Chapter 5: Protein Structure and Drug Discovery

Introduction

The great variety of three-dimensional structures and functions of proteins arise in molecules that share underlying common features. Chemically, proteins are like strings of Christmas tree lights: each protein consists of a linear (i.e. unbranched) polymer mainchain with different amino acid sidechains attached at regular intervals (Fig. 1.6). The wire linking the string of lights corresponds to the repetitive mainchain or backbone, and the variable sequence of colours of the lights corresponds to the individuality of the sequence of sidechains.

The amino acid sequence of a protein is specified by the nucleotide sequence of a gene. The three-dimensional structures of protein molecules are determined, without further participation of nucleic acids, by the one-dimensional sequences of their amino acids. Proteins fold spontaneously to their native conformations.

How does the amino acid sequence encode the three-dimensional structure? Any possible folding of the mainchain places different residues into contact. The interactions of the sidechains and mainchain, with one another and with the solvent, and the restrictions placed on sidechain mobility, determine the relative stabilities of different conformations. This is a consequence of the second law of thermodynamics, which states that systems at constant temperature and pressure find an equilibrium state that is a compromise between comfort (low enthalpy, H) and freedom (high entropy, S), to give a minimum Gibbs free energy $G = H - TS$, in which T is the absolute temperature. (In human relationships, marriage is just such a compromise.)

Proteins have evolved so that one folding pattern of the mainchain is thermodynamically significantly better than other conformations. This is the native state. If we could calculate sufficiently accurately the energies and entropies of different conformations, and if we could computationally examine a large enough set of possible conformations to be sure of including the correct one, it would be possible consistently to predict protein structures from amino acid sequences on the basis of *a priori* physicochemical principles. There has been progress towards this goal but it has not yet been achieved.

The mainchain of each protein in its native state describes a curve in space. We now know the structures of 15 000 proteins (including many identical or single-site mutants), and see in them a great variety of spatial patterns. The first problem in analysing these structures is one of presentation. Figure 5.1 illustrates, for the small protein acylphosphatase, the difficulty in interpreting a fully detailed, literal representation, and the kind of simplified pictures that computer programs produce to give us visual access to the material. An active cottage industry has produced many different simplified representations. A skilled molecular illustrator will combine them to show different parts of a structure in finely tuned degrees of detail.

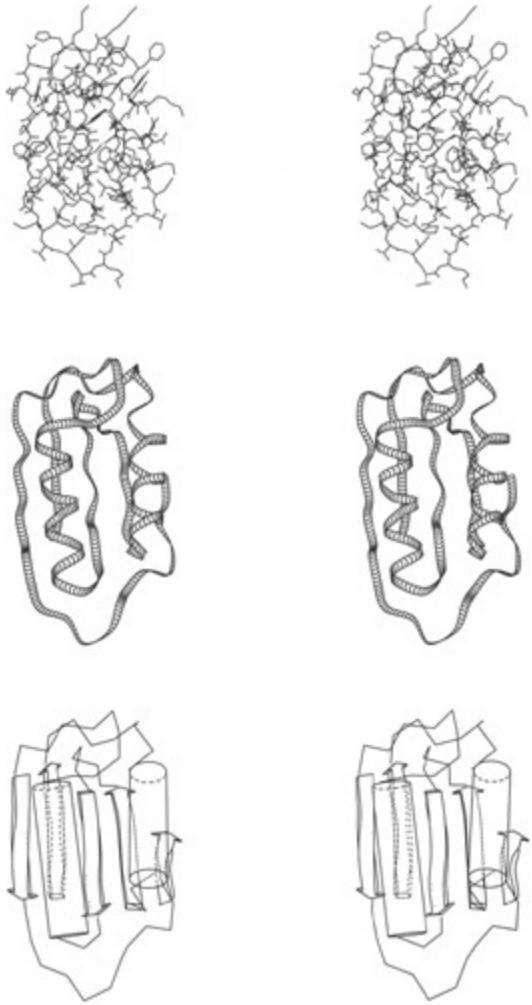


Figure 5.1: Proteins are sufficiently complex structures that it has been necessary to develop specialized tools to present them. This figure shows a relatively small protein, acylphosphatase, at three different degrees of simplification. Top: complete skeletal model. Centre: the course of the chain is represented by a smooth interpolated curve, the chevrons indicating the direction of the chain. Bottom: schematic diagram, in which cylinders represent helices and arrows represent strands of sheet. The solid objects in the picture are represented as 'translucent' by altering lines that pass behind them to broken lines. It is possible to superpose different representations visually by rotating the page by 90° and viewing in stereo (but not for too long!).

The central frame of Fig. 5.1 shows the course through space of the main-chain of acylphosphatase. Two regions at the front of the picture have the form of helices - like classic barber poles - with their axes almost vertical in the orientation shown. Acylphosphatase also contains four strands of sheet. These too are approximately vertical in orientation. The four strands interact laterally to stabilize their assembly into a β-sheet. In the bottom frame, helices and strands are represented as 'icons': helices as cylinders and strands of sheet as large arrows. The top frame of Fig. 5.1, showing the most detailed representation of the structure, including mainchain and sidechains, indicates the importance of simplification in producing an intelligible picture of even a small protein.

Protein stability and folding

Although it is not yet possible to predict the structures of proteins from basic physical principles alone, we do understand the general nature of the interactions that determine protein structures.

To form the native structure, the protein must optimize the interactions within and between residues, subject to constraints on the space curve traced out by the mainchain. Preferred conformations of the mainchain bias the folding pattern towards recurrent structural patterns: helices, extended regions that interact to form sheets, and several standard types of turns.

The Sasisekharan-Ramakrishnan-Ramachandran plot describes allowed mainchain conformations

To a good approximation, the mainchain conformation of each non-glycine residue is restricted to two discrete conformational states.
A fragment of the linear polypeptide chain common to all protein structures is shown in Fig. 5.2. Rotation is permitted around the N-Cα and Cα-C single bonds of all residues (with one exception: proline). The angles ϕ and ψ around these bonds, and the angle of rotation around the peptide bond, ω, define the conformation of a residue. The peptide bond itself tends to be planar, with two allowed states: *trans*, $\omega \approx 180°$ (usually) and *cis*, $\omega \approx 0°$ (rarely, and in most cases at a proline residue). The sequence of ϕ, ψ and ω angles of all residues in a protein defines the backbone conformation.

