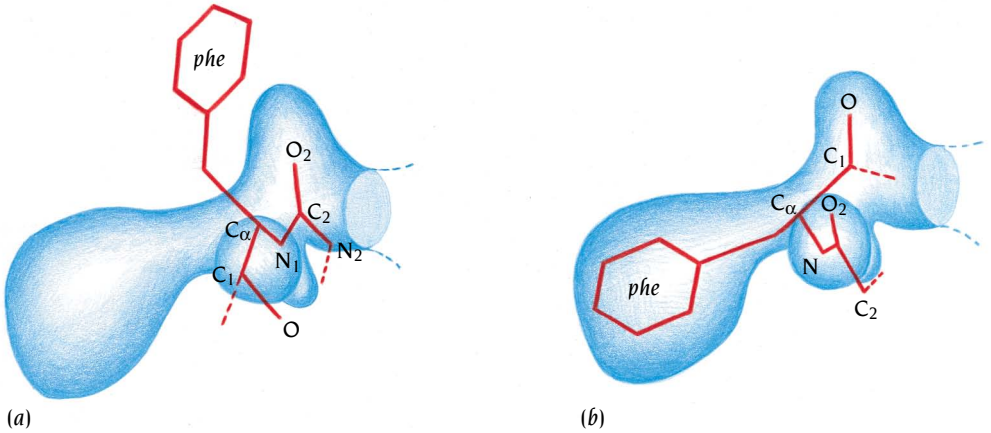
Figure 18.11 Electron-density maps at different resolution show more detail at higher resolution. (a) At low resolution (5.0 Å) individual groups of atoms are not resolved, and only the rodlike feature of an α helix can be deduced. (b) At medium resolution (3.0 Å) the path of the polypeptide chain can be traced, and (c) at high resolution (1.5 Å) individual atoms start to become resolved. Relevant parts of the protein chain (red) are superimposed on the electron densities (gray). The diagrams show one α helix from a small protein, myohemerythrin. [Adapted from W.A. Hendrickson in *Protein Engineering* (eds. D.L. Oxender and C.F. Fox.), p. 11. New York: Liss, 1987.]

units; the smaller this number is, the higher the resolution and therefore the greater the amount of detail that can be seen (Figure 18.11).

From a map at low resolution (5 Å or higher) one can obtain the shape of the molecule and sometimes identify α-helical regions as rods of electron density. At medium resolution (around 3 Å) it is usually possible to trace the path of the polypeptide chain and to fit a known amino acid sequence into the map. At this resolution it should be possible to distinguish the density of an alanine side chain from that of a leucine, whereas at 4 Å resolution there is little side chain detail. Gross features of functionally important aspects of a structure usually can be deduced at 3 Å resolution, including the identification of active-site residues. At 2 Å resolution details are sufficiently well resolved in the map to decide between a leucine and an isoleucine side chain, and at 1 Å resolution one sees atoms as discrete balls of density. However, the structures of only a few small proteins have been determined to such high resolution.

Building the initial **model** is a trial-and-error process. First, one has to decide how the polypeptide chain weaves its way through the electron-density map. The resulting chain trace constitutes a hypothesis, by which one tries to match the density of the side chains to the known sequence of the polypeptide. This sounds easy, but it is not; a map showing continuous density from N-terminus to C-terminus is rare. More usually one produces a number of matches between the electron density and discontinuous regions of the sequence that may initially account for only a small fraction of the molecule and may be internally inconsistent. When a reasonable chain trace has finally been obtained, an initial model is built to give the best fit of the atoms to the electron density (Figure 18.12). Today, computer graphics are exploited both for chain tracing and for model building to present the data and manipulate the models.

Figure 18.12 The electron-density map is interpreted by fitting into it pieces of a polypeptide chain with known stereochemistry such as peptide groups and phenyl rings. The electron density (blue) is displayed on a graphics screen in combination with a part of the polypeptide chain (red) in an arbitrary orientation (a). The units of the polypeptide chain can then be rotated and translated relative to the electron density until a good fit is obtained (b). Notice that individual atoms are not resolved in such electron densities, there are instead lumps of density corresponding to groups of atoms. [Adapted from A. Jones *Methods Enzym.* (eds. H.W. Wyckoff, C.H. Hirs, and S.N. Timasheff) 115B: 162, New York: Academic Press, 1985.]



Errors in the initial model are removed by refinement

The initial model will contain errors. Provided the protein crystals diffract to high enough resolution (better than ~2.5 Å), most of the errors can be removed by crystallographic refinement of the model. In this process the model is changed to minimize the difference between the experimentally observed diffraction amplitudes and those calculated for a hypothetical crystal containing the model instead of the real molecule. This difference is expressed as an **R factor**, residual disagreement, which is 0.0 for exact agreement and around 0.59 for total disagreement.

In general, the R factor is between 0.15 and 0.20 for a well-determined protein structure. The residual difference rarely is due to large errors in the model of the protein molecule, but rather it is an inevitable consequence of errors and imperfections in the data. These derive from various sources, including slight variations in conformation of the protein molecules and inaccurate corrections both for the presence of solvent and for differences in the orientation of the microcrystals from which the crystal is built. This means that the final model represents an average of molecules that are slightly different both in conformation and orientation, and not surprisingly the model never corresponds precisely to the actual crystal.

The atoms of a protein's structure are usually defined by four parameters, three coordinates that give their position in space and one quantity, B, which is called the temperature factor. For well refined, correct structures these B-values are of the order of 20 or less. High B-values, 40 or above, in a local region can be due to flexibility or slight disorder, but also serve as a warning that the model of this region may be incorrect.

In refined structures at high resolution (around 2 Å) there are usually no major errors in the orientation of individual residues, and the estimated errors in atomic positions are around 0.1–0.2 Å provided the amino acid sequence is known. Hydrogen bonds both within the protein and to bound ligands can be identified with a high degree of confidence.

At medium resolution (around 3 Å) it is possible to make serious errors in the interpretation of the electron-density map, and there are, unfortunately, a number of them in the literature. Errors usually arise because elements of secondary structure are wrongly connected by the loop regions. Alpha helices and β strands in the interior of the protein are rigid in structure and well defined in the electron-density map. The loop regions, however, are usually more flexible, and therefore the corresponding electron density is less well defined. It is easy to make errors in such regions in the preliminary interpretations of electron-density maps at medium resolution. These errors are usually caught and corrected before publication since such models will not refine properly and are likely to be incompatible with existing biochemical data.

However, some models containing serious errors have been published. They all have been based on data to only medium resolution together with insufficient phase information, which gives large errors in the electron density. It should, therefore, be kept in mind that unrefined structures with R values higher than 0.30 at medium resolution may contain errors, although the overwhelming majority of such published structures have survived subsequent refinement at high resolution.

