1
Sequence, Function, and Structure Relationships

1.1
Introduction

Life is the ability to metabolize nutrients, respond to external stimuli, grow, reproduce, and, most importantly, evolve. Most of these functions are performed by proteins, organic macromolecules involved in nearly every aspect of the biochemistry and physiology of living organisms. They can serve as structural material, catalysts, adaptors, hormones, transporters, regulators. Chemically, proteins are linear polymers of amino acids, a class of organic compounds in which a carbon atom (called Cα) is bound to an amino group ($–NH_2$), a carboxyl group (–COOH), a hydrogen atom (H), and an organic side group (called R). The physical and chemical properties unique to each amino acid result from the properties of the R group (Figure 1.1).

In a protein, the amino group of one amino acid is linked to the carboxyl group of its neighbor, forming a peptide (C–N) bond. There are two resonance forms of the peptide bond (i.e. two forms that differ only in the placement of electrons), as illustrated in Figure 1.2. Atoms involved in single bonds share one pair of electrons whereas two pairs are shared in double bonds. The latter are planar, i.e. not free to rotate. The resonance between the two forms shown in Figure 1.2 makes the peptide bond intermediate between a single and a double bond and, as a consequence, all peptide bonds in protein structures are found to be almost planar – the Cα, C, and O atoms of one amino acid and the N, H, and Cα of the next lie on the same plane. Although this rigidity of the peptide bond reduces the degrees of freedom of the polypeptide, the dihedral angles around the N–Cα and the Cα–C bonds are free to vary and their values determine the conformation of the amino acid chain.

Proteins assume a three-dimensional shape which is usually responsible for their function. The consequence of this tight link between structure and function and of the evolutionary pressure to preserve function has a very important effect – in contrast with ordinary polymers (e.g. polypeptides with random sequences) that typically form amorphous globules, proteins usually fold to unique structures. In other words they spontaneously assume a unique three-dimensional structure specified, as we will see, by their amino acid sequence. For example, enzymes accelerate chemical reactions by stabilizing their high energy intermediate and this

Protein Structure Prediction. Edited by Anna Tramontano

ISBN: 3-527-31167-X

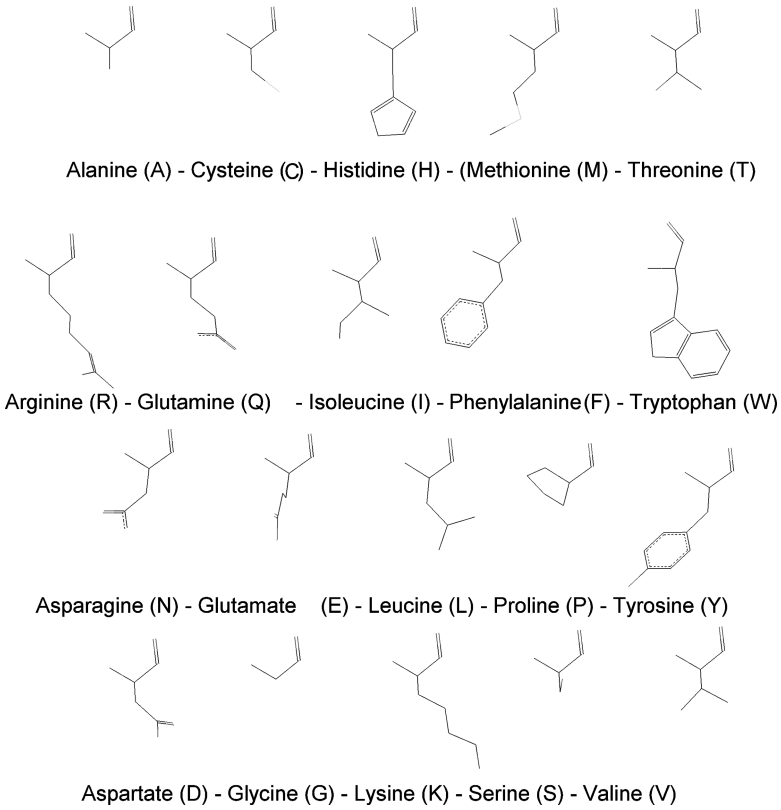


Figure 1.1 The twenty naturally occurring amino acids.

is achieved by correct relative positioning of appropriate chemical groups. Our body contains many proteins that catalyze the hydrolysis of peptide bonds in proteins (the inverse of the polymerizing reaction used to build proteins) to provide the body with a steady supply of amino acids. The substrates of these reactions are either proteins from the diet or "used" proteins inside the body. Digestion begins in the stomach where the acidic environment unfolds, i.e. destructures, the proteins and an enzyme called pepsin (Figure 1.3) starts chopping the proteins into pieces. Later, in the intestines, several protein-cutting enzymes, for example trypsin (also shown in Figure 1.3), cut the protein chains into shorter pieces. In subsequent steps, other enzymes reduce these shorter pieces to single amino acids.

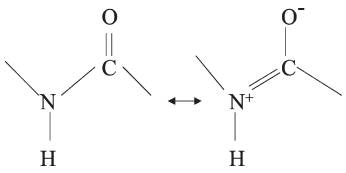
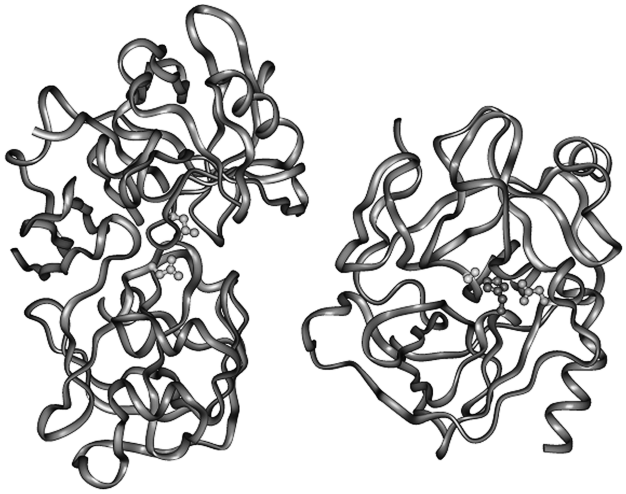


Figure 1.2 The two resonance structures of the peptide bond. Because of delocalization of the electrons, the C–N bond has the character of a partial double bond and this limits its freedom of rotation.



IGDEPLENYL DTEYFGTIGI GTPAQDFTVI FDTGSSNLWV
PSVYCSSLAC SDHNQFNPDD SSTFEATSQE LSITYGTGSM
TGILGYDTVQ VGGISDTNQI FGLSETEPGS FLYYAPFDGI
LGLAYPSISA SGATPVFDNL WDQGLVSQDL FSVYLSSNDD
SGSVVLLGGI DSSYYTGSLN WVPVSVEGYW QITLDSITMD
GETIACSGGC QAIVDTGTSL LTGPTSAIAN IQSDIGASEN
SDGEMVISCS SIASLPDIVF TINGVQYPLS PSAYILQDDD
SCTSGFEGMD VPTSSGELWI LGDVFIRQYY TVFDRANNKV
GLAPVA

IVGGYTCGAN TVPYQVSLNS GYHFCGGSLI NSQWVVSAAH
CYKSGIQVRL GEDNINVVEG NEQFISASKS IVHPSYNSNT
LNNDIMLIKL KSAASLNSRV ASISLPTSCA SAGTQCLISG
WGNTKSSGTS YPDVLKCLKA PILSDSSCKS AYPGQITSNM
FCAGYLEGGK DSCQGDSGGP VVCSGKLQGI VSWGSGCAQK
NKPGVYTKVC NYVSWIKQTI ASN

Figure 1.3 The three-dimensional structures of pepsin (left, PDB code: 1PSN) and trypsin (right, PDB code: 3TPI). These two enzymes cleave peptide bonds with a different mechanism. The first uses two aspartic acids, the second a triad formed by a histidine, a serine, and an aspartic acid. Their amino acid sequence is shown at the bottom of the figure in a one-letter code. Note that, for both enzymes, the amino acids forming the active site (underlined) are distant in the linear sequence and are brought together by the three-dimensional structure of the enzymes.

Both pepsin and trypsin belong to a class of enzymes called proteases, the first performs its job by taking advantage of the presence, in a cleft of the protein structure of two residues of aspartic acid; in the second catalysis is achieved by cooperation of three amino acids, a serine, a histidine, and an aspartic acid. Amino acids near the active site are responsible for recognition and correct positioning of the substrate. These functional amino acids, far apart in the linear amino acid sequence (Figure 1.3), are brought together in exactly the right position by the protein three-dimensional structure.

Similarly, recognition of foreign molecules is mediated by several proteins of the immune system, the most popular being antibodies. Antibodies bind other molecules, called antigens, by means of an exposed molecular surface complementary to the surface of the antigen, which can be a protein, a nucleic acid, a polysaccharide, etc. The binding surface is formed by amino acids from different parts of the molecule (Figure 1.4).

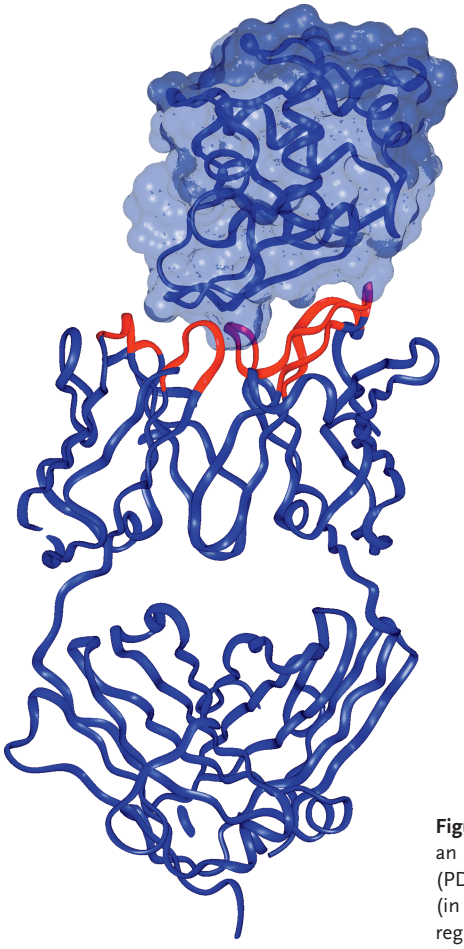


Figure 1.4 The three-dimensional structure of an antibody bound to its cognate molecule (PDB code: 3HFL). Note that the binding region (in red) is formed by amino acids from different regions of the linear sequence.

1.2 Protein Structure

Most readers will already be familiar with the basic concepts of protein structure; we will, nevertheless, review here some important aspects of this subject. The sequence of amino acids, i.e. of the R-groups, along the chain is called the primary structure. Secondary structure refers to local folding of the polypeptide chain. Tertiary structure is the arrangement of secondary structure elements in three dimensions and quaternary structure describes the arrangement of a protein's subunits. As we have already mentioned, the peptide bond is planar and the dihedral angle it defines is almost always 180°. Occasionally the peptide bond can be in the cis conformation, i.e. very close to 0°.

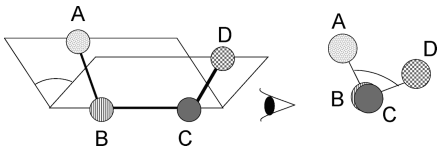


Figure 1.5 A dihedral angle between four points A, B, C, and D is the angle between two planes defined by the points A, B, C and B, C, D, respectively.

Question: What is a dihedral angle?

»The dihedral angle is the angle between two planes. In practice, if you have four connected atoms and you want to measure the dihedral angle around the central bond, you orient the system in such a way that the two central atoms are superimposed and measure the resulting angle between the first and last atom (Figure 1.5).«

The simplest arrangement of amino acids that results in a regular structure is the alpha helix, a right-handed spiral conformation. The structure repeats itself every 5.4 Å along the helix axis. Alpha helices have 3.6 amino acid residues per turn and

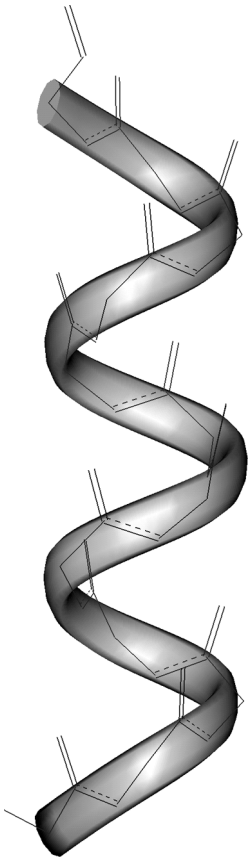


Figure 1.6 An alpha helix. Each backbone oxygen atom is hydrogen-bonded to the nitrogen of a residue four positions down the chain.

the separation of the residues along the helix axis is 5.4/3.6 or 1.5 Å, i.e. the alpha helix has a rise per residue of 1.5 Å. Every main-chain C=O group forms a hydrogen bond with the NH group of the peptide bond four residues away (Figure 1.6).