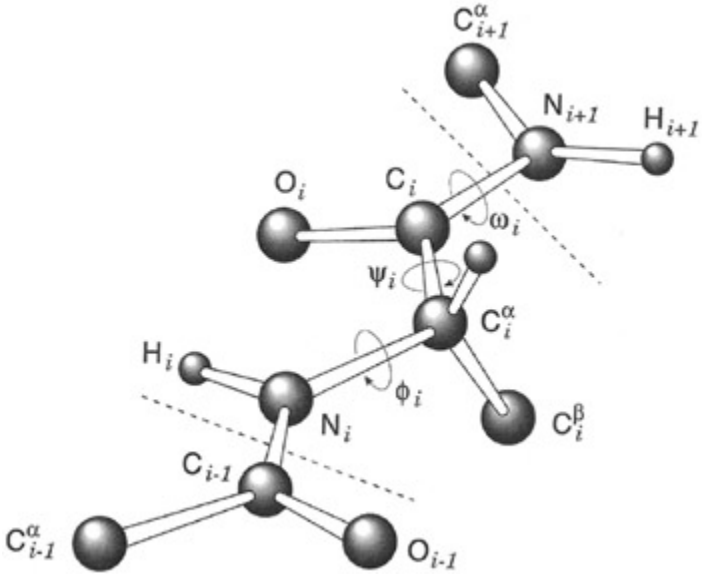


Figure 5.2: Definition of conformational angles of the polypeptide backbone.

The principle that two atoms cannot occupy the same space limits the values of conformational angles. The allowed ranges of ϕ and ψ, for $\omega = 180°$, fall into defined regions in a graph called Sasisekharan-Ramakrishnan-Ramachandran plot - usually shortened to 'Ramachandran plot' (see Fig. 5.3). Solid lines in the figure delimit energetically-preferred regions of ϕ and ψ; broken lines in the figure delimit sterically-disallowed regions. The conformations of most amino acids fall into either the α_R or β regions. Glycine has access to additional conformations. In particular it can form a left-handed helix: α_L. Figure 5.3 shows the typical distribution of residue conformations in a well-determined protein structure. Most residues fall in or near the allowed regions, although a few are forced by the folding into energetically less-favourable states.

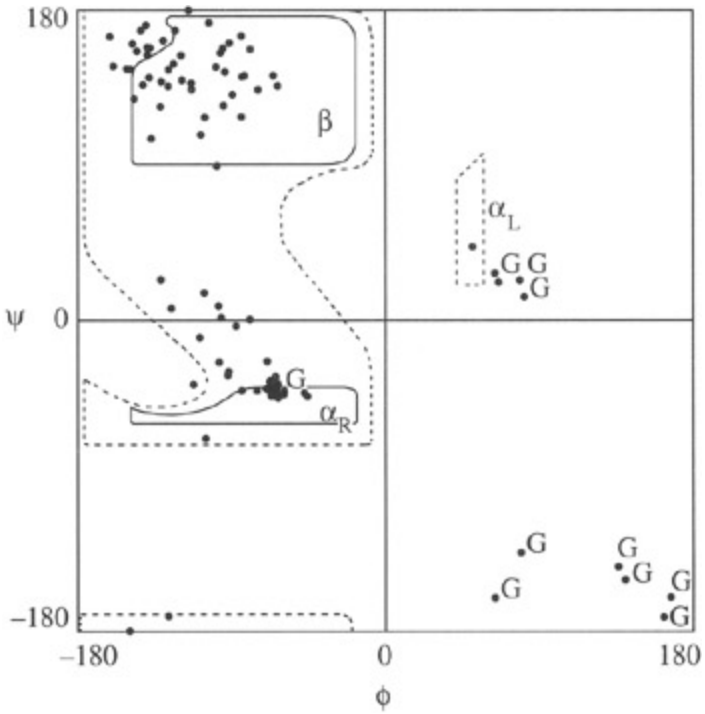


Figure 5.3: A Sasisekharan-Ramakrishnan-Ramachandran plot of acylphosphatase (PDB code 2ACY). Note the clustering of residues in the α and β regions, and that most of the exceptions occur in Glycine residues (labelled G).

The allowed regions generate standard conformations. A stretch of consecutive residues in the α conformation (typically 6–20 in native states of globular proteins) generates an α-helix. Repeating the β conformation generates an extended β-strand. Two or more β-strands can interact laterally to form β-sheets, as in acylphosphatase (Fig. 5.1). Helices and sheets are 'standard' or 'prefabricated' structural pieces that form components of the conformations of most proteins. They are stabilized by relatively weak interactions, *hydrogen bonds,* between mainchain atoms (Fig. 1.7). In some fibrous proteins all of the residues belong to one of these types of structure: wool contains α-helices; silk β-sheets. Amyloid fibrils, formed in disease states by many proteins, also contain extensive β-sheets.
Typical globular proteins contain several helix and/or sheet regions, connected by *turns.* Usually the ends of helix or strand regions appear on the surface of a domain of a protein structure. They are connected by turns, or loops: regions in which the chain alters direction to point back into the structure. Many but not all turns are short, surface-exposed regions that tend to contain charged or polar residues.

How does the mainchain choose among the possible allowed conformations? What is unique about each protein is the sequence of its sidechains. Therefore, interactions involving sidechains must determine the mainchain conformation.

The sidechains

Sidechains offer the physicochemical versatility required to generate all the different folding patterns. The sidechains of the twenty amino acids vary in:

- *Size.* The smallest, glycine, consists of only a hydrogen atom; one of the largest, phenylalanine, contains a benzene ring.
- *Electric charge.* Some sidechains bear a net positive or negative charge at normal pH. Asp and Glu are negatively charged, Lys and Arg are positively charged. (Charged residues of opposite sign can form attractive pairwise interactions called *salt bridges.*)
- *Polarity.* Some sidechains are polar; they can form hydrogen bonds to other polar sidechains, or to the mainchain, or to water. Other sidechains are electrically neutral. Some of these contain chemical groups related to ordinary hydrocarbons such as methane or benzene. Because of the thermodynamically unfavourable interaction of hydrocarbons with water, these are called 'hydrophobic' residues. Congregation of hydrophobic residues in protein interiors, predicted by W.J. Kauzmann before the first protein structures were determined, is an important contribution to protein stability. This effect is analogous to the formation of droplets of oil in salad dressing (see Box, p. 223: The Hydrophobic Effect).