Recent technological advances have greatly influenced protein crystallography

In the early days of protein crystallography the determination of a protein structure was laborious and time consuming. The diffracted beams were obtained from weak x-ray sources and recorded on films that had to be manually scanned and measured. The available computers were far from adequate for the problem, with a computing power roughly equal to present-day pocket calculators. Computer graphics were not available, and models of the protein had to be built manually from pieces of steel rod. To determine the

structure of even a small protein molecule, therefore, required many years of work and entailed time-consuming bottlenecks at almost every stage.

The situation is radically different today. The diffraction pattern can now be recorded on electronic area detectors coupled to powerful microcomputers that immediately interpret and process the recorded signals. Data collection that only a few years ago required many months of work is now done in a few days. If the in-house x-ray source is too weak for the problem, there are synchrotron sources available in several centers around the world that provide x-ray beams that are brighter by several orders of magnitude. Powerful computers in the laboratory provide the crystallographer with immediate access to almost all the computing power he or she needs. The electron-density maps are interpreted, and models of the protein molecules are built by the crystallographer sitting in front of a computer graphics screen. He or she is greatly aided by sophisticated software that involves semiautomatic methods for the model building using knowledge from databases of previously determined and refined protein structures.

These technical advances have greatly facilitated the use of crystallography for protein structure determination. One significant problem, however, still remains: obtaining crystals that diffract to high resolution. Some protein molecules give excellent crystals after the first few trials, others may require several months of screening for the proper crystallization conditions, and many have so far resisted all attempts to crystallize them. Fortunately, it is now possible to determine the structure of small protein molecules in solution by nuclear magnetic resonance (NMR) methods, and of large complexes by a combination of x-ray or NMR studies of individual or smaller pieces and fiber diffraction or electron microscopy studies of the complete complex.

X-ray diffraction can be used to study the structure of fibers as well as crystals

As described in Chapter 2, the first complete protein structure to be determined was the globular protein myoglobin. However, the α helix that was recognized in this structure, and which has emerged as a persistent structural motif in the many hundreds of globular proteins determined subsequently, was first observed in x-ray diffraction studies of fibrous proteins.

Polymer molecules that have a high degree of regularity in their monomer sequence tend to assume helical rather than globular conformations, and fibers are formed when these helices become aligned with each other. This is well illustrated by the proteins keratin and collagen and by the nucleic acid, DNA. The regularity in the DNA double-helix is so high that fibers drawn from a concentrated DNA gel are highly ordered; the long, thread-like DNA molecules extend parallel to the length of the fiber with a high degree of regularity in their side-by-side packing extending over many molecules. These regularly packed molecules form structures know as **crystallites**, and a typical fiber of 100 microns diameter contains a large number of such crystallites separated by less ordered regions where the molecules, while still largely parallel to the fiber length, are much less regularly packed. Because the crystallites are in random orientation about the direction of the fiber length, a diffraction pattern recorded from the fiber is similar to the pattern obtained when a single crystal is rotated 360° about the vertical axis while the data is being recorded. Therefore the diffraction pattern from a crystalline fiber can be analyzed using standard crystallographic techniques. The power of crystalline fiber diffraction analysis is illustrated by the detailed stereochemical information on the A and B conformations of DNA, reviewed in Chapter 7.

The polymer molecules in fibers typically assume helical structures with one pitch of the helix forming the repeating unit. This symmetry is the origin of the characteristic "cross-like" variation in overall intensity across the diffraction pattern from helical molecules. As the degree of regularity in the arrangement of repeating units in an array decreases, the diffraction spots become broader. For a completely irregular array, such as a crystalline powder,

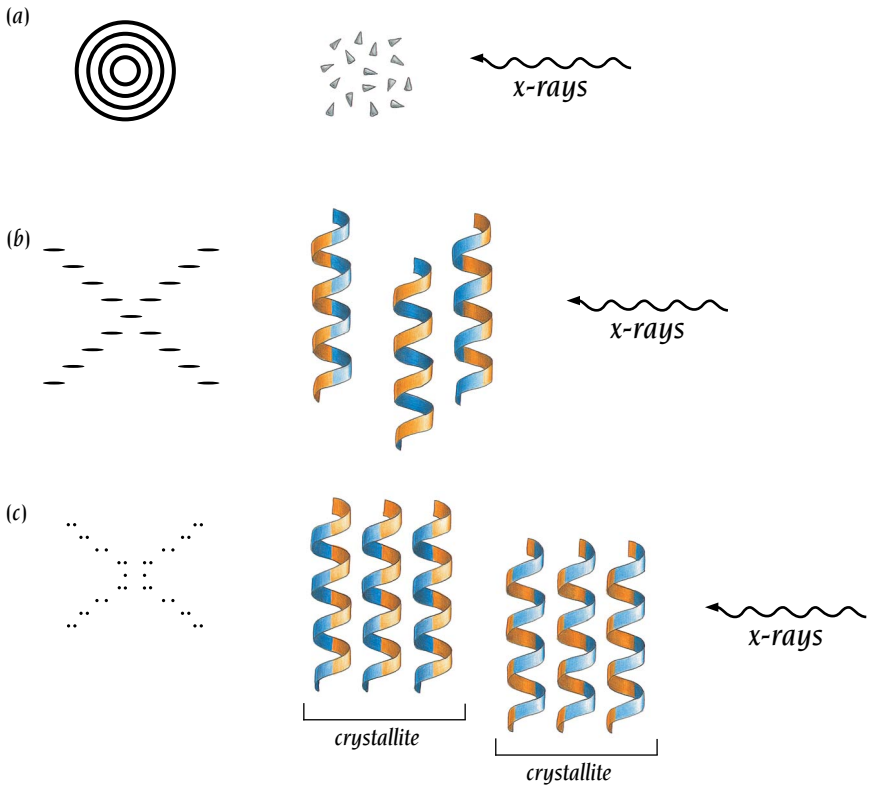


Figure 18.13 Typical fiber diffraction patterns (left) for differently ordered arrays of molecules (right). (a) A randomly oriented array, such as a powder of small crystals. The molecules within each crystal are highly ordered, but the crystals are in random orientation with respect to each other. (b) A fiber formed from an array of poorly ordered molecules. Three molecules within the array are shown (in a typical x-ray specimen there would be many more); they are parallel but have little regularity in their side-by-side packing. (c) A fiber consisting of crystallites of well ordered molecules. Two crystallites are shown, each with only three of the many molecules they would typically contain. The molecules within each crystallite are aligned with a high degree of order in their side-by-side packing and relative orientation, but the crystallites are rotated with respect to each other. A similar pattern is obtained for a single crystal that is rotated during the data collection.

the diffraction pattern reduces to the scattering from a single molecule averaged over all orientations and is observed as a continuous distribution without any characteristic diffraction spots (Figure 18.13a). However, the chains of extended helical molecules in a fiber tend to remain parallel even when there is no regularity in the orientation about the helix axis. This parallelism is reflected in the overall intensity distribution within the diffraction pattern and in particular the restriction of diffraction to layer lines (see Figure 18.13b,c). The main features in the cross-like diffraction pattern of a helical structure can be illustrated by considering diffraction by a single slit that has the shape of a sine curve and that therefore corresponds to a projection of a helix. The diffraction patterns for projections of two helices differing in pitch length are illustrated in Figure 18.14. The separation between the layer lines is determined by the helix pitch: as the helix pitch increases the layer lines move closer together (compare parts a and b of Figure 18.14).