Let us recall that a hydrogen bond is an intermolecular interaction formed between a hydrogen atom covalently bonded to an electronegative atom (for example oxygen or nitrogen) and a second electronegative atom that serves as the hydrogen-bond acceptor. The donor atom, the hydrogen, and the acceptor atom are usually co-linear. The alpha helix has 3.6 residues per turn and thirteen atoms enclosed in the ring formed by the hydrogen bond, it can also be called a 3.6(13) helix. Another type of helix is observed in protein structures, although much more rarely; this is the 3(10) helix. This arrangement contains three residues per turn and ten atoms in the ring formed by the hydrogen bond. In alpha helices, the peptide planes are approximately parallel with the helix axis, all C=O groups point in one direction, and all N–H groups in the opposite direction. Because of the partial charge on these groups, negative for CO and positive for NH, there is a resulting dipole moment in the helix. Side-chains point outward and pack against each other. The dipoles of a 3(10) helix are less well aligned and the side-chain packing less favorable, therefore it is usually less stable. Typically, in alpha helices the angles around the N–Cα and Cα–C bonds, called ϕ and ψ angles, are approximately –60° and –50°, respectively.

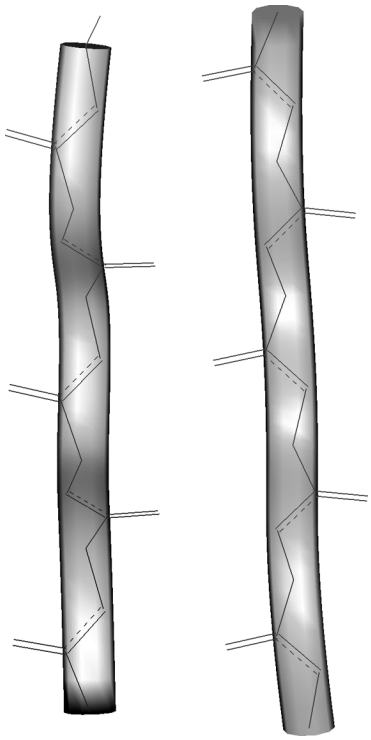


Figure 1.7 Two beta strands forming an antiparallel beta sheet. Oxygen and nitrogen atoms of different strands are hydrogen bonded to each other.

Another secondary structure element commonly observed in proteins is the beta sheet, an arrangement of two or more polypeptide chains (beta strands) linked in a regular manner by hydrogen bonds between the main chain C=O and N–H groups. The R groups of neighboring residues in a beta strand point in opposite directions forming a layered structure (Figure 1.7). The strands linked by the hydrogen bonds in a beta sheet can all run in the same direction (parallel sheet) or in opposite directions (antiparallel sheet). Beta sheets can be mixed, including both parallel and antiparallel pairs of strands. Most beta sheets found in proteins are twisted – each residue rotates by approximately 30° in a right-handed sense relative to the previous one.

The plot shown in Figure 1.8 is called a Ramachandran plot. This can be obtained by considering atoms as hard spheres and recording which pairs of ϕ and ψ angles do not cause the atoms of a dipeptide to collide. Allowed pairs of values are represented by dark regions in the plot whereas sterically disallowed regions are left white. The lighter areas are obtained by using slightly smaller radii of the spheres, i.e. by allowing atoms to come a bit closer together. Disallowed regions usually involve steric hindrance between the first carbon atom of the side-chain, the Cβ, and main-chain atoms. As we will see, the amino acid glycine has no side-chain and can adopt ϕ and ψ values that are unfavorable for other amino acids.

Question: Do we observe amino acids with dihedral angles in disallowed regions of the Ramachandran plot in experimental protein structures?

»Yes, we do. Even in very well refined crystal structures of proteins at high resolution, some ϕ and ψ angles fall into disallowed regions. The reader should keep in mind that the reason some combinations of angles are rarely observed is because they are energetically disfavored, not mathematically impossible. The loss of energy because of an unfavorable dihedral angles combination can be compensated by other interactions within the protein.«

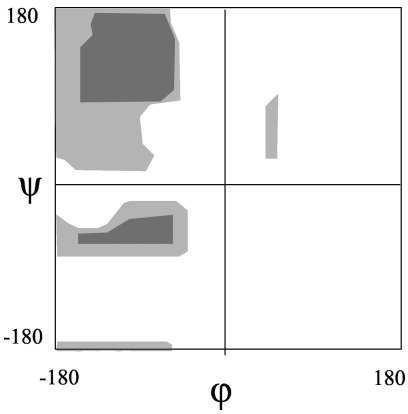


Figure 1.8 A Ramachandran plot is a graph reporting the values of phi and psi angles in protein structures. Darker areas indicate favorable combinations of angles, lighter gray areas are less favored, but still possible.

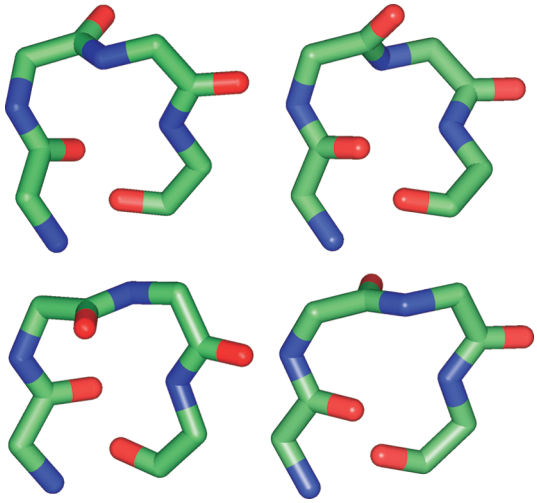


Figure 1.9 The four types of beta turn described in Table 1.1, types I and I′ are shown on the top, types II and II′ on the bottom.

Regions without repetitive structure connecting secondary structure elements in a protein structure are called loops. The amino acid chain can reverse its direction by forming a reverse turn characterized by a hydrogen bond between one main chain carbonyl oxygen and the N–H group 3 residues along the chain (Figure 1.9). When such a secondary structure element occurs between two anti-parallel adjacent beta strands in a beta sheet is called a beta hairpin. Reverse turns are classified on the basis of the ϕ and ψ angles of the two residues in their central positions as shown in Table 1.1. Note that some turns require that one of their amino acids has ϕ and ψ angles falling in disfavored regions of the Ramachandran plot.

Table 1.1 Turns are regions of the protein chain that enable the chain to invert its direction. The ϕ and ψ angles of some commonly occurring turns are listed.

Turn type	ϕ_1	ψ_1	ϕ_2	ψ_2
I	–60	–30	–90	0
I′	60	30	90	0
II	–60	120	80	0
II′	60	–120	–80	0

Proteins can be formed from only alpha helical or from only beta sheet elements, or from both; the association of these elements within a single protein chain is called tertiary structure. Certain arrangements of two or three consecutive secondary structures (alpha helices or beta strands), are present in many different protein

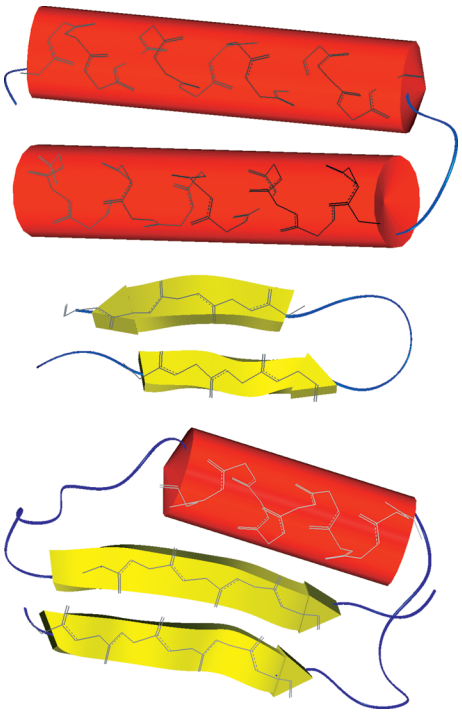


Figure 1.10 Supersecondary structures: alpha–loop–alpha, beta hairpin, and beta-alpha-beta unit.

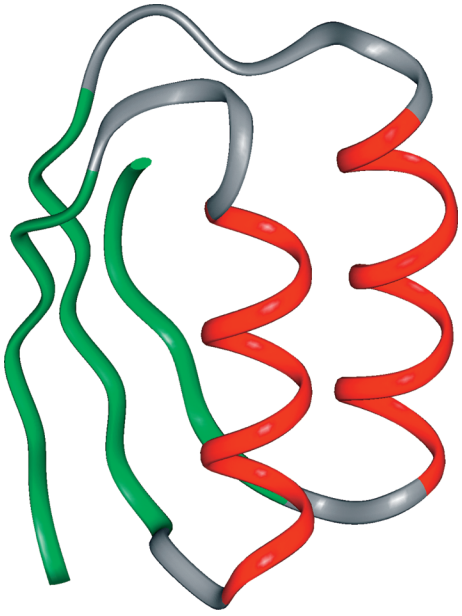


Figure 1.11 The Rossmann fold. The figure shows a region of the succinyl-Coa synthetase enzyme from the bacterium *Escherichia coli* (PDB code: 2SCU).

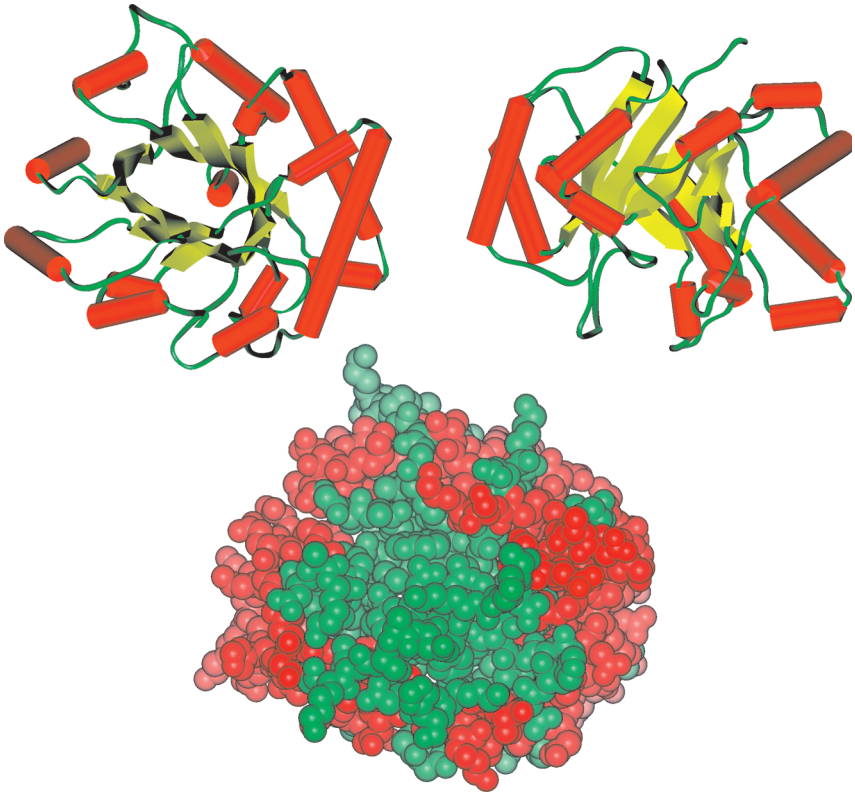


Figure 1.12 A TIM barrel (PDB code: 8TIM). The structure at the top left is the same as that at the top right rotated by 90° around an horizontal axis. On the bottom the structure is shown with all its non-hydrogen atoms. Atoms in green belong to the central beta barrel, atoms in red to the surrounding helices.

structures, even with completely different sequences; these are called supersecondary structures. They include the alpha–alpha unit (two antiparallel alpha helices joined by a turn); the beta–beta unit (two antiparallel strands connected by a hairpin); and the beta–alpha–beta unit (two parallel strands, separated by an alpha helix antiparallel to them (Figure 1.10). Sometimes the term "motif" is used to describe these supersecondary structures. Supersecondary structures are not necessarily present in a protein structure, however, which can be formed from several alpha helices or beta strands without containing any of the supersecondary structures described above. On the other hand, some combinations of the supersecondary structural motifs are observed relatively often in proteins. A very commonly found arrangement of helices is the four-helix bundle (two alpha–alpha units connected by a loop). Another common motif is the beta–alpha–beta–alpha–beta unit, called the Rossman fold (Figure 1.11). These arrangements are often called domains or folds. Some folds can be very large and complex and can be formed

from several supersecondary structures. One example is the TIM barrel fold; this is shared by many enzymes and formed from several beta–alpha–beta units (Figure 1.12).