- **The hydrophobic effect**

- The difference among the different amino acid sidechains in their preferences for aqueous or oil-like environments is one of the governing principles of protein structure.
- What is the hydrophobic effect? Phase separation in oil-water mixtures - for instance, salad dressing - is one common example; another is that gases (unlike most solids) are less soluble in water as the temperature increases. Readers with whistling tea kettles will have heard low levels of sound prior to proper boiling - this occurs when the dissolved air comes out of solution as the water is heated.
- What is the origin of the hydrophobic effect? Cold water is a highly structured liquid. It contains many hydrogen bonds, which account for its high heat of vaporization and low density. But water is even more highly ordered around solutes than in the pure liquid. Methane dissolved in water - it is only slightly soluble, but soluble enough to study - is surrounded by a cage of water molecules called a clathrate complex. As a result, dissolving methane in water makes the solvent even more ordered, lowering the entropy. The natural tendency toward states of higher entropy inhibits the dissolving of methane in water. This is why methane and other hydrocarbons are only very slightly water-soluble. The solubilities of non-polar gases decrease upon heating-from an already small value in cold water - because as the temperature increases entropy plays an even more important role in determining the equilibrium state.
- The hydrophobic effect in aqueous solutions of simple non-polar solutes was well known to physical chemists when W.J. Kauzmann, in 1959, recognized its importance for protein structure.
- The nonpolar sidechains of proteins are similar to oil-like solutes. Their interaction with water is unfavourable. Kauzmann predicted that they would be sequestered in protein interiors, away from the solvent. This *oil-drop model of protein interiors* was confirmed by the X-ray crystal structures of globular proteins. We now recognize also the importance of high packing densities in protein interiors, and that it is more accurate to regard the interior of a folded protein as more like a crystal than like an organic liquid. But the hydrophobic effect has lost none of its significance.
- As a consequence of the hydrophobic effect, charged residues are almost completely excluded from protein interiors; in rare cases they form internal salt bridges. Obviously, the backbone must traverse the interiors of the protein, and carries with it the polar N and 0 atoms, which can interact with other polar mainchain atoms and with polar sidechains such as threonine or asparagine. Thus, the interior is not completely oil-like. Conversely, the surface of a protein is not exclusively charged or polar. About half the residues on the surface of a protein are non-polar.
- *Shape and rigidity.* The overall shape of a sidechain depends on its chemical structure and on its degrees of internal conformational freedom.

Protein stability and denaturation

What are the chemical forces that stabilize native protein structures? What is the process by which a protein folds from an ensemble of denatured conformations to a unique native state?
To address these questions, biochemists have studied the denaturation of proteins in response to heat, or increasing concentrations of urea or guanidinium hydrochloride (commonly used denaturants). Some measurements are *static* - determination of the amount of native and denatured states at equilibrium under different conditions, or the heat released at points along the transition. Others are *kinetic* - measurement of rates of folding or unfolding, or identification of structures that appear transiently during the process.

One important message is that proteins are only marginally stable. The native state of globular proteins is typically only 20–60 kJ mol^{-1} (5–15 kcal mol^{-1}) more stable than the denatured state. This is the equivalent of about one or two water-water hydrogen bonds.

Precisely why proteins have marginal stability is unclear. Some people believe that it facilitates protein turnover. Others suggest that proteins are as stable as they need to be so 'why bother' (less informally: there is no selective advantage in) further optimizing the stabilizing interactions. We do know that the interactions that stabilize native proteins are capable of producing protein structures with much higher stabilities.
Suppose you are a globular protein in aqueous solution, and you want to achieve a stable native state. Your major problem is the great loss of conformational freedom, relative to the ensemble of denatured states, that is exacted from you in adopting a unique conformation. This entails a large reduction in entropy, which is thermodynamically unfavourable. One way in which you can compensate is to form a compact globular state, burying many residues in the interior away from contact with water. The release of water from interaction with

the non-polar atoms of the protein produces a compensating *increase* in entropy arising from the *hydrophobic effect* (see Box).

That's fine, but now you discover that to form the compact state you have buried many polar atoms, including but not limited to mainchain nitrogen and carbonyl oxygens. In the denatured state, these atoms make hydrogen bonds to water. When buried in the interior, their hydrogen bonding potential must somehow be satisfied. (Do not forget: one or two uncompensated hydrogen bonds and you have blown it; your native state would be unstable.) A fairly general-purpose solution that satisfies mainchain hydrogen-bonding potential is to form helices or sheets.
There is a bonus: Formation of helices and sheets also ensures that the mainchain is in a stereochemically acceptable conformation, as limited by the Sasisekharan-Ramakrishnan-Ramachandran plot. Residues in α-helices are all in the α conformation; residues in strands of β-sheet are all in the β conformation.
How do you decide which regions should form helices or strands? Enthalpically, helix and sheet are reasonably similar for most residues. However, entropically, some sidechains are more hindered in helices than in strands; these prefer strands. These effects bias the formation of secondary structures. Specific sequences providing sidechain-mainchain hydrogen bonds form *helix caps*, governing where α—helices begin and end.

How compact is the globular state required to be? You could achieve exclusion of water from your interior by fairly loose packing - as long as no channel is larger than 1.4 Å in radius (the size of a water molecule.) But the closer together you can squeeze your atoms, the better advantage you can take of Van der Waals forces, general forces of attraction between atoms that give matter its general cohesion. Protein interiors are densely packed: the fitting together of the sidechains is like a solved jigsaw puzzle. However, the puzzle pieces (the residues) are deformable, so the folding process is more complicated than the rigid matching of pieces in ordinary jigsaw puzzles.

In summary, you have to find a conformation of the chain that simultaneously solves all the following problems:

1. All residues must have stereochemically allowed conformations. This applies to both the mainchain and the sidechains. Steric collisions would raise the energy of the conformation and render it unstable.
2. Buried polar atoms must be hydrogen bonded to other buried polar atoms. If you miss out a few hydrogen bonds, the protein will prefer to form the denatured state in order to allow these polar atoms to hydrogen bond to solvent.
3. Enough hydrophobic surface must be buried, and the interior must be sufficiently densely packed, to provide thermodynamic stability.

For most proteins, there is a unique solution of all these problems, and this defines the native state. Some proteins change conformation when they bind ligands, or pass through metastable states, as part of their mechanisms of function.

The fact that one conformation of a protein - the native state - has substantially greater stability than others is complex but not mysterious. It is a question of optimizing the available interactions, and selecting sequences for which this optimum is unique and substantially lower than others. For most regions the local structure is determined by local interactions. Therefore, if the native state were not unique there would have to be more than one way to fit a given set of pieces together. Given the chain constraints it is easy for evolution to avoid this.

Protein folding
Suppose again that you are a protein, and that you are denatured. Now that you understand how your native state is stabilized, how would you go about finding it? Clearly you cannot try all conformations - many years ago C. Levinthal calculated that a simple conformational search, using reasonable numbers for speeds of internal rotations, would require much too much time to finish. Two circumstances conspire to make the *process* by which proteins fold to their native states a mysterious one.
First is the fact that proteins are only marginally stable. This implies that any quasi-stable intermediate in protein folding must be even less stable, else the folding process would get trapped in the intermediates. Indeed, for many proteins, measurements of fractions of molecules in native and denatured states as a function of temperature or denaturant concentration imply simple two-state Native ↔ Denatured equilibria in which undetectably few molecules are anything but native or denatured. This confirms that any putative intermediates can have no more than marginal stability. But this makes it difficult to follow the folding transition structurally.

The second circumstance that makes protein folding mysterious is that the denatured state is so heterogeneous that in the absence of stable intermediates there is no convenient way to visualize the complete pathway.