In addition to the reciprocal relationship between the helix pitch and layer line spacing, Figure 18.14 illustrates the reciprocal relationship between the orientation of the arms of the cross and the angle of climb of the helix: as the helix becomes steeper the arms of the cross become more horizontal.

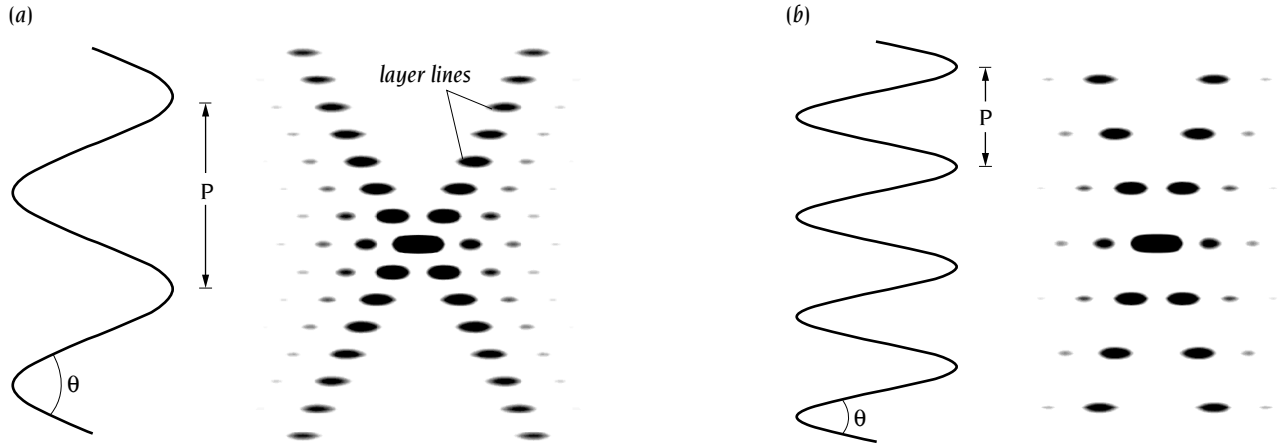


Figure 18.14 The diffraction pattern of helices in fiber crystallites can be simulated by the diffraction pattern of a single slit with the shape of a sine curve (representing the projection of a helix). Two such simulations are given in (a) and (b), with the helix shown to the left of its diffraction pattern. The spacing between the layer lines is inversely related to the helix pitch, P and the angle of the cross arms in the diffraction pattern is related to the angle of climb of the helix, θ. The helix in (b) has a smaller pitch and angle of climb than the helix in (a). (Courtesy of W. Fuller.)

The structure of biopolymers can be studied using fiber diffraction

A number of important protein molecules occur *in vivo* as fibrous structures. In keratin and collagen, groups of extended polymer molecules are wound around each other like strands in a rope, while muscle fibers contain a mixture of extended and globular components in a highly ordered array (see Chapter 14). The periodicities in these complex arrays of polymer molecules may be as large as hundreds or thousands of angstroms, whereas the periodicities of structural features in individual polymer molecules such as the pitch of the DNA double helix are typically a few tens of angstroms. The fibrous state offers a less constrained environment than a single crystal so that it is possible to follow changes in the x-ray fiber diffraction pattern as a consequence of structural transitions. In the case of DNA, changes of helix pitch and the number of nucleotide-pairs per pitch have been followed, while in muscle such time-resolved techniques have allowed changes in the diffraction pattern to be recorded during the contractile cycle.

The diffraction from helical molecules can be illustrated by the diffraction patterns from DNA (Figure 18.15). The pattern from the A form on the left has sharp diffraction spots across the whole pattern, indicating the regular packing of molecules in crystallites. In contrast, the pattern from the B form consists almost entirely of continuous diffraction along layer lines and the structure is said to be semi-crystalline. The pitch of the A form is 28 Å compared with 34 Å for the B form and this is reflected in Figure 18.15 by the smaller spacing between layer lines in the B form. The strong meridional diffraction near the top of the pattern from the B form is due to scattering from the stack of base pairs that form the core of the DNA double helix. Its position indicates that the distance between successive base pairs is 3.4 Å and therefore that there are 10 base pairs per helix pitch.

Knowing the helix pitch, it is possible to determine an approximate value for the radius of the helix in the B form from the inclination of the