Another layer of organization of protein structure is the domain level. The definition of a domain is rather vague. Some confusion also arises because the term is often also used in the context of the amino acid sequence, rather than of its three-dimensional structure. In general a domain can be defined as a portion of the polypeptide chain that folds into a compact semi-independent unit. Domains can be seen as "lobes" of the protein structure that seem to have more interaction between themselves than with the rest of the chain (Figure 1.13). Several proteins are formed from many repeated copies of one or a few domains; such proteins are called mosaic proteins and the domains are often referred to as "modules". A domain can be formed by only (or almost only) alpha helices or beta sheets, or by their combination. In the latter case the helices and strands can be packed against each other in the beta–alpha–beta supersecondary arrangement (alpha/beta domains) or separated in the structure (alpha + beta domain).

Finally, we talk about architecture of a protein when we consider the orientations of secondary structures and their packing pattern, irrespective of their sequential order, and we talk of protein topology when we also take into account the nature of the connecting loops and, therefore, the order in which the secondary structure elements occur in the amino acid sequence.

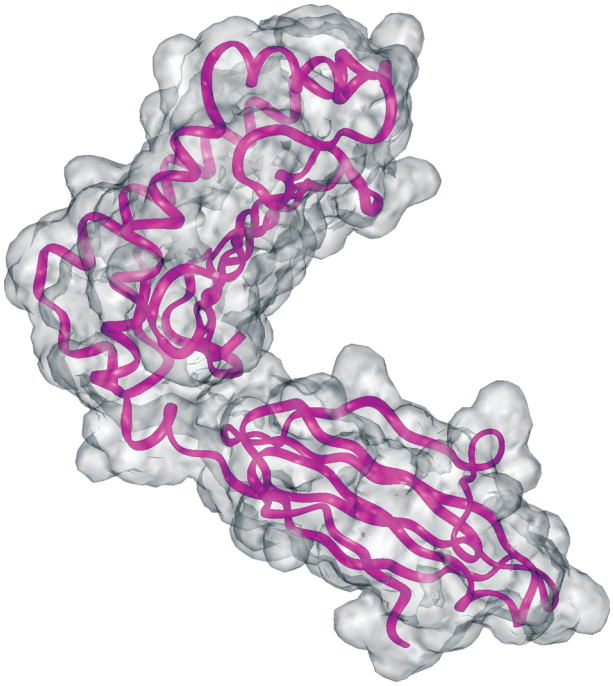


Figure 1.13 A two-domain protein chain (PDB code: 1HSA).

1.3
The Properties of Amino Acids

There are twenty naturally occurring amino acids. They can play different roles and it is important to survey their properties to be able to analyze and ultimately attempt to predict the structure and function of a protein.

The smallest amino acid is glycine, the side-chain of which is just a hydrogen atom. The lack of a side-chain makes this amino acid very flexible. We have already mentioned that this amino acid can assume "unusual" ϕ and ψ angles. We also saw that the structural requirements of turns often need an amino acid in this conformation and indeed these positions are often occupied by glycines. The observation that a glycine is always present in a given position in a family of evolutionarily related proteins often points to the presence of a tight turn in the region. The flexibility of glycine also implies that the loss of entropy associated with restricting its conformation in a protein structure is higher than for other amino acids, and the absence of a side-chain makes it less likely for this amino acid to establish favorable interactions with surrounding amino acids. Glycines are, therefore, rarely observed in both alpha helices or beta sheets.

The next amino acid, in order of size, is alanine. Here the side-chain is a CH_3 group. It is a small hydrophobic amino acid, without any reactive group, and rarely involved in catalytic function. Its small non polar surface and its hydrophobic character suggest, however, that this amino acid can be exposed to solvent, without large loss of entropy, and can also establish favorable hydrophobic interactions with other hydrophobic surfaces. In other words, it is an ideal amino acid for participating in interacting surfaces between proteins that associate transiently.

Cysteine is a small hydrophobic amino acid that can form disulfide bridges, i.e. covalent bonds arising as a result of the oxidation of the sulfhydryl (SH) group of the side-chains of two cysteine units when they are in the correct geometric orientation (Figure 1.14). Disulfide bridges enable different parts of the chain to be covalently bound. Because the intracellular environment is reducing, disulfide bridges are only observed in extracellular proteins. Cysteine can also coordinate metals and its SH group is rather reactive. In some viral proteases it takes the role of serine in serine protease active sites we have already described.

Serine is a small polar amino acid found both in the interior of proteins and on their surfaces. It is sometimes found within tight turns, because of its small size and its ability to form a hydrogen bond with the protein backbone. It is often

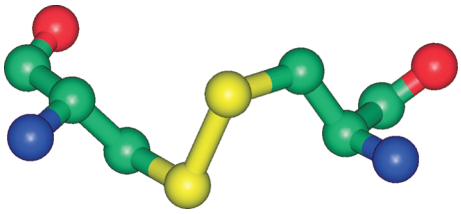


Figure 1.14 A disulfide bond. The yellow atoms are sulfur atoms.

observed in active sites, where it can act as a nucleophile as already mentioned for serine proteases. Another important property of this amino acid is that it is a substrate for phosphorylation – enzymes called protein kinases can attach a phosphate group to its side-chain. This plays important roles in many cellular processes and in signal transduction.

Another relatively small amino acid, rather similar to serine, is threonine. This amino acid can also be part of active sites and can be phosphorylated. An important difference, though, is that threonine is "beta branched", i.e. it has a substituent on its beta carbon and this makes it less flexible and less easy to accommodate in alpha helices. Beta-branched amino acids are indeed more often found in beta sheets.

Asparagine and glutamine are polar amino acids that generally occur on the surface of proteins, exposed to an aqueous environment, and frequently involved in active sites. Asparagine, for example, is found as a replacement for aspartate in some serine proteases. One peculiar property of asparagine is that it is often found in the left-handed conformation (positive ϕ and ψ angles) and can therefore play a role similar to that of glycine in turns. This is possibly because of its ability to form hydrogen bonds with the backbone.

Proline is unique because it is an imino acid rather than an amino acid. This simply means that its side-chain is connected to the protein backbone twice, forming a five-membered nitrogen-containing ring. This restricts its conformational flexibility and makes it unable to form one of the two main-chain hydrogen bonds that other amino acids can form in secondary structure elements; it is, therefore, often found in turns in protein structures. When it is in an alpha helix, it induces a kink in the axis of the helix. It is not a very reactive amino acid, but plays an important role in molecular recognition – peptides containing prolines are recognized by modules that are part of many signaling cascades. Proline can be found in the cis conformation (i.e. with the angle around the peptide bond close to 0° rather than 180°). The main chain nitrogen atoms of the other amino acids are bound to a hydrogen and a carbon atom whereas the situation in proline is more symmetrical with the atom bound to two carbon atoms. This means that the energy difference between the *cis* and *trans* conformations is smaller for this amino acid.

Leucine, valine, and isoleucine are hydrophobic amino acids, very rarely involved in active sites. The last two are beta branched and therefore often found in beta sheets and rarely in alpha *helices*.

Aspartate and glutamate are negatively charged amino acids, generally found on the surface of proteins. When buried, they are involved in salt bridges, i.e. they form strong hydrogen bonds with positively charged amino acids. They are frequently found in protein active sites and can bind cations such as zinc.

Lysine and Arginine are positively charged and can have an important role in structure. The first part of their side-chain is hydrophobic, so these amino acids can be found with part of the side-chain buried, and the charged portion exposed to solvent. Like aspartate and glutamate, lysine and arginine can form salt bridges and occur quite frequently in protein active or binding sites. They are, furthermore, often involved in binding negatively charged phosphates and in the interacting surfaces of DNA- or RNA-binding proteins.

At physiological pH, histidine can act as both a base or an acid, i.e. it can both donate and accept protons. This is an important property that makes it an ideal residue for protein functional centers such as the serine protease catalytic triad. Histidine can, furthermore, bind metals (e.g. zinc). This property is often exploited to simplify purification of proteins cloned and expressed in heterologous systems. The addition of a tail of histidines to the protein of interest confers on the protein the ability to chelate metals and this engineered property can be exploited for purifying the protein.

Methionine has a long and flexible hydrophobic side-chain. It is usually found in the interior of proteins. Like cysteine, it contains a sulfur atom, but in methionine the sulfur atom is bound to a methyl group, which makes it much less reactive.

Phenylalanine, tryptophan, and tyrosine are aromatic amino acids. The term "aromatic" was used by chemists to describe molecules with peculiar odors long before their chemical properties were understood. In chemistry, a molecule is called aromatic if it has a planar ring with $4n + 2$ π-electrons where n is a non-negative integer (Hückel' s Rule). In practice, these molecules, the prototype of which is benzene, have a continuous orbital overlap that gives them special optical properties. For example, tryptophan absorbs light at 280 nm and this property is routinely used to measure the concentration of proteins in a solution (assuming the protein contains at least one tryptophan). Also, if an aromatic residue is held rigidly in space in an asymmetric environment, it absorbs left-handed and right-handed polarized light differently. This effect, which can be measured by circular dichroism spectroscopy, is therefore sensitive to overall three-dimensional structure and can be used to monitor the conformational state of a protein. Another important property of amino acids with aromatic side-chains is that they can interact favorably with each other. The face of an aromatic molecule is electron-rich while the hydrogen atoms around the edge are electron-poor. This implies that off-set face-to-face and edge-to-face interactions between aromatic rings have both hydrophobic and electrostatic components. Tyrosine is also a substrate for phosphorylation, similarly to serine and threonine, although the enzymes responsible for phosphorylation of tyrosine are different from those that phosphorylate serine and threonine.

1.4 Experimental Determination of Protein Structures

Two experimental techniques are used to determine the three-dimensional structure of macromolecules at the atomic level – X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy. Although it is beyond the scope of this book to describe the details of these techniques, which are rather complex both theoretically and experimentally, it is important to have some basic understanding of their results, because, as we will see, most methods for prediction of protein structure are based on existing structural data.

X-ray crystallography is based on the fact that an ordered ensemble of molecules arranged in a crystal lattice diffracts X-rays (the wavelengths of which are of the order of interatomic distances) when hit by an incident beam. The X-rays are dispersed by the electrons in the molecule and interfere with each other giving rise to a pattern of maxima and minima of diffracted intensities which depends upon the position of the electrons (and hence of the atoms) in the ordered molecules in the crystal. The electron density of the protein, i.e. the positions of the protein atoms, determines the diffraction pattern of the crystal, that is the magnitudes and phases of the X-ray diffraction waves, and vice versa, through a Fourier transform function. In practice:

$$\varrho(x,y,z) = \frac{1}{V}\sum_{hkl}\vec{F}_{hkl} = \frac{1}{V}\sum_{h}\sum_{k}\sum_{l}F(h,k,l)e^{-2\pi i(hx+ky+lz)}$$

where $\varrho(x,y,z)$ is the electron density at position (x,y,z), V is the volume, $\vec{F}(h,k,l)$ is the vector describing the diffracted waves in terms of their amplitudes $F(h,k,l)$ and phases (the exponential complex term). The electron density at each point depends on the sum of all of the amplitudes and phases of each reflection. If we could measure the amplitude and phase of the diffracted waves, we could relatively easily compute the exact relative location of each atom in the diffracting molecules. Unfortunately the phase of the diffracted waves cannot be measured and, therefore, we must use "tricks" to guess their approximate value and reconstruct the image of the diffracting molecule.