Contrast protein folding with two other types of structure formation:

1. In assembling do-it-yourself furniture, one passes through a succession of well-defined intermediate states. First one screws A to B in the native-like conformation. The structure of the A–B fragment is determined and stabilized purely by the interactions between A and B. Were it not for gravity, a stable A–B intermediate would be formed. But proteins do not have the luxury of forming stable intermediates.
2. In assembling an arch from its voussoirs, the structure as a whole has no stability until the keystone is inserted. Only the completed arch has independent stability; there are no stable intermediates, and the only way to assemble the structure is by using scaffolding which is subsequently removed. But proteins do not have the luxury of using external scaffolding.

What proteins have to do is to work with unstable intermediates - like do-it-yourself furniture in the *presence* of gravity - and to get the job finished before the intermediates fall apart, or else keep reforming them and trying again.

Identification of transient structure during protein folding can be achieved experimentally by isotope exchange measurements. Prepare a sample of denatured protein in which all Hydrogen atoms are replaced by Deuterium. (It is possible to separate signals from H and D in NMR experiments.) At various times during re-folding, in separate experiments, expose the sample to a pulse of protons. After the native state is formed, detect where in the structure D ↔ H exchange occurred and when. Such studies justify the model that many proteins fold by initial formation of a 'molten globule' containing some native secondary structure, but without the tertiary structural interactions that lock the molecule into its final conformation. This is followed by a hierarchical condensation to form supersecondary structure, etc., leading eventually to accretion of the native state. For most proteins, there is no evidence for non-native structures as intermediates along productive folding pathways, although non-native structures - such as incorrect proline isomers - can divert and thereby slow down the folding process.

The conclusion is that structures of local regions are determined primarily by local interactions, and, although these interactions may be inadequate to stabilize local regions to the point where they can be isolated, they are good enough to provide a low-energy pathway for structure assembly.

Applications of hydrophobicity

Using a *hydrophobicity scale* that assigns a value to each amino acid, we can plot the variation of hydrophobicity along the sequence of a protein. This is called a *hydrophobicity profile.* Analysis of hydrophobicity profiles has been used to predict the positions of turns between elements of secondary structure, exposed and buried residues, membrane-spanning segments, and antigenic sites.

Example 5.1

Use of hydrophobicity profiles to predict the positions of turns between helices and strands of sheet.

Figure 5.4a shows the hydrophobicity profile of hen egg white lysozyme. It has pronounced minima at the following residues: 17, 44, 70, 93, and 117. Figure 5.4b shows the structure of hen egg white lysozyme, from which it is possible to check the correlation between turns in the structure and the positions of the minima in the hydrophobicity profile.

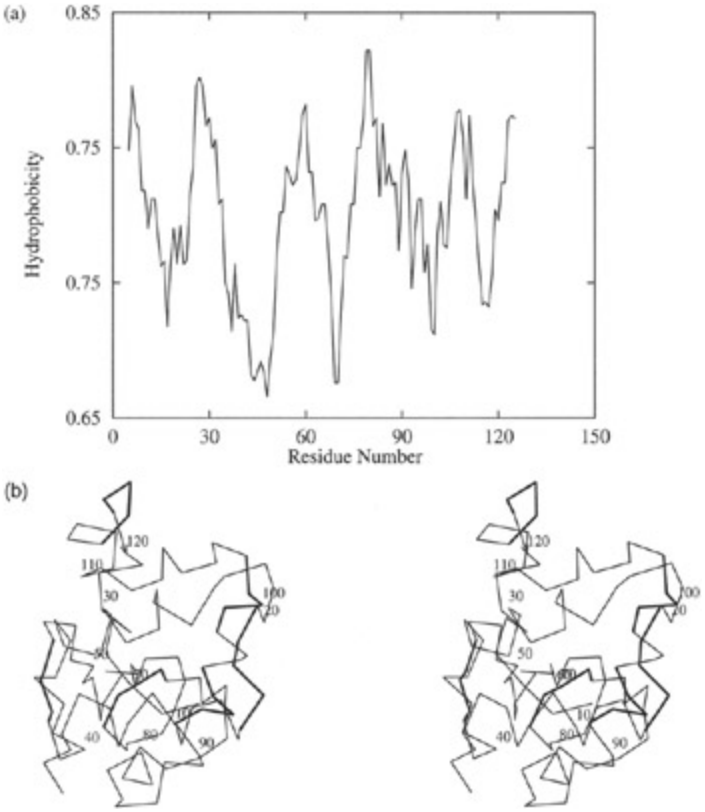


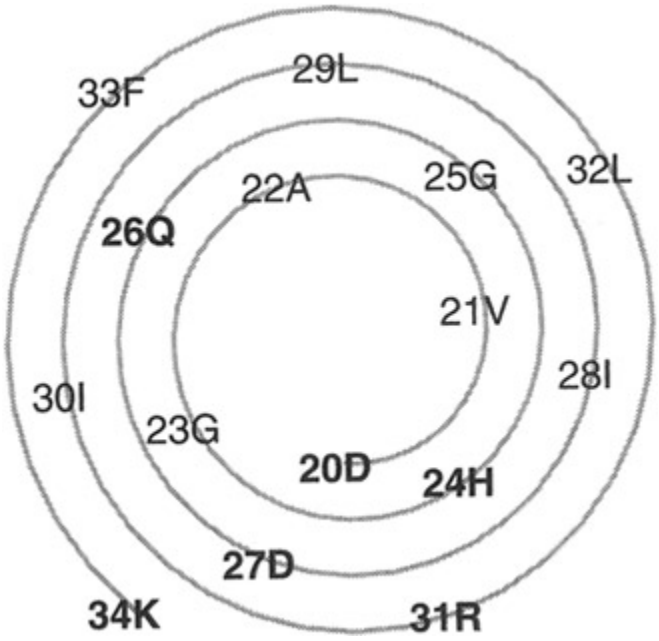
Figure 5.4: (a) Hydrophobicity profile of hen egg white lysozyme. (Produced using the Primary Structure Analysis tools available through http://www.expasy.ch.) (b) Sturcutre of hen egg while lysozyme. Regions corresponding to minima in the hydrophobicity plot are shown in thicker lines.

Four of the major minima in the hydrophobicity profile appear at or near the positions of turns. Another minimum occurs in a surface-exposed region, but in the structure this one is a strand of a β-sheet rather than a turn. One of the minima is within a helix. Conversely, many of the turns do not correspond to pronounced minima in the hydrophobicity plot. Hydrophobicity profiles provide useful information, but do not unambiguously predict all turns in a protein structure.

Example 5.2

The helical wheel. O.B. Ptitsyn observed that α-helices in globular proteins often have a 'hydrophobic face' turned inwards towards the protein interior, and a 'hydrophilic face' turned outwards towards the solvent. Each residue in an α-helix appears at a position 100° around the circumference from its predecessor. Therefore, to achieve Ptitsyn's effect, the sequence of residues should alternate between hydrophobic and hydrophilic with a periodicity of approximately four.