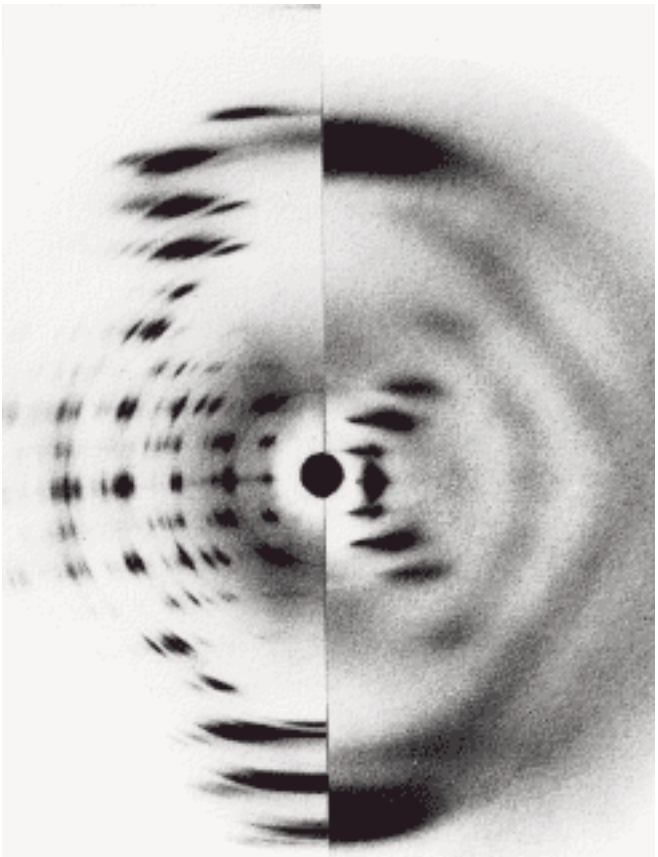
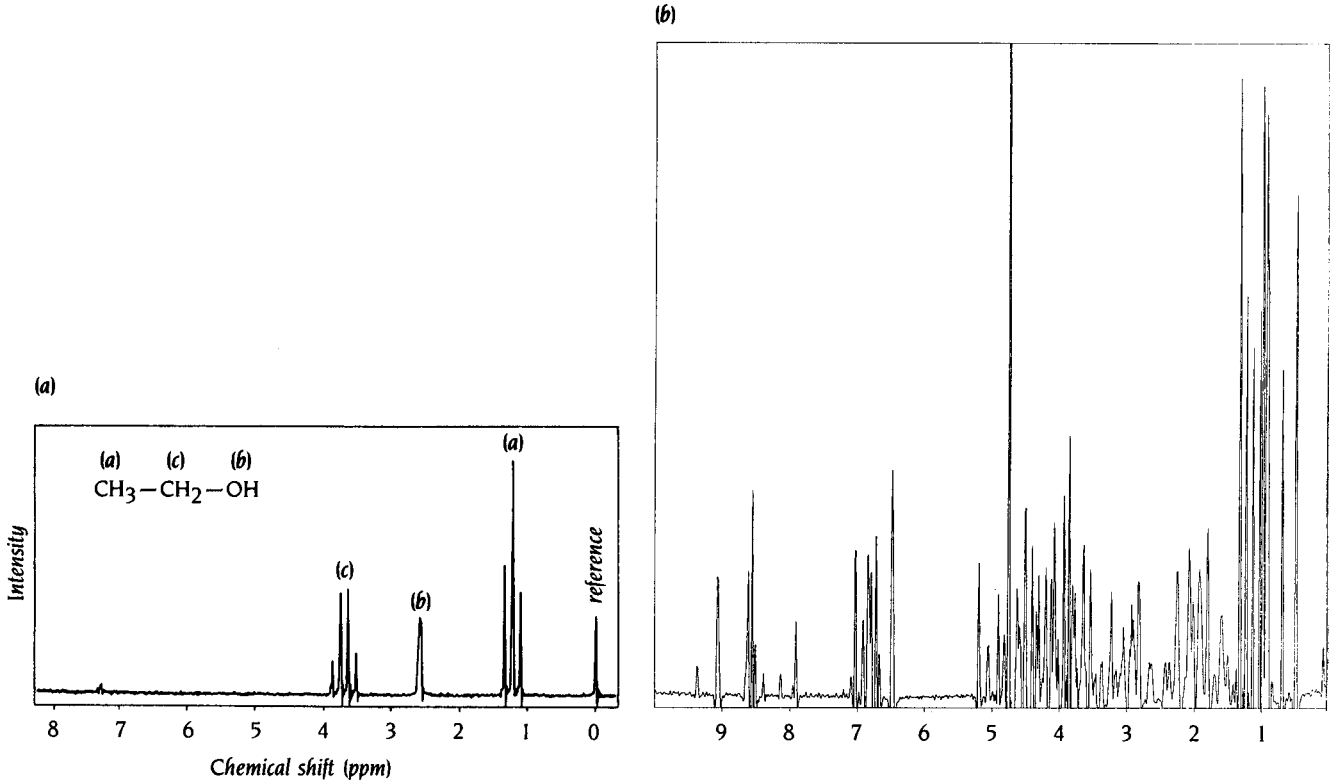


Figure 18.15 Diffraction patterns of DNA, showing the patterns obtained for both A-DNA (left half) and B-DNA (right half). (Courtesy of W. Fuller.)



arms of the cross in the B diffraction pattern. Arguments such as those outlined above, based on x-ray fiber diffraction patterns obtained by Maurice Wilkins and Rosalind Franklin, were used by Jim Watson and Francis Crick in the construction of their double-helical model for DNA.

Figure 18.16 One-dimensional NMR spectra. (a) ^{1}H-NMR spectrum of ethanol. The NMR signals (chemical shifts) for all the hydrogen atoms in this small molecule are clearly separated from each other. In this spectrum the signal from the CH_3 protons is split into three peaks and that from the CH_2 protons into four peaks close to each other, due to the experimental conditions. (b) ^{1}H-NMR spectrum of a small protein, the C-terminal domain of a cellulase, comprising 36 amino acid residues. The NMR signals from many individual hydrogen atoms overlap and peaks are obtained that comprise signals from many hydrogen atoms. (Courtesy of Per Kraulis, Uppsala, from data published in Kraulis et al., *Biochemistry* 28: 7241–7257, 1989.)

NMR methods use the magnetic properties of atomic nuclei

Certain atomic nuclei, such as ^{1}H, ^{13}C, ^{15}N, and ^{31}P have a magnetic moment or **spin**. The chemical environment of such nuclei can be probed by **nuclear magnetic resonance** (**NMR**) and this technique can be exploited to give information on the distances between atoms in a molecule. These distances can then be used to derive a three-dimensional model of the molecule. Most structure determinations of protein molecules by NMR have used the spin of ^{1}H, since hydrogen atoms are abundant in proteins. Small proteins can be analyzed by ^{1}H (proton) NMR but to study larger proteins and to obtain sufficient data to determine side chain conformations it is necessary to introduce ^{13}C and ^{15}N into the protein. This is usually done by producing the protein in microorganisms grown in media enriched with these isotopes. NMR studies of proteins containing one of the isotopes are called 3-D NMR, and when both ^{13}C and ^{15}N are present they are called 4-D NMR.

When protein molecules are placed in a strong magnetic field, the spin of their hydrogen atoms aligns along the field. This equilibrium alignment can be changed to an excited state by applying **radio frequency** (**RF**) pulses to the sample. When the nuclei of the protein molecule revert to their equilibrium state, they emit RF radiation that can be measured. The exact frequency of the emitted radiation from each nucleus depends on the molecular environment of the nucleus and is different for each atom, unless they are chemically equivalent and have the same molecular environment (Figure 18.16a). These different frequencies are obtained relative to a reference signal and are called **chemical shifts**. The nature, duration, and combination of applied RF pulses can be varied enormously, and different molecular properties of the sample can be probed by selecting the appropriate combination of pulses.