In effect, three methods are used to estimate the phases. Direct methods consist in using all possible values for the phases in the Fourier transform equation until an interpretable electron density is found; this is feasible for small molecules only. Interference-based methods can make use of multiple isomorphous replacement or anomalous scattering techniques. The first derives the phase by comparing the diffraction pattern of a protein crystal with that of crystals identical to the original one but for the presence of "heavy" atoms (i.e. atoms with many electrons and, therefore, very strong diffracting power) in specific positions of the molecules. The "anomalous scattering" technique instead derives initial phases by measuring diffraction data at several different wavelengths near the absorption edge of a heavy-atom. Finally, if we have a reasonable model for the molecule in the crystal, we can resort to the "molecular replacement" technique which computes approximated phases for the molecule in the crystal on the basis of the position of the atoms in the model. The availability of a high-quality three-dimensional model for a protein can therefore also be instrumental in obtaining its experimentally determined structure.

Given the diffracted intensities of a protein crystal and a set of "good" estimated phases, we can calculate the electron density that formed the observed pattern and position the atoms of the protein in the computed electron density (Figure 1.15). Important aspects of the whole procedure are that the protein under examination forms well ordered, well diffracting crystals and that the phase estimation procedure is successful in generating an interpretable electron-density map.

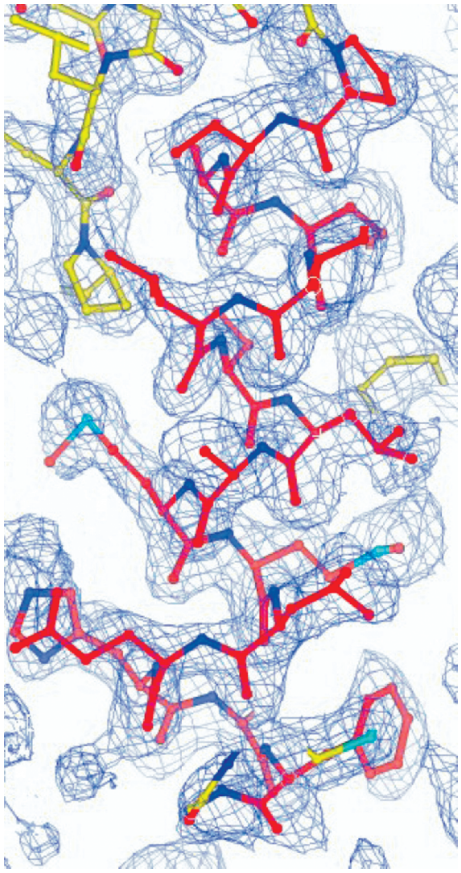


Figure 1.15 An electron-density map derived from an X-ray crystallography experiment. The atoms can be positioned in the map as shown in the figure, revealing an alpha helical structure.

Question: Is the quality of an X-ray determination of a protein structure comparable to that for small organic molecules?

»The quality of the structural data that can be obtained by protein crystallography is nowhere near the accuracy with which crystal structures of small molecules can be determined. This is because proteins can assume many different, although closely related, conformations and this limits the order of the molecules within the crystal. Also, protein crystals are usually only about half protein – the other half is occupied by solvent molecules. As we will see, the accuracy of small molecule crystallography can be used to derive parameters useful in modeling procedures.«

Just as in every experiment, in protein crystallography also the quality of the results improves with the ratio of the amount of data collected (the diffraction intensities) and the number of properties estimated (the positions of the atoms). In crystallog-

raphy, the inverse of this ratio is expressed by the term "resolution" which is expressed in Angstroms (1 Angstrom = 10^{-10} m). The lower is the resolution the better is the quality of the structure. A resolution of approximately 3Å enables secondary structural elements and the direction of the polypeptide chain to be clearly identified in the electron density map; with a resolution of 2.5Å the side-chains can be built into the map with reasonable precision.

Hydrogen atoms do not diffract very well, because they only have one electron, and they are usually not detectable by X-ray crystallography unless the resolution is really very good, approximately 1.0Å. This implies that, given a crystal structure with good but not exceptional resolution, we can only deduce the presence of hydrogen bonds by the position of the donor and acceptor atoms.

After reconstructing the structure, we can back compute the expected diffraction pattern and compare it with that observed. The *R* factor indicates how much the two patterns (theoretical and experimental) differ and is expressed as a percentage. This factor is linked to the resolution. As a rule of thumb, a good structure should have an *R* factor lower than the resolution divided by 10 (i.e. ≤30% for a 3.0Å resolution structure, ≤20% for a 2.0Å structure, etc). To avoid any bias, it is more appropriate to compare the expected data with data set aside and not used to reconstruct the structure. In this case the term is called "*R*free". For a correctly reconstructed structure, one expects the ratio *R*/*R*free to be >80%.

Of course atoms in a crystal also have thermal motion. We can estimate the extent of their motion by looking at their electron density and, indeed, crystallography assigns a value that describes the extent to which the electron density is spread out to each atom. This value, called the "temperature factor" or "Debye–Waller factor" or *B* factor is given by:

$$B = 8\pi^2 U^2$$

where *U* is the mean displacement of the atom (in Å), so high B factors indicate greater uncertainty about the actual atom position. For example, for $B = 20$ Å^2:

$$U = \frac{1}{\pi}\sqrt{\frac{B}{8}}\text{Å} = \frac{1}{\pi}\sqrt{\frac{20}{8}}\text{Å} \cong \frac{1}{3.14}\sqrt{2.5}\text{Å} \cong 0.5\text{Å}$$

and the uncertainty about the position of the atoms is 0.5 Å. Values of 60 or greater may imply disorder (for example, free movement of a side-chain or alternative side-chain conformations). As expected, atoms with higher *B* factors are often located on the surface of a protein whereas the positions of the atoms in the internal well packed core of the protein are less uncertain (Figure 1.16). Finally, the occupancy value for an atom represents the fraction of expected electron density that was actually observed in the experiment.

Nuclear magnetic resonance (NMR) is another very useful technique for determining the structure of macromolecules. This technique is based on the observation that several nuclei (e.g. H, ^{13}C, ^{15}N) have an intrinsic magnetic moment. If we place a concentrated homogeneous solution of a protein (or nucleic acid) inside a

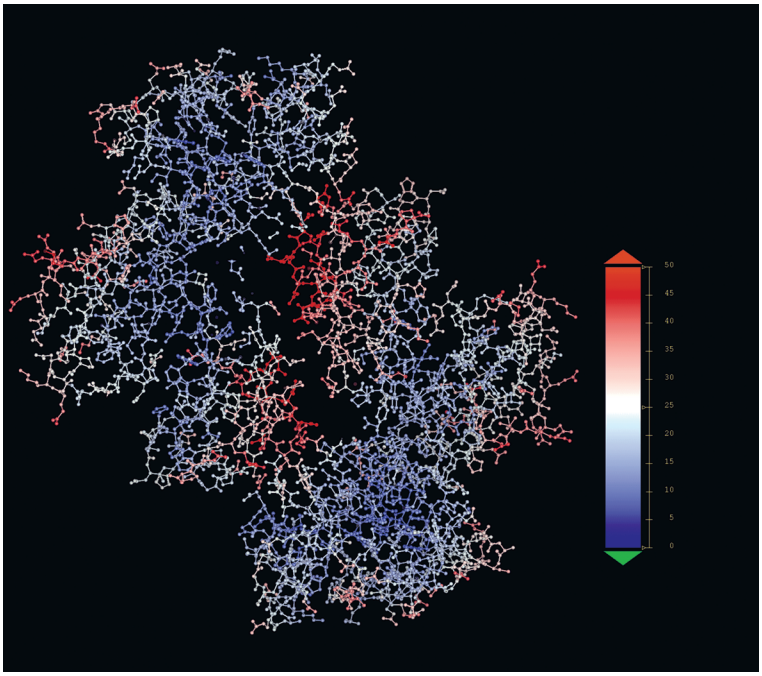


Figure 1.16 A protein structure colored according to the B-factor of its atoms. The color scheme is such that atoms with high B-factors are red and those with low B-factors blue.

very powerful magnetic field, the spin of the nuclei will become oriented in the direction of the external field. By applying radio-frequency magnetic fields to the sample, we can measure the energy absorbed at the frequency corresponding to the jump between two allowed spin orientations. Each atom has a characteristic resonance which depends on its structure, but it is also affected by the surrounding atoms. These subtle absorbance differences between the same atom in different environments make it possible to identify which resonance corresponds to which of the protein atoms.

If two atoms are close in space, magnetic interactions between their spins can be measured. The intensity of the interaction decays rapidly with the distance between the atoms (it is a function of r^{-6}, where r is the distance). This effect can be exploited to map short distances between interacting atoms. The result of the experiment is a set of lower and upper limits for the distance between pairs of atoms (constraints). If the number of constraints is sufficient, there will be a finite number of possible conformations of the protein compatible with the data. The more constraints we are able to measure, the more similar to each other will these structures be (Figure 1.17).

The number of constraints in an NMR experiment is strongly dependent on the flexibility of the protein and of its regions in solution: if a given region is very mobile, it will be very difficult to identify the neighbors of its atoms because they

will not spend enough time next to each other. In such cases, we cannot measure the interactions but we recover very valuable information about the intrinsic mobility of the protein structure.

Question: How do I evaluate the quality of an NMR structure and how does it compare with X-ray structure?

»NMR structures are usually reported with *rmsd* values (the square root of the average sum of the squared distances between corresponding atoms, see later) for the various structures compatible with the data. The lower the *rmsd*, the

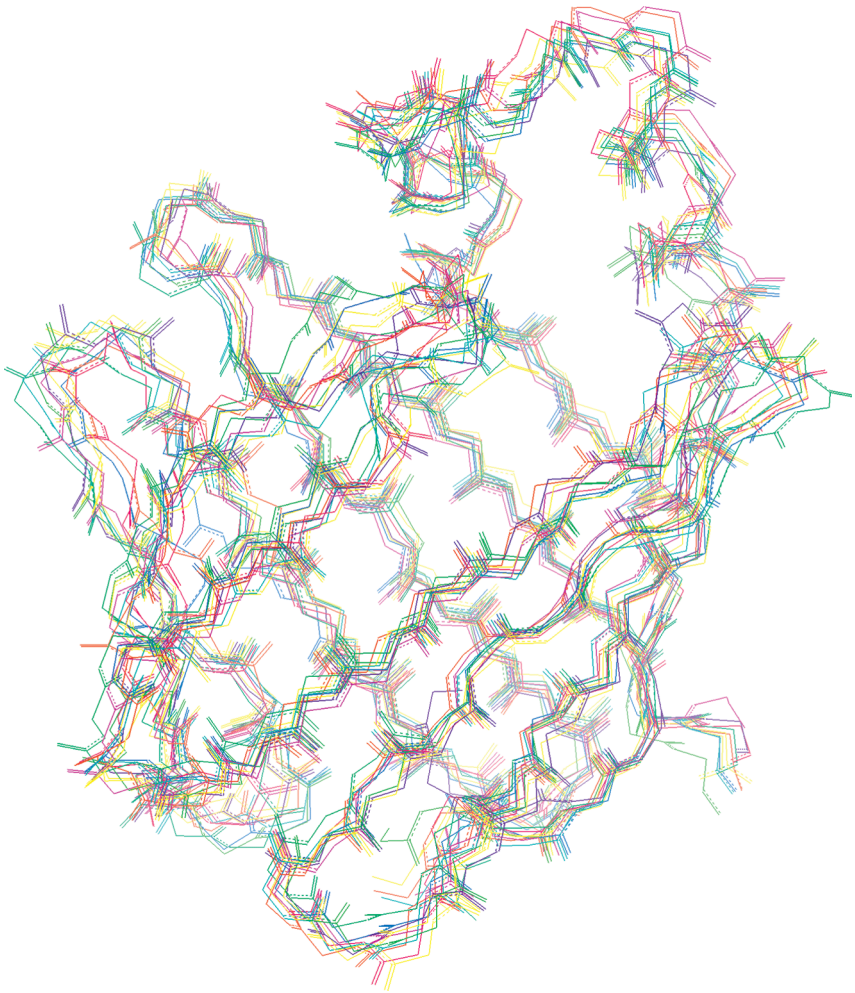


Figure 1.17 Several NMR-derived structures of the chicken fatty acid-binding protein. Note that the exposed regions are less well defined than the central core of the protein.

higher the accuracy of the measurements. The answer to the second part of the question depends on whether the question is posed to a crystallographer or to an NMR spectroscopist! There is no clear and definite answer because the two experiments give different, albeit related, information.«

Question: Does the crystal structure of a protein reflect its "true" native and functional structure?