To check this relationship, the residues can be projected onto a plane perpendicular to a helix axis, a diagram called a *helical wheel*. This example shows the sequence of an α-helix of sperm whale myoglobin. Charged and polar residues appear in boldface type; others in ordinary type.



The helix has a hydrophobic face-which points to the inside of the structure, and a hydrophilic face - which points outside. From such a pattern of hydrophobicity we can predict whether a region of an amino acid sequence is likely to form an a-helix in the native protein structure.

The box shows a PERL program to draw helical wheels.

Example 5.3
Detection of transmembrane helical segments. Many membrane proteins have the structure, first seen in bacteriorhodopsin, of seven helices traversing a membrane, connected by loops (see Fig. 5.5).

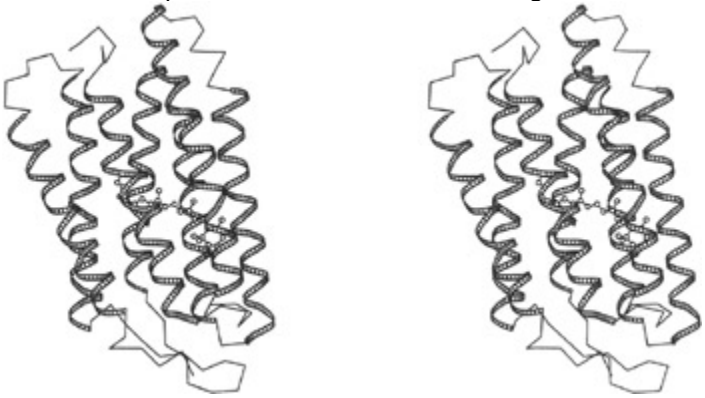
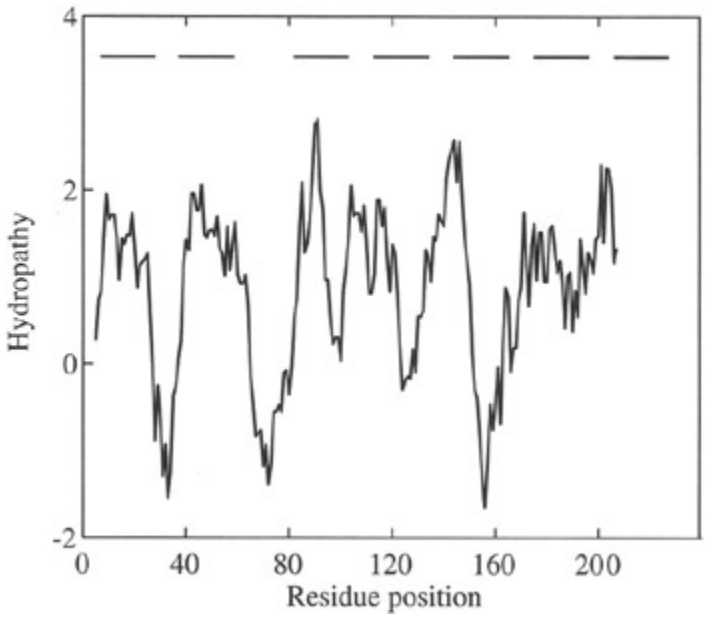


Figure 5.5: Bacteriorhodopsin from the bacterium *Halobacterium salinarum* (formerly *Halobacterium halobium*) [2BRD] viewed in the plane of the membrane. The ligand shown in ball-and-stick representation is the chromophore, retinal.

Residues within the membrane-spanning segments are almost exclusively hydrophobic, because the entire helix is embedded in a non-aqueous medium. They are separated by regions containing polar amino acids. Transmembrane helices are typically 15–30 residues long.
An unfiltered hydrophobicity plot of the amino acid sequence of *H. salinarum* bacteriorhodopsin shows seven regions of maxima, corresponding to the seven transmembrane helices (positions indicated by horizontal bars).




```
#!/usr/bin/perl

#helwheel.pl -- draw helical wheel

#usage: echo DVAGHGQDILIRLFKSH | helwheel.prl > output.ps

# or    echo 20DVAGHGQDILIRLFKSH | helwheel.prl > output.ps

#       the numerical prefix sets the first residue number


# The output of this program is in PostScript (TM),

#      a general-purpose graphical language


# The next section prints a header for the PostScript file


print «EOF;

%!PS-Adobe-

%%BoundingBox: (atend)

%1 0 0 setrgbcolor

%newpath

%37.5 161 moveto 557.5 161 lineto 557.5 681 lineto 37.5 681 lineto

%closepath stroke

297.5 421. translate 2 setlinewidth 1 setlinecap

/Helvetica findfont 20 scalefont setfont 0 0 moveto

EOF
```

```
#  Define fonts to associate with each amino acid

$font{"G"} = "Helvetica";      $font{"A"} = "Helvetica";      $font{"S"} = "Helvetica";
$font{"T"} = "Helvetica";      $font{"C"} = "Helvetica";      $font{"V"} = "Helvetica";
$font{"I"} = "Helvetica";      $font{"L"} = "Helvetica";      $font{"F"} = "Helvetica";
$font{"Y"} = "Helvetica";      $font{"P"} = "Helvetica";      $font{"M"} = "Helvetica";
$font{"W"} = "Helvetica";      $font{"H"} = "Helvetica-Bold"; $font{"N"} = "Helvetica-Bold";
$font{"Q"} = "Helvetica-Bold"; $font{"D"} = "Helvetica-Bold"; $font{"E"} = "Helvetica-Bold";
$font{"K"} = "Helvetica-Bold"; $font{"R"} = "Helvetica-Bold";

$_= <>;                              #  read line of input
chop(); $_ =~ s/\s//g;                  #  remove terminal carriage return and blanks

if {$_ =~ s/^(\d+)//)                   #  if input begins with integer
    {$resno = $1;}                      #  extract it as initial residue number
else {$resno = 1}                       #  if not, set initial residue number = 1

$radius = 50;                           #  initialize values for radius,
$x = 0; $y = -50; $theta = -90;            #  x, y and angle theta

# print light gray spiral arc as succession of line segments, 10 per residue

$npoints = 10*(length($_) - 1);

print "0.8 0.8 0.8 setrgbcolor\n";         #  set colour to light gray
print "newpath\n";                       #  draw spiral arc
printf("%8.3f %8.3f moveto\n",$x,$y);
foreach $d (1 .. $npoints) {               #  10 points per residue
   $theta += 10; $radius += 0.6;           #  increase radius and theta
   $x = $radius*cos($theta*0.01747737);    #  calculate new value of x
```

```
    $y = $radius*sin($theta*0.01747737);    #    and y
    printf("%8.3f %8.3f lineto\n",$x,$y);
}
print "stroke\n";

#  print residues and residue numbers

$radius = 50;                          #  reinitialize values for radius,
$x = 0; $y = -50; $theta = -90;            #  x, y and angle theta
print '0 setgray\n'                        #  set colour to black

foreach (split ("",$_)) {                 #  loop over characters from input line
    print "/$font($_) findfont";           #  set font appropriate
    print "20 scalefont setfont\n";        #  for this amino acid
    printf("%8.3f %8.3f moveto\n",$x,$y);   #  move to current point
    print " ($resno$_) stringwidth";       #  adjust position to center residue
    print " pop -0.5 mul -7 rmoveto\n";     #     identification on point on spiral
    print " ($resno$_) show\n";            #  print residue number and id
    print "% $theta $resno$_\n";
    $theta += 100; $radius += 6;           #  set new values of angle, radius
    $x = $radius*cos($theta*0.01747737);    #  compute new values of x
    $y = $radius*sin($theta*0.01747737);    #     and y
    $resno++;                              #  increase residue number
}

print "showpage\n";                        #  postscript signals to
print "%%BoundingBox:";                    #  print
$xl = 297.5 - 1.05*$radius;                #   x
$xr = 297.5 + 1.05*$radius;                #    and
$yb = 421.  - 1.05*$radius;                #     y
$yt = 421.  + 1.05*$radius;                #      limits
```

```
printf("%8.3f %8.3f %8.3f %8.3f\n",$xl,$xr,$yb,$yt);


print "showpage\n";

print "%%EOF\n";                    # and wind up
```