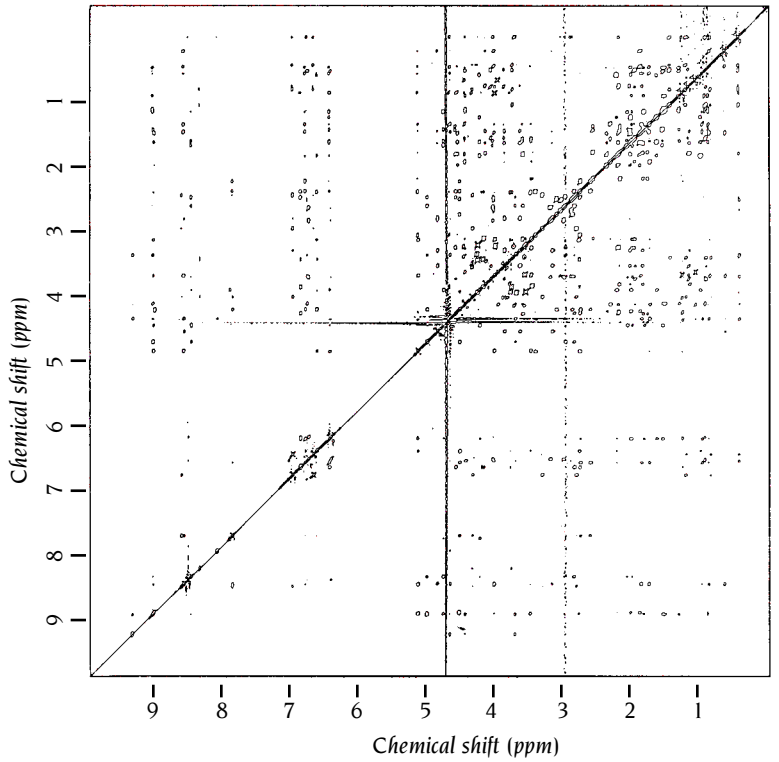


Figure 18.17 Two-dimensional NMR spectrum of the C-terminal domain of a cellulase. The peaks along the diagonal correspond to the spectrum shown in Figure 18.16b. The off-diagonal peaks in this NOE spectrum represent interactions between hydrogen atoms that are closer than 5 Å to each other in space. From such a spectrum one can obtain information on both the secondary and tertiary structures of the protein. (Courtesy of Per Kraulis, Uppsala.)

In principle, it is possible to obtain a unique signal (chemical shift) for each hydrogen atom in a protein molecule, except those that are chemically equivalent, for example, the protons on the CH_3 side chain of an alanine residue. In practice, however, such one-dimensional NMR spectra of protein molecules (see Figure 18.16b) contain overlapping signals from many hydrogen atoms because the differences in chemical shifts are often smaller than the resolving power of the experiment. In recent years this problem has been bypassed by designing experimental conditions that yield a two-dimensional NMR spectrum, the results of which are usually plotted in a diagram as shown in Figure 18.17.

The diagonal in such a diagram corresponds to a normal one-dimensional NMR spectrum. The peaks off the diagonal result from interactions between hydrogen atoms that are close to each other in space. By varying the nature of the applied RF pulses these off-diagonal peaks can reveal different types of interactions. A **COSY** (*co*rrelation *s*pectroscop*y*) experiment gives peaks between hydrogen atoms that are covalently connected through one or two other atoms, for example, the hydrogen atoms attached to the nitrogen and C_α atoms within the same amino acid residue (Figure 18.18a). An **NOE** (*n*uclear *O*verhauser *e*ffect) spectrum, on the other hand, gives peaks between pairs of hydrogen atoms that are close together in space even if they are from amino acid residues that are quite distant in the primary sequence (see Figure 18.18b).

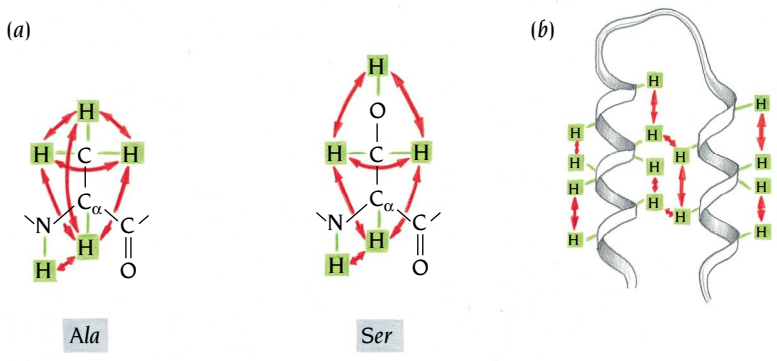


Figure 18.18 (a) COSY NMR experiments give signals that correspond to hydrogen atoms that are covalently connected through one or two other atoms. Since hydrogen atoms in two adjacent residues are covalently connected through at least three other atoms (for instance, HCα-C′-NH), all COSY signals reveal interactions within the same amino acid residue. These interactions are different for different types of side chains. The NMR signals therefore give a "fingerprint" of each amino acid. The diagram illustrates fingerprints (red) of residues Ala and Ser. (b) NOE NMR experiments give signals that correspond to hydrogen atoms that are close together in space (less than 5 Å), even though they may be far apart in the amino acid sequence. Both secondary and tertiary structures of small protein molecules can be derived from a collection of such signals, which define distance constraints between a number of hydrogen atoms along the polypeptide chain.

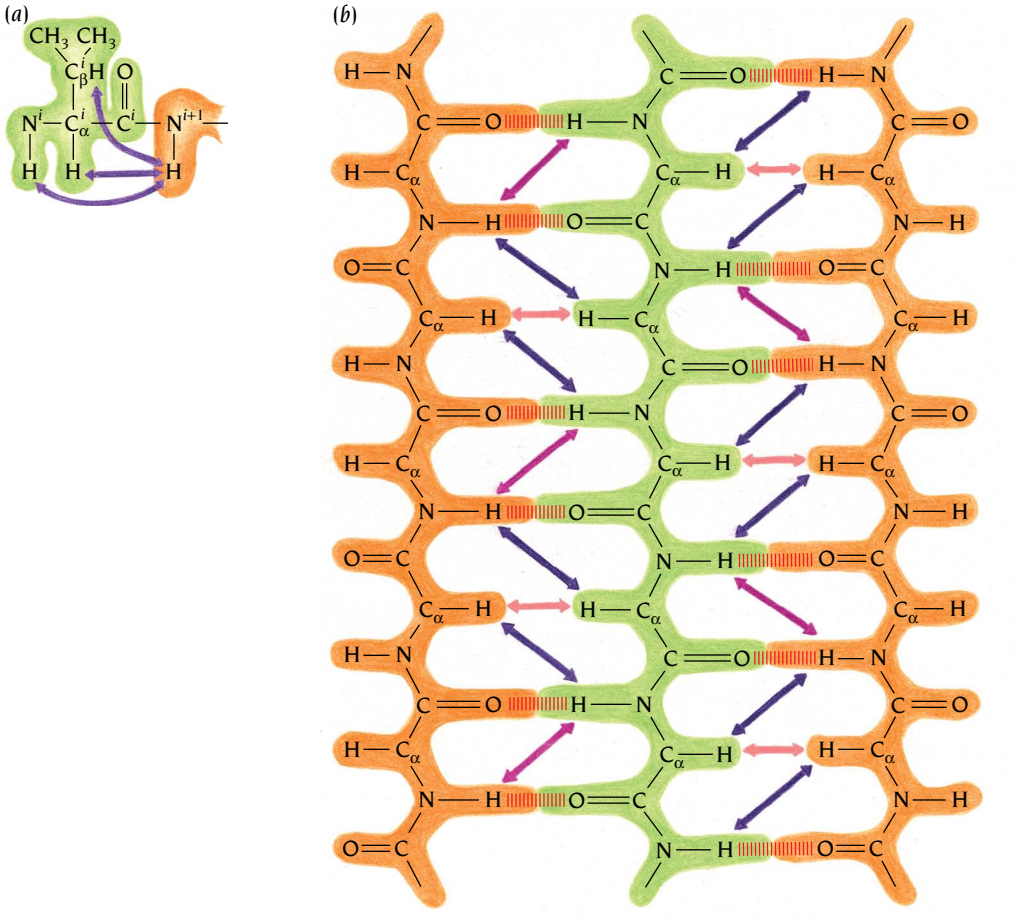
Two-dimensional NMR spectra of proteins are interpreted by the method of sequential assignment