»This question is often asked. Several lines of evidence point to a positive answer – structures of the same protein solved by both X-ray crystallography and NMR, or solved independently in different crystal forms, are the same within the experimental error. Furthermore, protein crystals are full of solvent (and for this reason very fragile) and it has often been shown that crystallized enzymes can function inside the crystal; they are therefore deemed to have the correct native functional structure.«

1.5 The PDB Protein Structure Data Archive

Structures determined by both X-ray and NMR are deposited in a data base called PDB. X-ray structure entries consist of a single structure; for NMR entries there is a variable number of structures, usually approximately 20, compatible with the data. Each entry is uniquely identified by a four-letter code. In the first part of a PDB entry there are the name of the molecule, the biological source, some bibliographic references, and the R and Rfree factors. There is also information about how chemically realistic the model is, i.e. how well bond lengths and angles agree with expected values (the values found in small molecules). For a good model, average deviations from expected values should be no more than 0.2 Å in bond lengths and 4° in bond angles. The SEQRES records contain the amino acid or nucleic acid sequence of residues in each chain of the macromolecule studied, whereas the HELIX, SHEET and TURN records list the residues where secondary structure elements begin and end and their total length.

Question: Where can I find the sequences of all the proteins of known structure?

»There is also a database of the sequences of known structures, usually called pdb, containing the sequences extracted from the SEQRES records. Be aware, however, that even if some parts of the protein are not visible in the electron density map, because they are too mobile or because the protein was partially degraded in the experiment, their sequence will still be included in the SEQRES field. In other words the sequence

corresponds to that of the studied molecule, not necessarily to the part of the molecule the structure of which is contained in the entry. The database of sequences of known structure called ASTRAL only includes the sequence of the part of the molecule that has been experimentally determined.«

After this initial part of the file, the actual coordinates are listed in records identified by the keyword ATOM. These include a serial number for the atom, the atom name, the alternative location indicator, used when the electron density for the atom was observed in two positions, the chain identifier, a residue sequence number and code, the x, y, and z orthogonal coordinates for the atom, the occupancy, and the temperature factor. For example the record:

```
ATOM 1281 N GLY Z 188A 29.353 66.969 17.508 1.00 28.84
```

describes the nitrogen atom of a glycine unit with residue number 188 and residue code A. The coordinates are $x = 29.353$, $y = 66.969$, $z = 17.508$. The occupancy is 1.00 (i.e. complete) and the B factor is 28.84 (corresponding to an uncertainty in the position of this atom of 0.6 Å).

Question: Which is the minimum occupancy of atoms reported in a PDB file?

»There is no lower limit to the value of the occupancy for an atom. It can be 0 if the position of an atom was guessed on the basis of the positions of the surrounding atoms. Be aware that none of the widely used structure-visualization packages highlights them automatically. It is always advisable, if one is working on a particular region of a protein, to verify the B factor and occupancy of its atoms.«

It is worth briefly describing the residue number and code, because these are often the cause of much frustration when trying to use a PDB file: the residue number is not necessarily consecutive. For example, trypsin is synthesized as a longer molecule the first 15 amino acids of which must be enzymatically removed to produce the active protein. The first residue number in the 3PTI entry for trypsin is indeed 16. A common numbering scheme is occasionally used for a family of evolutionarily related proteins, and in such circumstances the residue numbering follows the scheme. If one of the proteins of the family contains amino acids inserted among the commonly accepted numbering, the residue code is given a letter. In the 3TPI entry, for example, we find:

```
ALA 183
GLY 184A
TYR 184
LEU 185
GLU 186
```

GLY 187
GLY 188A
LYS 188

For NMR structures the headers do not, of course, include the R factors and the resolution. The ATOM fields are quite similar, the B factor is usually set to 0 and the sections referring to each of the models are included between the records MODEL and ENDMDL.

1.6 Classification of Protein Structures

Protein structures can be classified according to their similarity, in terms of secondary structure content, fold, and architecture. There are a few widely used

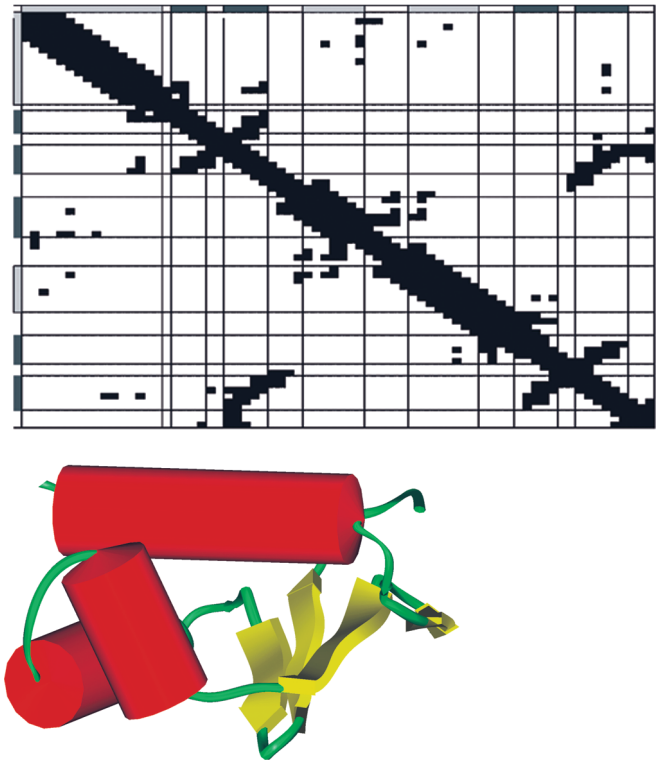


Figure 1.18 A distance map for the domain shown at the bottom (PDB code: 1RUN). The secondary structure of the protein is shown in the first line and first column of the matrix. Black and gray regions correspond to beta strands and alpha helices, respectively. Filled cells correspond to distances shorter than 6 Å.

classifications of protein structures which are extremely useful for navigating through the protein structural space. These are collected and made available to the community via Web servers and differ in the method used to obtain the classifications.

FSSP is a classification method based on comparison of the "distance matrices" of proteins. These are an alternative representation of protein structures (Figure 1.18) obtained by filling a matrix. Each row and each column of a matrix represent an amino acid and each cell contains the distance between the amino acid in the row and the amino acid in the column in the protein structure. Given two proteins, we can compare their distance matrices and derive a structural superposition between their atoms, i.e. the superposition that minimizes the distance between corresponding pairs of atoms. The resulting structural distance between the two proteins, defined as the root mean square of the average sum of the squared distances, is used by FSSP to cluster the known structures and to classify them.

CATH is another classification of protein structures based on use of a different algorithm to compute structural similarity. In this classification the two distance matrices that are compared contain the vectorial distance between pairs of atoms, rather than the scalar one. CATH provides a hierarchical classification of the structures, identifying four levels of similarity – Class, Architecture, Topology, and Homology. The Class is defined on the basis of the predominant type of secondary structure (all alpha, all beta, alpha and beta, and domains with little or no secondary structure). The Architecture describes the overall shape of the domain structure as determined by the orientations of the secondary structures ignoring the connectivity between the secondary structures. It is assigned on the basis of visual inspection of the proteins and of literature data. The Topology level depends on the structural distance between proteins, and evolutionarily related proteins are grouped at the Homology level, on the basis, as we will see, of sequence-based methods.

Finally, SCOP is another classification with a hierarchical organization including Class, Fold, Superfamily, and Family levels. The main Class types in SCOP are all alpha, all beta, alpha plus beta, and alpha/beta. A protein is assigned to one of the classes according its predominant secondary structure. The other classes include multi-domain, membrane, and cell surface proteins and peptides, small proteins, peptides, designed proteins, and low-resolution structures. The second level of classification, Fold, includes proteins with similar topological arrangements for which an evolutionary relationship cannot be identified, the third level (Superfamily) includes proteins that are believed to share a common ancestor. Proteins related by an unambiguous evolutionary relationship are grouped at the Family level. The classification in SCOP is essentially manual, although some automatic pre-processing is used to cluster clearly similar proteins.

None of these classifications is intrinsically better than any other and they usually agree with each other.

1.7 The Protein-folding Problem

The stability of each possible conformation of an amino acid chain depends on the free energy change between its folded and unfolded states:

$$\Delta G = \Delta H - T\Delta S$$

where ΔG, ΔH, and ΔS are the differences between the free energy, enthalpy, and entropy, respectively, of the folded and unfolded conformations. The enthalpy difference is the energy associated with atomic interactions within the protein structure (dispersion forces, electrostatic interactions, van der Waals potentials, and hydrogen bonding that we will describe in more detail later) whereas the entropy term describes hydrophobic interactions. Water tends to form ordered cages around non-polar molecules, for example the hydrophobic side-chains of an unfolded protein. On folding of the polypeptide chain, these groups become buried within the protein structure and shielded from the solvent. The water molecules are more free to move and this leads to an increase in entropy that favors folding of the polypeptide.

Question: What does an unfolded protein looks like?

»Although we generally assume that the unfolded chain is in a random coil conformation, i.e. that the angles of rotation about the bonds are independent of each other and all conformations have comparable free energies, the reader should be aware that, in reality, unfolded proteins tend to be less disordered and more compact than ideal random coils, because some regions of the polypeptide can interact more favorably with each other than with the solvent.«

In a cell proteins are synthesized on ribosomes, large molecular assemblies comprising proteins and ribonucleic acid molecules. Special adaptor molecules, tRNA molecules, recognize a triplet of bases on the messenger RNA, which in turn has been synthesized by following instructions contained in the genome, and adds the appropriate amino acid to the nascent chain. The synthesis of an average protein takes approximately a minute; the time required for folding, i.e. for achieving the "working" native structure, is comparable. Some slow steps of the reaction, for example formation of disulfide bonds, are accelerated by specific enzymes. Other proteins are also involved in the folding process and their role is either to protect the nascent protein chain (shielding the hydrophobic regions that are exposed to solvent before folding occurs) or to provide a more protected environment for folding; there is no evidence that anything but the amino acid sequence determines the native protein structure in vivo.

In the nineteen-sixties the American chemist Christian Anfinsen and his co-workers performed a series of seminal experiments demonstrating that the native

conformation of a protein is adopted spontaneously or, in other words, that the information contained in the protein sequence is sufficient to specify its structure. The enzyme selected by Anfinsen for the experiment was ribonuclease A (RNase A), an extracellular enzyme of 124 residues with four disulfide bonds (Figure 1.19). As already mentioned, these are covalent bonds arising as a result of oxidation of the sulfhydryl (SH) groups of the side-chains of two cysteines, when they are close to each other. The result is an S–S bond between their sulfur atoms. In Anfinsen' s experiment, the S–S bonds were first reduced to eight –SH groups (by use of mercaptoethanol, a reducing agent with the chemical formula HS–CH_2–CH_2–OH); the protein was then denatured by adding urea in high concentration (8 Molar). (The urea molecule enhances the solubility of nonpolar compounds in water and therefore reduces the strength of the stabilizing hydrophobic interactions that hold the protein structure together.) Under these conditions the enzyme is inactive and becomes a flexible random polymer. In the second phase of the experiment the urea was slowly removed (by dialysis); the –SH groups were then oxidized back to S–S bonds. We expect that if the protein is able to assume its

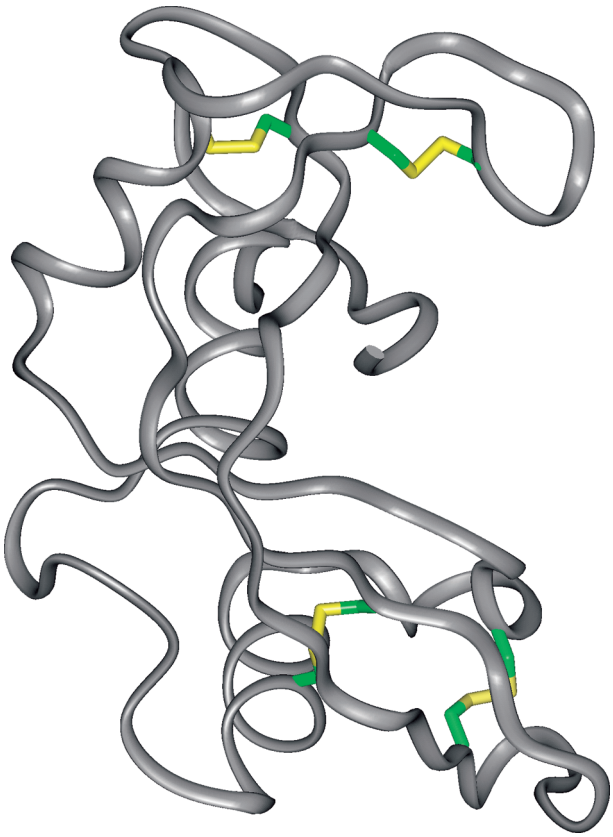


Figure 1.19 The structure of ribonuclease A (PDB code: 1AFK). Note the four disulfide bridges.

correct tertiary structure, the correct pairs of cysteines are close to each other so that the correct disulfide bonds form and the protein regains its activity. Indeed, the refolded protein regained more than 90 % of the activity of the untreated enzyme.