A number of programs available on the Web offer specialized methods for transmembrane helix prediction.

Web Resource: Transmembrane Helix Prediction

TMHMM (A. Krogh and E. Sonnammer) - based on a Hidden Markov Model:
http://www.cbs.dtu.dk/krogh/TMHMM/
PHDhtm (B. Rost):
http://dodo.bioc.columbia.edu/predictprotein
Membrane protein explorer (S. White):
http://blanco.biomol.uci.edu/mpex/

Superposition of structures, and structural alignments

Some aspects of sequence analysis carry over fairly directly into structural analysis, some must be generalized, and others have no analogues at all.

As in the case of sequences, a fundamental question in analysing structures is to devise and compute a measure of similarity. If two molecules have identical or very similar structures, we can imagine superposing them so that corresponding points are as close together as possible. Then the average distance between corresponding points is a measure of the structural similarity. In practice it is conventional to report the root-mean-square deviation of the corresponding atoms:

$$\text{r.m.s. deviation} = \sqrt{\sum d_i^2 / n}$$

where d_i is the distance between the ith pair of points after optimal fitting, and n is the number of points.

This assumes that we have pre-specified the correspondence between the points.
If the correspondence is not known, we must first determine it and only then calculate the r.m.s. deviation of the alignable substructures. If each point corresponds to an atom representing the successive residues of a protein or nucleic acid structure (the Cα atoms of proteins or the phosphorus atoms of nucleic acids), the problem is literally a question of alignment (= assignment of residue-residue correspondences) (see Box, next page). Indeed, determination of residue-residue correspondences via structural superposition of two or more proteins is a powerful method of sequence alignment. Because structure tends to diverge more conservatively than sequence during evolution, structure alignment is a more powerful method than pairwise sequence alignment for detecting homology and aligning the sequences of distantly-related proteins.

Example 5.4

Structural alignment of γ-chymotrypsin and *Staphylococcus aureus* epidermolytic toxin A.
Chymotrypsin and *S. aureus* epidermolytic toxin A are both members of the chymotrypsin family of proteinases.

Figure 5.6 shows a structural superposition of PDB entries 8GCH (γ-chymotrypsin) (solid lines) and 1 AGJ (*S. aureus* epidermolytic toxin A) (broken lines) The molecules share the common chymotrypsin-family serine proteinase folding pattern, and the Ser-His-Asp catalytic triad (thicker lines).

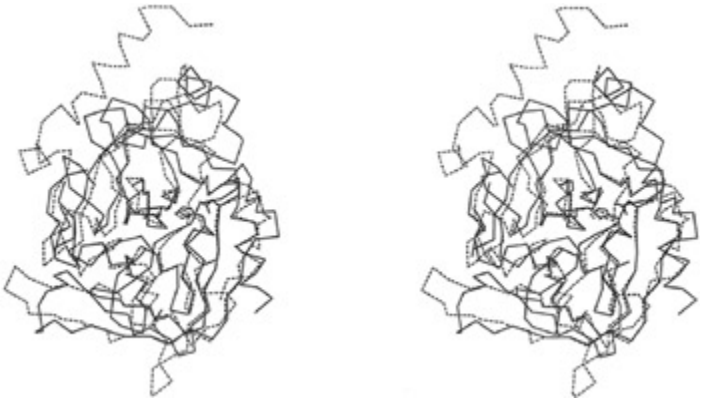


Figure 5.6: Structural superposition of γ-chymotrypsin [8GCH] (solid lines) and *S. aureus* epidermolytic toxin A [1ACJ] (broken lines). The sidechains of the catalytic triads are shown. Observe that the region around the active site is the best-conserved part of the protein.

A sequence alignment derived from the superposition follows:

8gch CGVPAIQPVLIVNG----------EEAVP--GS----WPWQVSLQ-DKTG

1agj ----------EVSAEEIKKHEEKWNKYYGVNAFNLPKELFSKVDEKDR-QKYPYNTIGNVFVK-G-

8gch FH--FCGGSLINE-NWVVTAAHC-GV-T---T-SDVVVAGEFDQG---SSSEKI--QKLKIAKVFK-NS-

1agj --QTSATGVLIG-KNTVLTNRHIAK-FANGDPSKVSFRPSI-NTDDNGNT-E-TPYGEYEVKEILQEP-F

8gch KYNSLTINNDITLLKLST-----AAS--FSQTVSAVCLPSASD--DFAAGTTCVTTGWG-LTRYNTPD-R

1agj GAG-----VDLALIRLKPDQNGVSL-GDK---ISPAKIGT---SNDLKDGDKLELIGYPFDH----KVNQ

9gch LQQASLPLL-SNTNCKKYWGTKIKDAM--ICAGASGV-SSCMGDSGGPLVCKKNGAWTLVGIVSWGSSTC

1agj MHRSEIELTTLS----------RGLRYY----GFTVPGNSGSGIFNSN---GELVGIHSSK----

8gch STST--------- PGVYARVTA-LVNWVQQTLAAN-

1agj ----VSHLDREHQINYGVGIGNYVKRIINEKN---E

The resemblance between these two sequences is well within the 'twilight zone'. It could not be derived correctly from standard pairwise alignment of the two sequences alone.

Determination of similarity and alignment in computational chemistry

1. Similarity of two sets of atoms with known correspondences:

$$p_i \leftrightarrow q_i,\ i = 1,...,N.$$

The analogue, for sequences, is the Hamming distance: mismatches only.