Two-dimensional NOE spectra, by specifying which groups are close together in space, contain three-dimensional information about the protein molecule. It is far from trivial, however, to assign the observed peaks in the spectra to hydrogen atoms in specific residues along the polypeptide chain because the order of peaks along the diagonal has no simple relation to the order of amino acids along the polypeptide chain. This problem has in principle been solved in the laboratory of Kurt Wüthrich in the ETH, Zürich, where the method of **sequential assignment** was developed.

Sequential assignment is based on the differences in the number of hydrogen atoms and their covalent connectivity in the different amino acid residues. Each type of amino acid has a specific set of covalently connected hydrogen atoms that will give a specific combination of **cross-peaks**, a "fingerprint," in a COSY spectrum (see Figure 18.18a). From the COSY spectrum it is therefore possible to identify the H atoms that belong to each amino acid residue and, in addition, determine the nature of the side chain of that residue. However, the order of these fingerprints along the diagonal has no relation to the amino acid sequence of the protein. For example, when the fingerprint in one specific region of the COSY spectrum of the *lac*-repressor segment was assigned to a Ser residue, it was not known whether this fingerprint corresponded to Ser 16, Ser 28, or Ser 31 in the amino acid sequence.

The sequence-specific assignment, however, can be made from NOE spectra (see Figures 18.17 and 18.18b) that record signals from H atoms that are close together in space. In addition to the interactions between H atoms that are far apart in the sequence, these spectra also record interactions between H atoms from sequentially adjacent residues, specifically, interactions from the H atom attached to the main chain N of residue number $i + 1$ to H atoms bonded to N, C_α, and C_β of residue number i (Figure 18.19a).

Figure 18.19 (a) Adjacent residues in the amino acid sequence of a protein can be identified from NOE spectra. The H atom attached to residue $i + 1$ (orange) is close to and interacts with (purple arrows) the H atoms attached to N, C_α, and C_β of residue i (light green). These interactions give cross-peaks in the NOE spectrum that identify adjacent residues and are used for sequence-specific assignment of the amino acid fingerprints derived from a COSY spectrum. (b) Regions of secondary structure in a protein have specific interactions between hydrogen atoms in sequentially nonadjacent residues that give a characteristic pattern of cross-peaks in an NOE spectrum. In antiparallel β-sheet regions there are interactions between C_α-H atoms of adjacent strands (pink arrows), between N-H and C_α-H atoms (dark purple arrows), and between N-H atoms of adjacent strands (light purple arrows). The corresponding pattern of cross-peaks in an NOE spectrum identifies the residues that form the antiparallel β sheet. Parallel β sheets and α helices are identified in a similar way.



These signals in the NOE spectra therefore in principle make it possible to determine which fingerprint in the COSY spectrum comes from a residue adjacent to the one previously identified. For example, in the case of the *lac*-repressor fragment the specific Ser residue that was identified from the COSY spectrum was shown in the NOE spectrum to interact with a His residue, which in turn interacted with a Val residue. Comparison with the known amino acid sequence revealed that the tripeptide Ser-His-Val occurred only once, for residues 28–30.

In practice, it is difficult to make unique assignments for longer pieces than di- or tri-peptides, since NOE signals also occur between residues close together in space but far apart in the sequence. Therefore, the peptide segments that have been uniquely identified by NMR are usually matched with corresponding segments in the independently determined amino acid sequence of the protein. Thus knowledge of the amino acid sequence is just as essential for the correct interpretation of NMR spectra as it is for the interpretation of electron-density maps in x-ray crystallography. Whereas x-ray crystallography directly gives an image of a three-dimensional model of the protein molecule, NMR spectroscopy identifies H atoms in the protein that are close together in space, and this information is then used to derive, indirectly, a three-dimensional model of the protein.

Distance constraints are used to derive possible structures of a protein molecule

The final result of the sequence-specific assignment of NMR signals, preferably done using interactive computer graphics, is a list of **distance constraints** from specific hydrogen atoms in one residue to hydrogen atoms in a second residue. The list contains a large number of such distances, which are usually divided into three intervals within the region 1.8 Å to 5 Å, depending on the intensity of the NOE peak. This list immediately identifies the secondary structure elements of the protein molecule because both α helices and β sheets have very specific sets of interactions of less than 5 Å between their hydrogen atoms (see Figure 18.19b). It is also possible to derive models of the three-dimensional structure of the protein molecule. However, usually a set of possible structures rather than a unique structure (Figure 18.20) is obtained, each of the possible structures obeying the distance constraints equally well. The sets of possible structures, which are frequently seen in NMR articles, do not, therefore, represent different actual conformations of a protein molecule present in solution. Rather they are simply the different structures that are compatible with data obtained by current methods. The primary source of this ambiguity is an insufficient number of measured distance constraints. Because of this ambiguity, the accuracy of an NMR structure is not constant over the whole molecule and is also difficult to quantify.

In addition to the problem of ambiguity, there are other limitations to the use of NMR methods for the determination of protein structures. The most severe concerns the size of the protein molecules whose structures can be determined. Currently, the upper limit is molecules with molecular weights of around 25 kDa, but this limit will be increased in the future by using improved methods and equipment. Furthermore, the method requires highly concentrated protein solutions, on the order of 1–2 mM, with the additional requirement that the protein molecules must not aggregate at these concentrations. In addition, the pH of the solution should be lower than about 6 for proton NMR experiments. The exchanges of the NH protons in the main chain become so fast at higher pH that it is very difficult to observe them with NMR, and the signals from these hydrogen atoms are essential for the sequential assignment procedure.