> *Question: How can we be sure that the protein was really unfolded after adding urea and mercaptoethanol?*
>
> »Anfinsen and co-workers also performed a control experiment to demonstrate that RNase A was completely unfolded in 8 Molar urea. RNase A was reduced and denatured as above, but in the second phase the enzyme was first oxidized to form S-S bonds, and only afterwards was the urea removed. If the protein is really in a random conformation in 8 Molar urea, it is likely that the cysteines are in different relative positions in different molecules and will randomly pair giving rise to scrambled sets of disulfide bonds. Because there are eight cysteine residues in ribonuclease, there are $7 \times 5 \times 3 \times 1 = 105$ different ways of forming disulfide bonds, only one of which is correct and leads to the formation of a functional enzyme. The experiment indeed showed that only about 1 % of the activity could be recovered in this control experiment. Later the same experiment was successfully repeated using a chemically synthesized protein chain, i.e. a protein that had never seen a cell or a ribosome.«

These experiments demonstrated that proteins can, indeed, adopt their native conformation spontaneously, but immediately raised a fundamental problem known as the Levinthal paradox – if the same native state is achieved by various folding processes both in vivo and in vitro, we must conclude that the native state of a protein is thermodynamically the most stable state under "biological" conditions, i.e. the state in which the interactions between the amino acids of the protein are the most energetically favorable compared with all other possible arrangements the chain can assume. But an amino acid chain has an enormous number of possible conformations (at least 2^{100} for a 100-amino-acid chain, because at least two conformations are possible for each residue). It can be computed that the amino acid chain would need at least ~2^{100} ps, or ~10^{10} years to sample all possible conformations and find the most stable structure.

Levinthal concluded that a specific folding pathway must exist and that the native fold is simply the end of this pathway rather than the most stable chain fold. In other words, Levinthal concluded that the folding process is under kinetic rather than under thermodynamic control and that the native structure corresponds not to the global free energy minimum but rather to one which is readily accessible. The hypothesis underlying Levinthal' s reasoning is that the energetically favorable contacts that stabilize the structure arise only when the chain is folded or nearly folded. In other words, the protein chain must first lose all its entropy (being locked in a given conformation) and, only when the correct conformation is reached, can

the entropy loss be compensated by the gain in enthalpy. A wealth of literature addresses the Levinthal paradox and we will not dwell on the details here, except to say that, in general, the paradox can be solved by thinking of the folding process as a sequential process in which the entropy decrease is immediately or nearly immediately compensated by an energy gain and that, in this hypothesis, the time-scale computed for the folding process approximates that observed in nature.

1.8
Inference of Function from Structure

The Structural Genomics Initiatives promise to deliver between 10 000 and 20 000 new protein structures within the next few years and, as we will see in this book, many more protein structures will be modeled. The challenge is obviously to exploit this large amount of structural data to predict the functions of these proteins. Proteins sharing a common evolutionary origin (homologous proteins) have similar structures, as we will see shortly, and, occasionally, proteins that do not seem to share an evolutionary relationship might turn out to share the same topology. One can expect to gain insight into a protein's function from analysis of other, structurally similar, proteins. There are at least three difficulties to be overcome in this process:

- homologous proteins might have originated by gene duplication and subsequent evolution and therefore have acquired a different function;
- some folds are adopted by proteins performing a variety of function; and, finally,
- the protein of interest might have a novel, not yet observed, fold.

What can we learn from the analysis of a protein structure? We can certainly identify which residues are buried in the core of the protein and which are exposed to solvent. The structure will also tell us the quaternary structure of the protein – the structure observed in the crystal is often that which is biologically active, although there are exceptions that might create difficulties.

The presence of local structural motifs with functional roles can be detected by analyzing the structure. For example, the presence of a helix–turn–helix motif suggests that the protein binds DNA. Two alpha helices intertwined for approximately eight turns with leucine residues occurring every seven residues are the dimerization domains of many DNA-binding proteins. A motif in which a zinc atom is bound to two cysteines and two histidines separated by twelve residues is called a zinc finger and is found in DNA and RNA binding proteins. Other shorter and non-contiguous local arrangements can be identified and associated with a function, for example the arrangement of serine, histidine and aspartate in serine proteases (Figure 1.3).

When no known local functional motif can be detected, it is still possible to analyze clefts on the surface of the protein (in more than 70% of proteins the largest cleft contains the catalytic site) and highlight the presence of amino acid side-chains that are likely to be involved in catalytic activity. Biochemical knowledge can help us to postulate a catalytic mechanism.

For non-enzymes, the problem is much harder to solve. Detecting the protein–protein interaction sites is very difficult and there is not yet a completely satisfactory method, although analysis of the hydrophobicity of the surface in conjunction with automatic learning approaches is leading to some success.

When other members of the evolutionary family are known, analysis of the conservation and variability of amino acids facilitate estimation of the functional importance of different parts of the structure. Any approach used to detect function from structure has a major limitation, however: the molecular function of a protein does not tell us very much about its biological role. If we predict a protease activity for an enzyme, even if we can identify the likely substrate, we are still left with the question of its biological role, because these enzymes participate in many processes, from digestion to blood coagulation, from host defense to programmed cell death.

It should also be mentioned that, recently, more and more proteins (called moonlight proteins) have been found to perform more than one function, often totally unrelated to each other. This might be frustrating, but should not be surprising – there is no reason evolution should not take advantage of different surface regions of proteins to endow them with different activities.

Another significant proportion of proteins seem to be intrinsically disordered and assume their native structure only when they meet and bind with their partners (natively unfolded proteins).

> *Question: Is the property of being disordered functionally important ?*
>
> »The property is often evolutionarily conserved and is, therefore, deemed to be functionally relevant. The reason might be that their flexibility enables these proteins to bind several targets or to provide a large interacting surface in big complexes. This might also be a clever way of engineering high specificity but low affinity. A large interaction surface usually confers both properties but, if the protein has to expend energy for folding before binding, specificity can still be achieved without large affinity. Other explanations can be invoked for this behavior, for example the lifetime of an unfolded protein in a cell is probably shorter and this can provide a regulatory mechanism. More simply, there is no reason why evolution should select against these proteins, because selective pressure acts on the function of the protein and is not concerned with what the protein does when not involved in its functional interactions, assuming it does not have a deleterious effect on the cell.«

The existence of moonlight and natively unfolded proteins makes the problem of inferring the function of novel proteins even more complex and, indeed, this is one of the fields that is attracting more attention at the present. It is easy to predict that

many new more powerful methods will be developed in the near future, taking advantage both of the wealth of data that is being accumulated and of novel approaches. The problem is somewhat recent – before the start of structural genomics initiatives, the determination of the structure of a protein was usually the final step of its characterization and was aimed at understanding the details of its functional mechanism or interactions rather than to infer its biological function from scratch. Only recently are we facing the challenge of having an available structure and no functional information.

1.9 The Evolution of Protein Function

In 1859 Darwin published "The Origin of Species", a book that laid the foundation of evolutionary theory. The careful observations he made during his travels led him to realize that the taxonomy of species could be explained by postulating gradual changes occurring generation after generation and to propose that changes might result in competitive advantage for the organism as members of a population better fitted to survive leave more offspring. The traits of successful individuals then become more common, whereas traits that do not increase, or even reduce, the fitness become rarer or disappear altogether. Evolution, therefore, acts to transform species in the direction of better fitness for the environment. Darwin also had the intuition that even traits that do not, by themselves, confer any selective advantage, might become predominant in a population if they attract the preference of sexual partners. At about the same time Mendel discovered that the traits of the partners are not blended in the offspring – on the contrary, specific characters are sorted and inherited. The foundations for a molecular theory of evolution only needed identification of the material carrying the characteristics, i.e. the DNA, and this happened approximately half a century later.

At the time it did seem surprising that a simple molecule such as DNA, which is, after all, only a polymer comprising a limited set of different nitrogen-containing bases (only four, as it happens) each attached to a sugar and a phosphate group, could explain the diversity between an amoeba and a man. Only fifty years later, however, the diffraction data collected by Rosalind Franklin enabled James Watson and Francis Crick to build a structural model of DNA. The structure of this molecule immediately suggested how DNA can be replicated and copied. This is probably the only example in history in which knowledge of the structure of a macromolecule has immediately provided information about a novel functional mechanism.

What remained to be understood was how the DNA could code for proteins, i.e. what was the code linking the four-character alphabet of a DNA molecule with the twenty-letter alphabet of proteins. The path that led to unraveling this code was much harder than the previous steps, it took years of study and experiments to obtain the genetic code (Table 1.2), i.e. the correspondence between each triplet of bases of the DNA and the coded amino acid. With rare exceptions, the genetic code

is universal, it is used by bacteria, plants, animals, more proof, if needed, of the theory of evolution.

Table 1.2 The genetic code.

	U	C	A	G	
	Phe	Ser	Tyr	Cys	U
U	Phe	Ser	Tyr	Cys	C
	Leu	Ser	**STOP**	**STOP**	A
	Leu	Ser	**STOP**	Trp	G
	Leu	Pro	His	Arg	U
C	Leu	Pro	His	Arg	C
	Leu	Pro	Gln	Arg	A
	Leu	Pro	Gln	Arg	G
	Ile	Thr	Asn	Ser	U
A	Ile	Thr	Asn	Ser	C
	Ile	Thr	Lys	Arg	A
	Met	Thr	Lys	Arg	G
	Val	Ala	Asp	Gly	U
G	Val	Ala	Asp	Gly	C
	Val	Ala	Glu	Gly	A
	Val	Ala	Glu	Gly	G

A story related to the genetic code is very instructive for highlighting that biology is a complex field and that beautiful elegant theories might fall short of describing how evolution has shaped life. Francis Crick was one of the contributors to the long and labor-consuming path that led to the discovery of the genetic code. In 1957, however he devised the following solution to the problem of mapping the four DNA bases to the twenty amino acids. We have four bases and twenty amino acids. There are only 16 possible combinations of two characters out of an alphabet of four (4^2), therefore a coding system that associates a pair of bases with each amino acid could only encode sixteen amino acids. If we use a triplet of bases to code for an amino acid, then we have 64 (4^3) possible combinations. The problem is how to map 64 triplets to twenty amino acids. Let us assume that only a subset of the 64 possible triplets codes for amino acids and that they are such that, when two are placed next to each other, only one "reading frame" can be meaningful. For example, if the codons CGU and AAG code for an amino acid, then none of the triplets GUA and UAA can be coding sequences, so if the DNA contains the sequence CGUAAG there is no ambiguity in how it should be read and translated.