2. Similarity of two sets of atoms with unknown correspondences, but for which the molecular structure - specifically the linear order of the residues - restricts the possiblities. In the case of proteins or nucleic acids we are limited to correspondences in which we retain the order along the chain:

$$p_{i(k)} \leftrightarrow q_{j(k)},\ k = 1,...,\ k \leq N,\ M$$

with the constraint that: $k_1 > k_2 \Rightarrow i(k_1) > i(k_2)$ **and** $j(k_1) > j(k_2)$**.** This can be thought of as corresponding to the Levenshtein distance, or to sequence alignment with gaps. The result of such a calculation is an alignment of parts of the sequences.

3. Similarities between two sets of atoms with unknown correspondence, with no restrictions on the correspondence:

$$p_{i(k)} \leftrightarrow q_{j(k)}$$

This problem arises in the following important case: Suppose two (or more) molecules have similar biological effects, such as a common pharmacological activity. It is often the case that the structures share a common constellation of a relatively small subset of their atoms that is responsible for the biological activity. These atoms are called a ***pharmacophore.*** The problem is to identify them: to do so it is useful to be able to find the maximal subsets of atoms from the two molecules that have a similar structure.

DALI (Distance-matrix ALIgnment)

As proteins evolve, their structures change. Among the subtle details that evolution has strongly tended to conserve are the patterns of contacts between residues. That is, if two residues are in contact in one protein, the residues aligned with these two in a related protein are also likely to be in contact. This is true even in very distant homologues, and even if the residues involved change in size. Mutations that change the sizes of packed buried residues produce adjustments in the packing of the helices and sheets against one another.
L. Holm and C. Sander applied these observations to the problem of structural alignment of proteins. If the inter-residue contact pattern is preserved in distantly-related proteins, then it should be possible to *identify* distantly-related proteins by detecting conserved contact patterns.

Computationally, one makes matrices of contact patterns in two proteins (this is very easy), and then seeks the maximal matching submatrices (this is hard). Using carefully chosen approximations, Holm and Sander wrote an efficient program called DALI that is now in common use for identifying proteins with folding patterns similar to that of a query structure. The program runs fast enough to carry out routine screens of the entire Protein Data Bank for structures similar to a newly-determined structure, and even to perform a classification of protein domain structures from an all-against-all comparison. Holm and Sander have found several unexpected similarities not detectable at the level of pairwise sequence alignment.
An example of DALI's 'reach' into recognition of very distant structural similarities is its identification of the relation between mouse adenosine deaminase, *Klebsiella aerogenes* urease, and *Pseudomonas diminuta* phosphotriesterase (see Fig. 5.7).

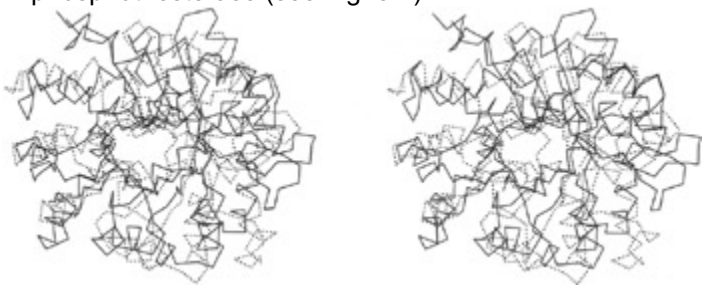


Figure 5.7: The regions of common fold, as determined by the program DALI by L. Holm and C. Sander, in the TIM-barrel proteins mouse adenosine deaminase [1FKX] (solid lines) and *Pseudomonas diminuta* phosphotriesterase [1PTA] (broken lines). In the alignment shown in this figure, the sequences have only 13% identical residues - closer to midnight than to the twilight zone.

DALI is available over the Web. You can submit coordinates to the site http://www2.ebi.ac.uk/dali/, and receive the set of similar structures and their alignments with the query structure.

Evolution of protein structures

Included in the 15000 protein structures now known are several families in which the molecules maintain the same basic folding pattern over ranges of sequence similarity from near-identity down to well below 20% conservation. The serine proteinases (*γ*-chymotrypsin and *S. aureus* epidermolytic toxin A, Fig. 5.6) and the adenosine deaminase-phosphotriesterase family (Fig. 5.7) are examples.

The general response to mutation is structural change. It is characteristic of biological systems that the objects we observe to have a certain form arose by evolution from related objects with similar but not identical form. They must, therefore, be robust, in having the freedom to tolerate some variation. We can take advantage of

this robustness in our analysis: By identifying and comparing related objects, we can distinguish variable and conserved features, and thereby determine what is crucial to structure and function.

Natural variations in families of homologous proteins that retain a common function reveal how structures accommodate changes in amino acid sequence. Surface residues not involved in function are usually free to mutate. Loops on the surface can often accommodate changes by local re-folding. Mutations that change the volumes of buried residues generally do not change the conformations of individual helices or sheets, but produce distortions of their spatial assembly. The nature of the forces that stabilize protein structures sets general limitations on these conformational changes. Additional constraints derived from function vary from case to case.
Families of related proteins tend to retain common folding patterns. However, although the general folding pattern is preserved, there are distortions which increase as the amino acid sequences progressively diverge. These distortions are not uniformly distributed throughout the structure. Usually, a large central *core* of the structure retains the same qualititative fold, and other parts of the structure change conformation more radically. Consider the letters `B` and `R`. As structures, they have a common core which corresponds to the letter `P`. Outside the common core they differ: at the bottom right `B` has a loop and `R` has a diagonal stroke.

Systematic studies of the structural differences between pairs of related proteins have defined a quantitative relationship between the divergence of the amino acid sequences of the core of a family of structures and the divergence of structure. As the sequence diverges, there are progressively increasing distortions in the mainchain conformation, and the fraction of the residues in the core usually decreases. Until the fraction of identical residues in the sequence drops below about 40–50%, these effects are relatively modest. Almost all the structure remains in the core, and the deformation of the mainchain atoms is on the average no more than 1.0 Å. With increasing sequence divergence, some regions re-fold entirely, reducing the size of the core, and the distortions of the residues remaining within the core increase in magnitude.
A correlation between the divergence of sequence and structure applies to all families of proteins. Figure 5.8a shows the changes in structure of the core, expressed as the root-mean-square deviation of the mainchain atoms after optimal superposition, plotted against the sequence divergence: the percentage of conserved amino acids of the core after optimal alignment. The points correspond to pairs of homologous proteins from many related families. (Those at 100% residue identity are proteins for which the structure was determined in two or more crystal environments, and the deviations show that crystal packing forces - and, to a lesser extent, solvent and temperature - can modify slightly the conformation of the proteins.) Figure 5.8b shows the changes in the fraction of residues in the core as a function of sequence divergence. The fraction of residues in the cores of distantly related proteins can vary widely: in some cases the fraction of residues in the core remains high, in others it can drop to below 50% of the structure.