How well do NMR-derived structures agree with those determined by x-ray methods? The structures of some different globular proteins that have been independently obtained by the two methods—such as bovine pancreatic

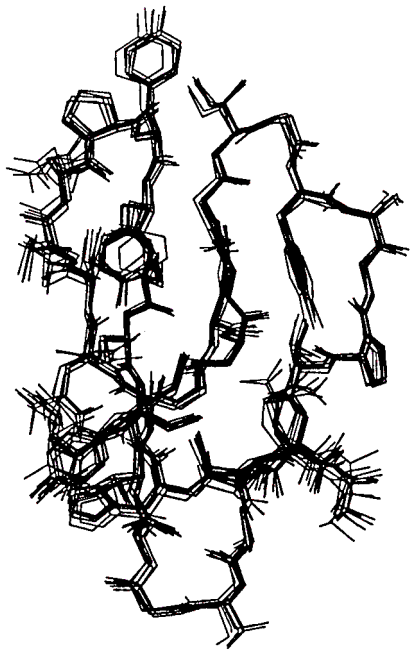


Figure 18.20 The two-dimensional NMR spectrum shown in Figure 18.17 was used to derive a number of distance constraints for different hydrogen atoms along the polypeptide chain of the C-terminal domain of a cellulase. The diagram shows 10 superimposed structures that all satisfy the distance constraints equally well. These structures are all quite similar since a large number of constraints were experimentally obtained. (Courtesy of P. Kraulis, Uppsala, from data published in P. Kraulis et al., *Biochemistry* 28: 7241–7257, 1989, by copyright permission of the American Chemical Society.)

trypsin inhibitor (see Figure 2.14a), plastocyanin (see Figure 2.11c) and thioredoxin from *E. coli* (see Figure 2.7)—show that NMR and x-ray crystallography give nearly identical results. The minor differences that exist are of the same order of magnitude as usually seen between x-ray structures of unrelated crystal forms of the same protein or determinations made under different experimental conditions. In other words, they are mostly small differences in loop regions of the main chain and different conformations of exposed side chains.

The situation is different for other examples—for example, the peptide hormone glucagon and a small peptide, metallothionein, which binds seven cadmium or zinc atoms. Here large discrepancies were found between the structures determined by x-ray diffraction and NMR methods. The differences in the case of glucagon can be attributed to genuine conformational variability under different experimental conditions, whereas the disagreement in the metallothionein case was later shown to be due to an incorrectly determined x-ray structure. A re-examination of the x-ray data of metallothionein gave a structure very similar to that determined by NMR.

NMR and x-ray crystallography are in many respects complementary. X-ray crystallography deals with the structure of proteins in the crystalline state, while NMR determines the structure in solution. The time scales of the measurements are different: NMR is more suitable for investigation of various dynamic processes such as those during folding, while x-ray crystallography is more suitable for characterization of protein surfaces and the water structure around the protein. X-ray crystallography remains the only method available to determine the structure of large protein molecules, whereas NMR is the method of choice for small protein molecules that might be difficult to crystallize.

Biochemical studies and molecular structure give complementary functional information

Our current knowledge of the relation between structure and function of protein molecules is insufficient to deduce the function of a protein from its structure alone, although, as we have seen, structural homology with proteins of known function can sometimes allow this. It is necessary to combine biochemical studies with structural information. Biochemical and cell biological studies can tell us if a protein is a receptor, a transport molecule, or an enzyme and, in addition, which ligands can bind to it, as well as the functional effects of such ligand binding. Studies of the three-dimensional structure of complexes between specific ligands and the protein will then give detailed information on how the active site is constructed and which amino acid residues are involved in ligand binding. Examples that we have described include protein–DNA interaction in Chapters 8, 9, and 10, sugar binding to a sugar transport protein in Chapter 4, and binding of inhibitors to enzymes that cleave peptide bonds in Chapter 11.

The specific role of each amino acid residue for the function of the protein can be tested by making specific mutations of the residue in question and examining the properties of the mutant protein. By combining in this way functional studies in solution, site-directed mutagenesis by recombinant DNA techniques, and three-dimensional structure determination, we are now in a position to gain fresh insights into the way protein molecules work.

Conclusion

The three-dimensional structure of protein molecules can be experimentally determined by two different methods, x-ray crystallography and NMR. The interaction of x-rays with electrons in molecules arranged in a crystal is used to obtain an electron-density map of the molecule, which can be interpreted in terms of an atomic model. Recent technical advances, such as powerful computers including graphics work stations, electronic area detectors, and

very strong x-ray sources from synchrotron radiation, have greatly facilitated the use of x-ray crystallography.

Crystallization of proteins can be difficult to achieve and usually requires many different experiments varying a number of parameters, such as pH, temperature, protein concentration, and the nature of solvent and precipitant. Protein crystals contain large channels and holes filled with solvents, which can be used for diffusion of heavy metals into the crystals. The addition of heavy metals is necessary for the phase determination of the diffracted beams.

X-ray structures are determined at different levels of resolution. At low resolution only the shape of the molecule is obtained, whereas at high resolution most atomic positions can be determined to a high degree of accuracy. At medium resolution the fold of the polypeptide chain is usually correctly revealed as well as the approximate positions of the side chains, including those at the active site. The quality of the final three-dimensional model of the protein depends on the resolution of the x-ray data and on the degree of refinement. In a highly refined structure, with an R value less than 0.20 at a resolution around 2.0 Å, the estimated errors in atomic positions are around 0.1 Å to 0.2 Å, provided the amino acid sequence is known.

Biological fibers, such as can be formed by DNA and fibrous proteins, may contain crystallites of highly ordered molecules whose structure can in principle be solved to atomic resolution by x-ray crystallography. In practice, however, these crystallites are rarely as ordered as true crystals, and in order to locate individual atoms it is necessary to introduce stereochemical constraints in the x-ray analysis so that the structure can be refined by molecular modeling.

In NMR the magnetic-spin properties of atomic nuclei within a molecule are used to obtain a list of distance constraints between those atoms in the molecule, from which a three-dimensional structure of the protein molecule can be obtained. The method does not require protein crystals and can be used on protein molecules in concentrated solutions. It is, however, restricted in its use to small protein molecules.

Selected readings

Bernstein, F.C., et al. The protein data bank: a computer-based archival file for macromolecular structures. *J. Mol. Biol.* 112: 535–542, 1977.

Blundell, T.L., Johnson, L.N. *Protein Crystallography*. London: Academic Press, 1976.

Branden, C.-I., Jones, A. Between objectivity and subjectivity. *Nature* 343: 687–689, 1990.

Clore, G.M., Gronenborn, A.M. Determination of three-dimensional structures of proteins and nucleic acids in solution by nuclear magetic resonance spectroscopy. *CRC Crit. Rev. Biochem.* 24: 479–564, 1989.

Drenth, J. *Principles of protein x-ray crystallogrphy*. Berlin: Springer-Verlag, 1994.

Eisenberg, D., Hill, C.P. Protein crystallography: more surprises ahead. *Trends Biochem. Sci.* 14: 260–264, 1989.