In this hypothesis, none of the triplets AAA, CCC, GGG, and UUU can be coding, because if they were the sequences AAAAAA, CCCCCC, GGGGGG and UUUUUU would be ambiguous, so we are left with 60 codons. Next, only one of the codons that are cyclic permutations of each other can be coding. Let us consider

the codons ACG, CGA and GCA. Ambiguities arise if more than one of these is used. For example, if we use ACG and CGA, the sequence ACGACG is ambiguous. This implies that we can only select one codon every three; we are therefore left with only 20 out of the 60 codons. Crick and colleagues did realize that, apart for its elegance and simplicity, there was no other support for this hypothesis and indeed they wrote: "We put it forward because it gives the magic number – 20 – in a neat manner and from reasonable physical postulates." The theory was very elegant and equally wrong, demonstrating that evolution selects a working alternative, and does not seem to be interested in elegant minimal design!

Each cell of an organism, with rare exceptions, carries a complete copy of the genetic material. Bacteria usually have only one copy of the genetic material organized as a circle made of double-stranded DNA. More complex species have cells with nuclei and reproduce sexually. Most of the DNA in such organisms resides in the cell nucleus and is arranged in several chromosomes that occur in pairs. One member of each pair comes from the mother and one comes from the father. Some DNA is also stored in separate organelles within the cell, called mitochondria and chloroplasts.

Darwin viewed evolution in terms of the genealogical relationships among species or major groups of organisms over a long time span. The impressive progress in molecular biology enables us to study evolution in molecular terms, by looking at the change in the genetic make-up of a population and at the differences between species in terms of difference in their DNA sequence.

Replication, either in the process of creating new somatic cells (mitosis) or in the process of creating germ cells (meiosis), is extremely accurate, and there are several mechanisms to ensure its fidelity; errors are inevitable, however. Environmental factors, for example high-energy radiation, can, moreover, cause random damage to the DNA molecule. Mechanisms exist for repairing the damage, but sometimes they introduce errors. These can be of two types – replacements of DNA bases by others or deletions or insertions of any number of bases. A base replacement may or may not affect a protein sequence. The change may occur in an intron or in another region of the DNA that does not code for a protein. When it occurs in a protein–coding region, the replacement might lead to a codon that is translated into the same amino acid as the original, because of the redundancy of the genetic code. Alternatively, an amino acid residue in the original protein may be replaced by a different amino acid in the mutated protein (missense mutation) or the mutation can involve a stop codon. If a codon for an amino acid residue is changed to a stop codon, the protein will be terminated prematurely and will usually be non-functional (nonsense mutation) whereas if a stop codon mutates into a codon for an amino acid residue the translation continues, elongating the amino acid chain until the next stop codon is encountered.

Large insertions or deletions in the coding regions of a protein almost always prevent production of a useful protein. Short deletions or insertions in a coding region of any number of bases other than a multiple of three usually have a drastic effect – they cause a shift in the reading frame during translation, resulting in a meaningless change in the amino acid sequence in the C-terminal direction from

the point of mutation. When the insertion or deletion involves multiples of three bases, it does not affect the sequence of the protein outside the site of the insertion or deletion and may or may not affect its function.

A gene, or a whole chromosomal region, might be duplicated, leading to a situation in which two copies of the same gene are present. If there is no selective pressure, the two copies may evolve independently – one copy may continue to code for the protein performing the original function whereas the other may evolve by mutation into an entirely different protein with a new function. New combinations of existing genes are occasionally produced at the beginning of meiosis when the chromatids, or arms, of homologous chromosomes break and reattach to different chromosomes (crossing-over). It is easy to see how these mechanisms can account for variability within a species and differences between different species.

One can also speculate about the mechanisms by which new species arise. A species is defined as a set of individuals that, in the wild, would mate and produce fertile offspring. A new species can therefore originate when some individuals, for whatever reason, do not mate with the rest of the population for a sufficient length of time. These individuals can follow a different evolutionary path that might result in genetic incompatibility with the original group. This can be because of physical separation between groups of individuals, or acquisition of different lifestyle, or spreading of the individuals over a huge geographical range. Study of evolution and of the relationships between species and their proteins is of paramount importance in modern biology; most of what we can infer about the function and structure of biological elements comes from analysis of their differences and similarities with the corresponding elements in different species. The possibility of comparing the sequence of entire genomes has also resulted in the possibility of highlighting which parts of the genome are under evolutionary pressure and are, therefore, deemed to be functionally important. It is not surprising that a plethora of tools and theories has been developed to highlight evolutionary relationships, some of which will be described in this book in the appropriate context. Here we will review some elements of the terminology that are commonly used.

Phylogeny is an inferred pattern of evolutionary relationships between different groups of organisms. Usually we depict a phylogeny as a rooted tree in which the length of the branches is proportional to the divergence time and each leaf represents a species. We also use a tree representation to indicate the evolutionary relationships between genes or proteins. In this representation each leaf is a gene or a protein and the lengths of the branches are proportional to the accumulated changes between the molecules. It should be kept in mind that, although molecular trees derived from protein sequences are related to phylogenetic trees, the former refer to the observed difference, for example in functional regions, and do not, therefore, necessarily relate to a proper phylogenetic tree, because different evolutionary pressure can result in different rates of evolution for different genes or proteins. For example, the rate of mutation of hemoglobin is approximately one change per site every billion years whereas fibrinopeptides can accumulate nine time this number of mutations in the same period of time.

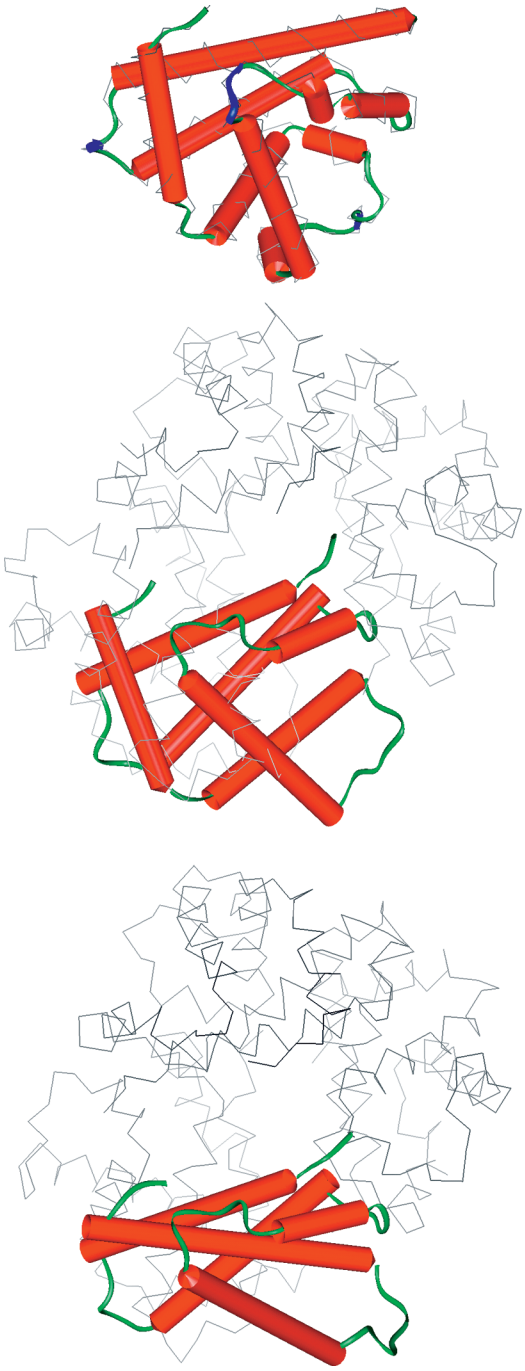


Figure 1.20 Three homologous chains: myoglobin, and alpha and beta globin. Note that these three structures are similar and evolutionarily related, but they are paralogous, i.e. they have arisen by gene duplication.

Sometimes molecular trees are unrooted, i.e. they depict differences, but make no hypothesis about the location or properties of the ancestor elements. For some applications, such a tree is good enough, but – once again – it is not a tool to derive evolutionary times.

Two elements (whether genes, proteins, anatomical structures) that derive from a common evolutionary ancestor are called "homologous". In anatomy and in protein analysis homology does not guarantee common functionality – the anterior wings of a bat and a human arm are homologous but have a different function. If two anatomical parts resemble each other, they are called analogous. The usual example here is the eye of vertebrates and that of the squid, they seem similar and have a similar function, but have a different evolutionary origin. We will call two protein structures analogous if they resemble each other (i.e. they have the same topology) but there is no evidence of common evolutionary origin. When we refer to genes or proteins, the concept of homology must be further specified. Two proteins (or genes) believed to have diverged from each other because of speciation events are called orthologous whereas two proteins (or genes) that are homologous, i.e. derived from a common ancestor, but have arisen after a duplication event are called paralogous. This is by no means a semantic distinction but one of the major issues in protein bioinformatics, because it has a very relevant impact on the prediction of the biological function of newly discovered proteins.

Let us use an example to illustrate the issue. Myoglobin is a monomeric protein the function of which is to store oxygen in the muscle. Human and chimp myoglobin are orthologous – they descend from the same ancestral protein via speciation and they also have the same function. Hemoglobin, the tetrameric oxygen transporter is composed by two pairs of alpha and beta chains. Myoglobin, alpha hemoglobin, and beta hemoglobin are all homologous, they descend from a common ancestor, as is apparent from their structural similarity (Figure 1.20), but they are paralogous, because they have arisen by the duplication of an ancestral globin gene. If two homologous genes are found in the same genome, it is easy to see that they are paralogous. Paralogous genes can, however, also be present in different genomes – human myoglobin and chimp alpha globin are paralogous. As illustrated by this example, because of the possibility of paralogous relationships, the finding that two genes are homologous does not necessarily imply they share the same function.

1.10
The Evolution of Protein Structure

If a base-substitution event occurs in a protein-coding region of a genome, the net effect can be the substitution of one residue of the encoded protein with a different one. What is the effect on the structure of the protein of a single amino acid replacement? There are only two possibilities. One is that the fine balance between the gain and loss of free energy of folding is compromised, there is no single global energy minimum for the new sequence, and it does not fold any more. Because

proper folding is required for function, the most likely outcome is that the organism is not viable and the mutation is not propagated in the population. The second alternative is that the energy landscape of the new sequence changes, but it still contains a free energy global minimum and the corresponding native structure is still able to perform the same function as the original protein.

How likely is it that the new conformation is very different from the original? Statistics and physics both tell us this is extremely unlikely and that the most probable outcome is that the new sequence assumes a structure very similar to that of the original protein. In other words the substituted amino acid is accommodated into the structure with only local perturbation and without dramatic global changes in structure and function. Indeed substantial changes in protein architecture, because of a single, evolutionarily accepted mutation, have not yet been observed. Therefore, when residue substitutions and short insertions/deletions accumulate in members of an evolutionarily related family of proteins, they will cause local

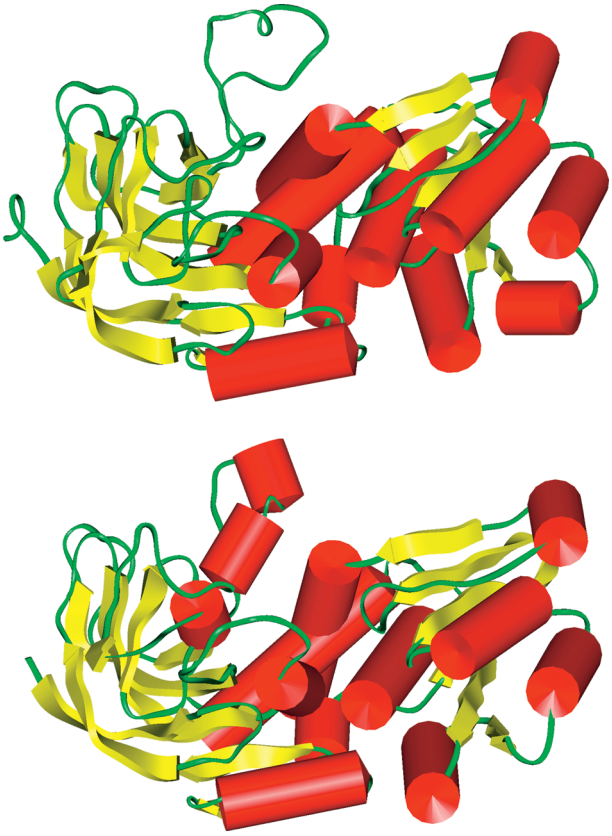


Figure 1.21 Two homologous proteins sharing 30% sequence identity: alcohol dehydrogenase from horse liver (PDB code: 1YE3) and *Acinetobacter calcoaceticus* (1F8F).

structural perturbations without affecting the general shape, or topology, of the protein. Of course, the greater the number of mutations (or, equivalently, the further the proteins are in the evolutionary scale), the larger will be the difference between the protein structures. We can quantify this qualitative observation by measuring the relationship between the different sequences of homologous proteins and their structural divergence, as we will see in the next section. We will merely mention here that it is accepted that pairs of evolutionarily related proteins sharing at least 30% sequence identity have a similar fold (Figure 1.21).