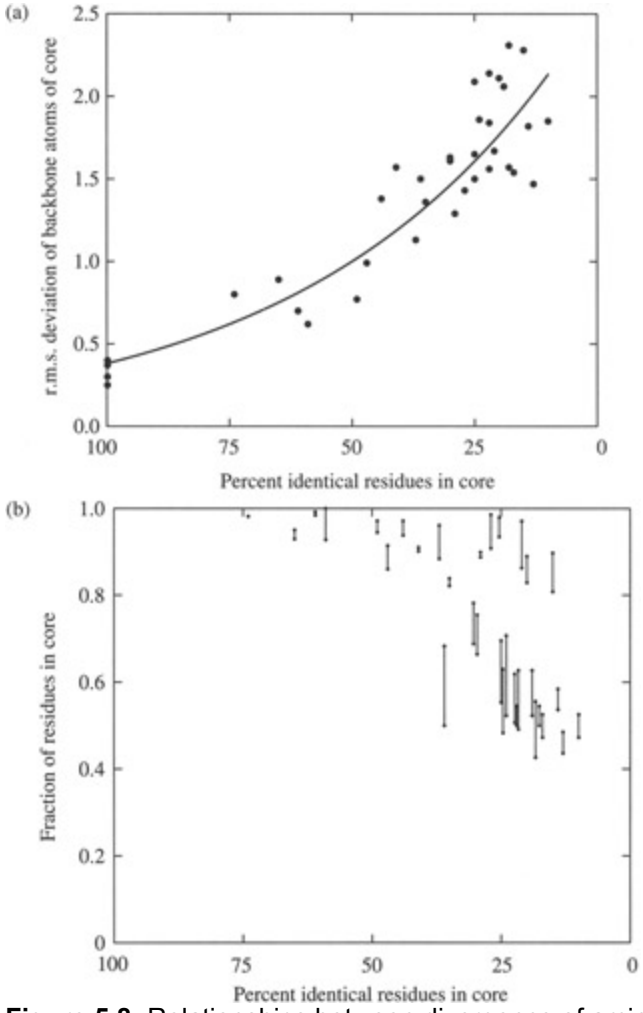


Figure 5.8: Relationships between divergence of amino acid sequence and three-dimensional structure of the core, in evolving proteins. (a) Variation of r.m.s. deviation of the core with the per cent identical residues in the core. (b) Variation of size of the core with the per cent identical residues in the core. This figure shows results calculated for 32 pairs of homologous proteins of a variety of structural types. (Adapted from Chothia, C. and Lesk, A.M. (1986) 'Relationship between the divergence of sequence and structure in proteins,' *The EMBO Journal 5*, 823–6.)

Classifications of protein structures

Organization of protein structures according to folding pattern imposes a very useful logical structure on the entries in the Protein Data Bank. It affords a basis for structure-oriented information retrieval. Several databases derived from the PDB are built around classifications of protein structures. They offer useful features for exploring the protein structure world, including: search for keyword or sequence, navigation among similar structures at various levels of the classification hierarchy, presentation of structure pictures, probing the databank for structures similar to a new structure, and links to other sites. These databases include SCOP (Structural Classification of Proteins), CATH (Class, Architecture, Topology, Homologous superfamily), FSSP/DDD (Fold classification based on Structure-Structure alignment of Proteins/Dali Domain Dictionary), and CE (The Combinatorial Extension Method).

SCOP

The SCOP (Structural Classification of Proteins) database organizes protein structures in a hierarchy according to evolutionary origin and structural similarity. At the lowest level of the SCOP hierarchy are individual *domains,* extracted from the Protein Data Bank entries. Sets of domains are grouped into *families* of homologues, for which the similarities in structure, sequence, and sometimes function imply a common evolutionary origin. Families containing proteins of similar structure and function, but for which the evidence for evolutionary relationship is suggestive but not compelling, form *superfamilies.* Superfamilies that share a common folding topology, for at least a large central portion of the structure, are grouped as *folds.* Finally, each fold group falls into one of the general *classes.* The major classes in SCOP are α, β, α+β, α/β, and

miscellaneous 'small proteins', which often have little secondary structure and are held together by disulphide bridges or ligands.
The box shows the SCOP classification of flavodoxin from *Clostridium beijerinckii* (Plate II). For illustrations of the degree of similarity of proteins grouped together at different levels of the hierarchy, and discussion of other classification schemes, see *Introduction to Protein Architecture: The Structural Biology of Proteins,* chapter 4.

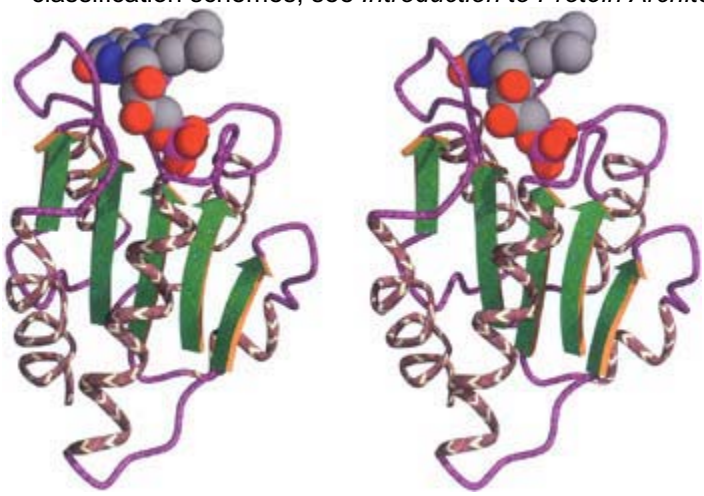


Plate II: Flavodoxin from *Clostridium beijerinckii*, binding cofactor FMN [5NLL]. Large arrows represent strands of sheet. Placement of this structure in a hierarchical classification of protein structures according to the SCOP database is described on page 236.

The SCOP release of July 2001 contained 13 220 PDB entries, split into 31474 domains. The distribution of entries at different levels of the hierarchy is:

Class	**Number of**		
	families	**superfamilies**	**folds**
All-α proteins	337	224	138
All-β proteins	276	171	93
α/β proteins	374	167	97
$\alpha + \beta$ proteins	391	263	184
Multi-domain proteins	35	28	28
Membrane and cell surface proteins	28	17	11
Small proteins	116	77	54
Total	1557	947	605

Numerous other web sites offering classifications of protein structures are indexed at: http://www.bioscience.org/urllists/protdb.htm and http://www2.ebi.ac.uk/msd/Links/fold.shtml.)

SCOP classification of Flavodoxin from Clostridium beijerinckii

1. *Root:* SCOP
2. *Class:* Alpha and beta proteins (α/β)

Mainly parallel β-sheets (β-α-β units)

3. *Fold:* Flavodoxin-like

3 layers, *α/β/α;* parallel *β*-sheet of 5 strands, order 21345

4. *Superfamily:* Flavoproteins
5. *Family:* Flavodoxin-related binds FMN
6. *Protein:* Flavodoxin
7. *Species: Clostridium beijerinckii*

From: http://scop.mrc-lmb.cam.ac.uk/scop