Ferre-D'Amare, A.R., Burley, S.K. Use of dynamic light scattering to assess crystallizability of macromolecules and macromolecular assemblies. *Structure* 2: 357–359, 1994.

Hendrickson, W. Determination of macromolecular structures from anomalous diffraction of synchrotron radiation. *Science* 254: 51–58, 1991.

Kleywegt, G.J., Jones, T.A. Where freedom is given, liberties are taken. *Structure* 3: 535–540, 1995.

McPherson, A. *The Preparation and Analysis of Protein Crystals*. New York: Wiley, 1982.

McPherson, A., et al. The science of macromolecular crystallization. *Structure* 3: 759–768, 1995.

Rodgers, D.W. Cryocrystallography. *Structure* 2: 1135–1140, 1994.

Walter, R.L., et al. High resolution macromolecular structure determination using CCD detectors and synchrotron radiation. *Structure* 3: 835–844, 1995.

Wilson, H.R. *Diffraction of X-rays by Proteins, Nucleic Acids and Viruses*. London: Edward Arnold, 1966.

Wright, P. What can two-dimensional NMR tell us about proteins? *Trends Biochem. Sci.* 14: 255–260, 1989.

Wüthrich, K. *NMR of Proteins and Nucleic Acids*. New York: Wiley, 1986.

Wüthrich, K. Protein structure determination in solution by nuclear magnetic resonance spectroscopy. *Science* 243: 45–50, 1989.

Wyckoff, H.W., Hirs, C.H.W., Timasheff, S.N. Diffraction methods for biological macromolecules. *Methods Enzymol.* 114: 330–386, 1985.

Protein Structure on the World Wide Web

The World Wide Web has transformed the way in which we obtain and analyze published information on proteins. What only a few years ago would take days or weeks and require the use of expensive computer workstations can now be achieved in a few minutes or hours using personal computers, both PCs and Macintosh, connected to the internet. The Web contains hundreds of sites of interest to molecular biologists, many of which are listed in **Pedro's BioMolecular Research Tools (http://www.fmi.ch/biology/research_tools.html)**. Many sites provide free access to databases that make it very easy to obtain information on structurally related proteins, the amino acid sequences of homologous proteins, relevant literature references, medical information and metabolic pathways. This development has opened up new opportunities for even non-specialists to view and manipulate a structure of interest or to carry out amino-acid sequence comparisons, and one can now rapidly obtain an overview of a particular area of molecular biology. We shall here describe some Web sites that are of interest from a structural point of view. Updated links to these sites can be found in the ***Introduction to Protein Structure*** **Web site (http://www.ProteinStructure.com/)**.

Many Web sites offer the opportunity to view the structures of proteins interactively, using a protein's atomic coordinates to produce images of different types that can be rotated and zoomed. Both the atomic coordinates and the computer software required for this can be freely either accessed from the Web or downloaded to desktop computers for off-line use. Three commonly used and free programs to view proteins structures on personal computers are RasMol, Chime and Mage. **RasMol (http://www.umass.edu/microbio/rasmol/)** runs on personal computers and produces interactive molecular images from a molecule's atomic coordinates. **Chime (http//www.mdli.com/tech/chemscape.html)** is a plug-in for Web browsers that allows interactive RasMol-like images to be embedded within Web pages. **Kinemages (http://www.faseb.org/protein/kinemages/kinpage.html)** are interactive molecular images produced by the program Mage. Many on-line journals use the kinemage format to display protein structures, and kinemages illustrating many of the protein structures discussed in this book can be found in the *Introduction to Protein Structure* Kinemage supplement disks (for further information, see http://www.Protein Structure.com).

The **Brookhaven Protein Data Bank, PDB (http://www.pdb.bnl.gov)**, is the primary store of experimentally determined atomic coordinates of proteins. Each coordinate set has a unique identification code that can be

retrieved together with the coordinates and other information about the structure, including interactive images, by searching for the protein's name or publication details. The ENTREZ search engine at the **National Center for Biotechnology Information (NCBI; http://www3.ncbi.nlm.nih.gov/Entrez/)** allows the searching of integrated databases for, amongst other things, protein sequence data, protein structures and bibliographic data. This site also provides the Vector Alignment Tool (VAST), allowing one to find and view similar structures, and the Basic Alignment Search Tool (BLAST), allowing one to find similar sequences.

There are also several databases that have arranged all known protein structures into some classification scheme. **SCOP (http://scop.mrc lmb.cam.ac.uk/scop/)** is a database of structural domains arranged in a hierarchical manner according to structural and evolutionary relatedness. The SCOP site allows structures to be viewed interactively and enables the search for proteins with similar amino acid sequences using BLAST. **CATH (http://www.biochem.ucl.ac.uk/bsm/cath/)** arranges domains according to class (secondary structure composition), architecture (orientations of the secondary structures) and topology (shape and connectivity of secondary structure). The CATH site also includes a useful glossary of terms used in the description of protein structures. Protein topologies are extensively discussed in **TOPS (http://www3.ebi.ac.uk/tops/)**, a site that searches for specific topologies and which can determine the topology of a new protein. **FSSP (http://www2.ebi.ac.uk/dali/fssp/)** is based on an all-against-all comparison of known structures. FSSP allows one to view similar structures superimposed and to obtain sequences of homologous proteins. A network service that allows the comparison of three-dimensional structures is **DALI (http://www2.ebi.ac.uk/dali/)**. The coordinates of a new structure can be submitted to the site and DALI will check these coordinates against known structures to reveal biologically interesting similarities.

The ExPASy molecular biology Web site of the **Swiss Institute of Bioinformatics (http://www.expasy.ch)** covers many aspects of protein sequence and structure. It includes SWISS-PROT, which is an annotated protein sequence database, and SWISS-MODEL, an automated knowledge-based protein modeling server that allows one to model the three-dimensional structure of a protein whose sequence is known, based on the known structure of a homologous protein. Two very useful databases concerning the compilation and multiple sequence alignment of homologous domains are **Pfram (http://www.sanger.ac.uk/Pfam/)** and **ProDom (http://protein.toulouse.inra.fr/prodom.html)**.

There are now a large number of Web sites devoted to particular fields of research. The **Nucleic Acid Database atlas (http://ndbserver.rutgers.edu/NDB/ndb.html)** provides a resource for viewing many DNA, RNA, protein–DNA and protein–RNA structures that have been solved by x-ray crystallography. In addition, many researchers produce Web pages that are concerned with their own particular research interests. As examples, an interesting site to learn about viruses is http://www.bocklabs.wisc.edu/Welcome.html, and an introduction to the disulfide bond-coupled folding pathway of bovine pancreatic trypsin inhibitor can be found at http://www.biology.utah.edu/People/regfaculty/~goldenberg/GoldenbergLab/research/bpti.html.

Index

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.

Page numbers in **bold** refer to a major text discussion; page numbers with an F refer to a figure; page numbers with a T refer to a table.