> *Question: Is the sequence-to-structure relationship limited to naturally evolved proteins or does it reflect an intrinsic property of amino acid sequences?*
>
> »It is important to understand that the sequence–structure relationship already described occurs when there is the requirement, as in natural evolution, that each step of the evolutionary path must be functional. If we were to artificially introduce several mutations into a protein structure without guaranteeing that each step yields a functional mutant, we might be able to change its structure dramatically. Indeed, in 1994 two scientists (Creamer and Rose) issued the "Paracelsus Challenge" – they offered a prize of $1000 to the first group to successfully convert one protein fold into another while retaining at least 50% sequence identity with the original fold (this challenge was named after Paracelsus, the 16th century Swiss alchemist). At least three groups have published reports in response to this challenge.«

Insertions and deletions, even when they do not cause frame shifts, are difficult to accommodate in a protein structure because a substantial part of the molecule, the internal core, is tightly packed. The periodicity of secondary structure elements also implies that insertion or deletion of one amino acid can change the pattern of interaction of the whole region very markedly. Not surprisingly, most of the observed amino acid insertions and deletions are located at the surface of the protein structure and outside secondary structure elements.

Fusion of two or more initially independent genes leads to the production of multidomain proteins with new combinations of functions in a single protein. In eukaryotes, this process is thought to be facilitated by the presence of introns (intervening sequences in genes that are not coding). These represent regions in which genes can easily be recombined – if one exon from a gene coding for a protein region with a given function is inserted into an intron region of a gene for a protein carrying a different function, the new hybrid protein might be capable of both functions and serve a new physiological role. It should be mentioned, however, that there is no evidence that exons preferentially encode structural or functional units.

1.11
Relationship Between Evolution of Sequence and Evolution of Structure

To analyze the relationship between sequence and structural divergence in quantitative terms, we must define a measure of distances in sequence and structure space. Several unsolved problems connected with this issue will be discussed later in this book. For the moment let us assume we know how to find the correspondence between the amino acids of two evolutionarily related proteins that reflects their evolutionary history. In other words, let us assume that, given two proteins, we can construct a matrix such as that shown in Figure 1.22 in which the first row contains the amino acid sequence of the first protein, possibly with inserted spaces, and the second contains the amino acid sequence of the second protein, again possibly including spaces. Two amino acids in the same column are assumed to originate from the same amino acid of the ancestor protein. The spaces represent insertion and deletion events. Given this correspondence, called alignment, we can define the distance in sequence space simply as the fraction of amino acids that is different between the two proteins.

Sequence 1	...	M	Q	D	G	T	S	R	F	T	C	R	G	K	–	–	P	I	H	H	F
Sequence 2	...	L	S	E	G	N	H	K	L	S	C	R	H	D	Q	G	P	V	N	D	Y

Figure 1.22 Alignment of two fragments of the sequences of the proteins shown in Figure 1.21.

To measure distance in structure space, we use the root mean square deviation (*rmsd*) between corresponding atom pairs after optimum superposition. In practice, we apply the rigid-body translation $T = (T_x, T_y, T_z)$ and rotation $R = (R_x, R_y, R_z)$ to one of the proteins that minimizes the value:

$$rmsd(T,R) = \min_{T,R} \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left[(x_i - R_x x_i' + T_x)^2 + (y_i - R_y y_i' + T_y)^2 + (z_i - R_z z_i' + T_z)^2\right]}$$

The set of corresponding atom pairs should, once again, be that which reflects the evolutionary correspondence between amino acids. Obviously, when we are superimposing two proteins with different sequences, we can only use atoms that are common between any two protein structures, for example the Cα, or the atoms of the backbone, or the backbone plus the Cβ for all amino acids except glycine.

As already discussed, peripheral parts of the proteins can undergo local rearrangements that can be quite substantial. If there are insertions and deletions it is obviously impossible to compute the *rmsd* values for the inserted residues, but even if this is not so, the fact that the superposition procedure minimizes the sum of the squares of the deviations means the *rmsd* is dominated by the most divergent regions and, if we include them, the changes in the more conserved regions would be masked. This leads to the need to superimpose separately the conserved "cores"

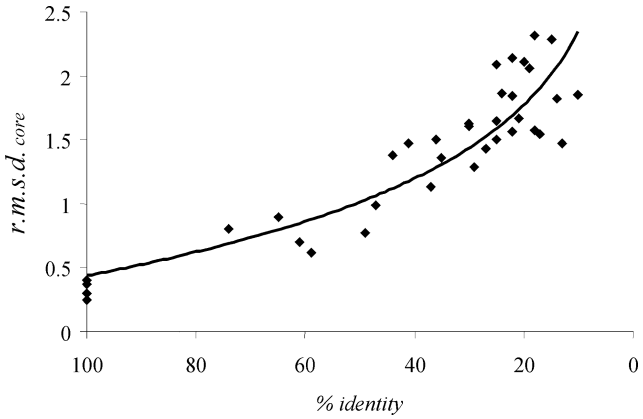


Figure 1.23 Relationship between sequence identity and structural similarity. The plot is obtained using the same set of proteins originally analyzed by Lesk and Chothia.

of evolutionarily related proteins and, therefore, calls for a definition of the core. Many empirical procedures are commonly used. Chothia and Lesk, for example, propose the following: given two related proteins, one first superimposes the main chain atoms of corresponding elements of the secondary structure and then continues to add residues at either ends of the elements until the distance between the alpha carbons of the last added residue deviates by more than 3 Å. Next, one jointly superimposes these "well fitting" regions and calculates the resulting *rmsd*.

Now that we know how to measure sequence distance and structural divergence, we can investigate the relationship between them. In a seminal paper Chothia and

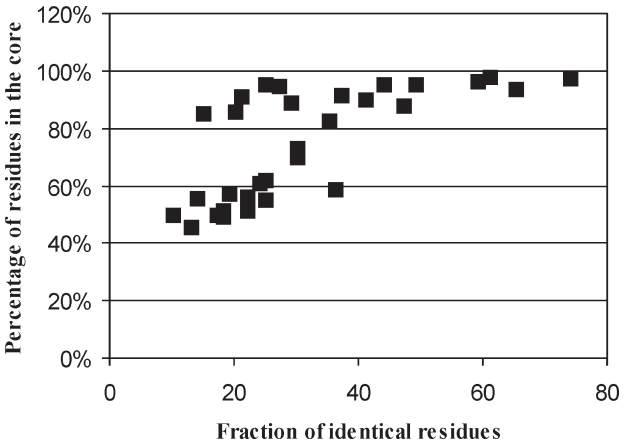


Figure 1.24 Relationship between sequence identity and the extent of the common structural core between pairs of homologous proteins. (Data from the original Chothia and Lesk analysis on thirty-two pairs of proteins).

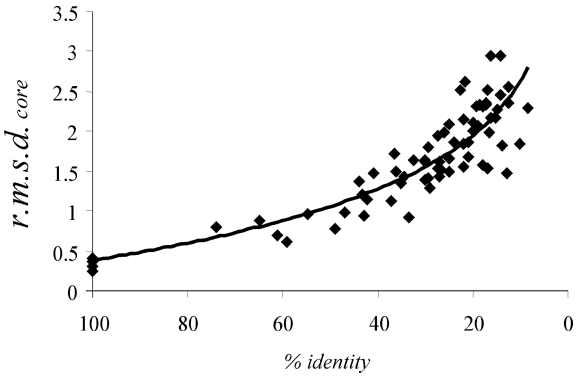


Figure 1.25 Relationships between sequence identity and structural similarity. The plot was obtained by using a larger set of proteins than in Figure 1.23, but the trend is essentially the same.

Lesk selected thirty-two pairs of homologous proteins of known structure, identified the common core within each pair with the procedure described above, and computed the *rmsd* values between the core of each pair as a function of the sequence identity between the two protein sequences (Figure 1.23). Their conclusions were that, as sequences diverge, the extent of the common core between two homologous proteins decreases. The common core contains almost all the residues when pairs of closely related proteins (with sequence identity >50%) are considered; when residue identity drops below 20% the structures might diverge quite substantially and the core can contain as little as 40% of the structure (although in some cases it can include most of the structure) (Figure 1.24). They found that the relationship between the structural divergence of the core and the sequence identity was in accordance with the equation:

$$rmsd_{\text{core}} = 0.40e^{\frac{(100 - \%identity)}{100}}$$

Although the original analysis by Chothia and Lesk was limited to 32 pairs of proteins only, a relationship with very similar parameters was obtained when the analysis was repeated using a much larger sample (Figure 1.25).

Suggested Reading

These two books guide the reader through a fascinating tour of protein architecture and show how the shape of a protein is linked to its function:

A.M. Lesk (**2001**) Introduction to Protein Architecture, Oxford University Press

A.M. Lesk: (**2004**) Introduction to Protein Science: Architecture, Function, and Genomics, Oxford University Press

Another excellent book that describes the principles of protein structure, with examples of key proteins in their biological context, is:

C.-I. Branden, J. Tooze (**1999**) Introduction to Protein Structure, 2nd edn, Garland, New York

Readers interested in learning more about crystallography and nuclear magnetic resonance spectroscopy can consult these two seminal books:

J. Drenth (**1994**) Principles of Protein X-ray Crystallography, Springer
K. Wüthrich (**1986**) NMR of Proteins and Nucleic Acids, John Wiley and Sons

Every biochemistry book contains at least a chapter dedicated to the structure and function of proteins. Readers can read the relevant chapters from:

J.M. Berg, L. Stryer, J. L. Tymoczko (**2002**) Biochemistry, 5th edn, W. H. Freeman
D.L. Nelson, M. M. Cox (**2004**) Lehninger Principles of Biochemistry, 4th edn, W. H. Freeman
D. Voet, J. Voet (**2004**) Biochemistry, 3rd edn, Wiley

The original paper describing the PDB data archive (http://www.pdb.org) is:

F.C. Bernstein, T. F. Koetzle, G. J. B. Williams, E. F. Meyer Jr, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, M. Tasumi (**1977**) The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. J. Mol. Biol. **112**, 535–542

Structural classification of proteins, SCOP (http://scop.mrc-lmb.can.ac.uk/scop), is described in:

A.G. Murzin, S. E. Brenner, T. Hubbard, C. Chothia (**1995**) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J. Mol. Biol. **247**, 536–540
CATH (http://www.biochem.ucl.ac.uk/bsm/cath/) C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, J. M. Thornton (**1997**) CATH – A hierarchic classification of protein domain structures. Structure **5**, 1093–1108
FSSP (http://www.bioinfo.biocenter.helsinki.fi:8080/dali/index.html) L. Holm, C. Sander (**1996**) Mapping the protein universe. Science **273**, 595–602

The original paper by Anfinsen is:

C.B. Anfinsen, E. Haber, M. Sela, F. White Jr (**1961**) The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain. PNAS **47**, 1309–1314

It is difficult to gain access to the original paper in which Levinthal described the paradox that took his name:

C. Levinthal (**1969**). In: P. Debrunner, J. C. M. Tsibris, E. Munck (Ed.) Mossbauer spectroscopy in biological systems, University of Illinois, Urbana, IL, pp. 22–24

However, several more recent papers describe the reasoning explained in the paper.

The worldwide initiatives in structure genomics are described at http://www.rcsb.org/pdb/strucgen.html where a good collection of informative background information about this project is also available.

I would also recommend reading "The Origin of Species" by Charles Darwin, because of its outstanding historical and scientific interest.

The paper by Francis Crick and colleagues, describing their theory on the comma free genetic code is:

F.H.C. Crick, J. S. Griffith, L. E. Orgel (**1957**) Codes without commas. PNAS USA **43**, 416–421

Finally, the analysis of the relationship between sequence and structural similarity in proteins by C. Chothia and A. M. Lesk can be found in:

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