CHAPTER 5

Protein structure and drug discovery

Chapter contents

Learning goals

1. To understand the concept of protein folding: the process by which the one-dimensional amino acid sequence encoded by a gene takes up a definite and biologically active three-dimensional conformation.
2. To recognize that steric considerations severely limit the conformations of the polypeptide chain, with the Sasisekharan-Ramakrishnan-Ramachandran plot showing the allowed states of the mainchain.
3. To get to know the 20 sidechains—the actors that play all the roles in all the proteins.
4. To understand the hydrophobic effect and its implications for the structures and energetics of folded proteins.
5. To generalize the ideas of sequence alignment to the alignment of protein sequences by structural superposition.
6. To know the relationship between divergence of sequence and divergence of structure in protein evolution.
7. To become familiar with classification of protein folding patterns, as presented for example, by the Structural Classification of Proteins (SCOP) database and web site.
8. To know some basic approaches to the prediction of protein structure from amino acid sequence, and the state of the art as revealed in the Critical Assessment of Structure Prediction (CASP) programmes.
9. To understand the basis of Hidden Markov Models, the most powerful methods now available for deducing affinities between proteins from their sequences.
10. To know the basic requirements for a successful drug, and understand some approaches to drug discovery and design.

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 (that is, 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 30 000 proteins (including many replicates 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.

The central frame of Fig. 5.1 shows the course through space of the mainchain of acylphosphatase. Two regions at the front of the picture have the form of helices—like classic barber's 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.

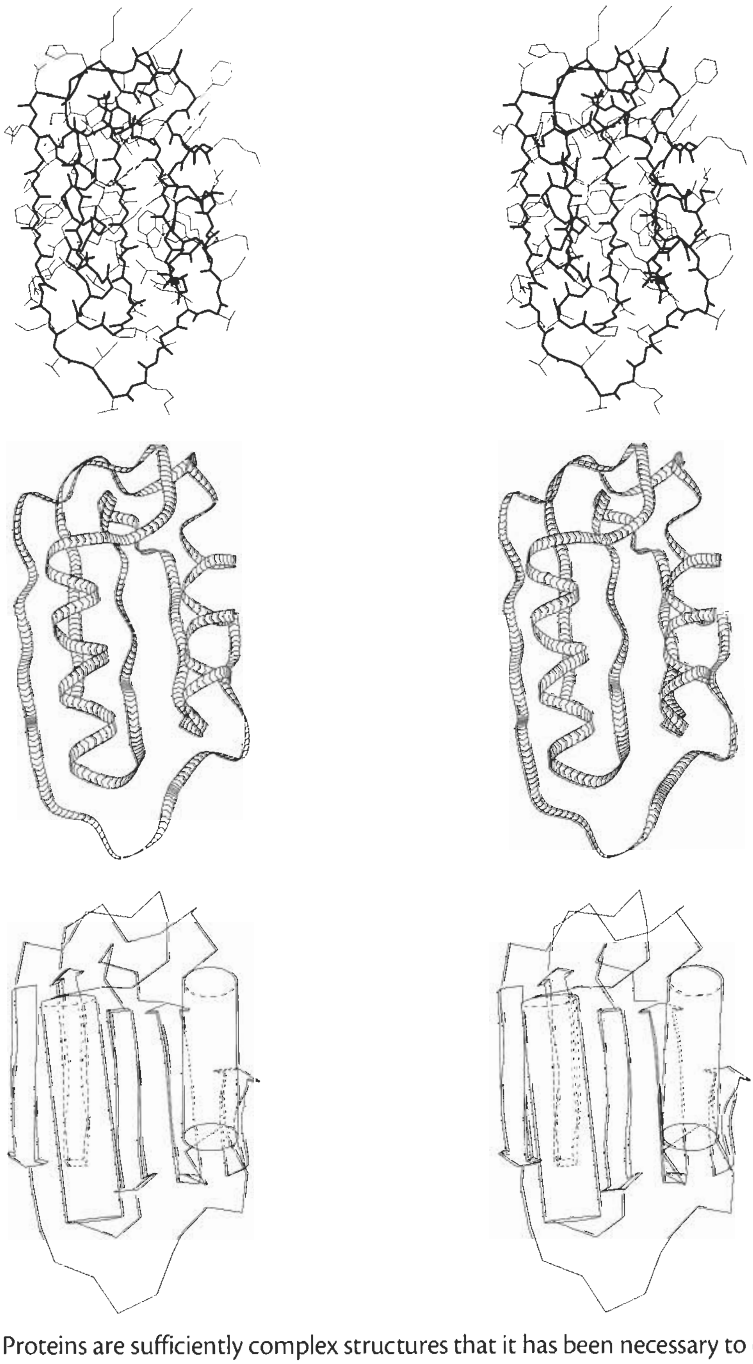


Fig. 5.1 Proteins are sufficiently complex structures that it has been necessary to develop specialized tools to present them. This figure shows a relatively small protein, acylphosphatase, at three different degrees of simplification. Top: complete skeletal model; mainchain bolder than sidechains. Centre: 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 90° and viewing in stereo (but not for too long!).

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 ψ, ϕ and ω angles of all residues in a protein defines the backbone conformation.

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

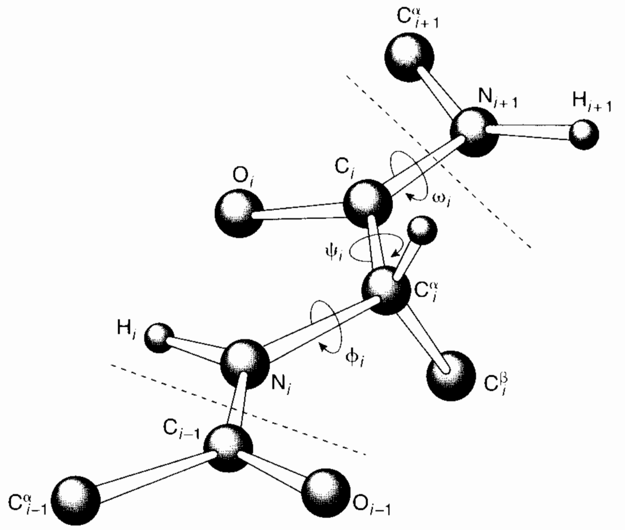


Fig. 5.2 Definition of conformational angles of the polypeptide backbone.

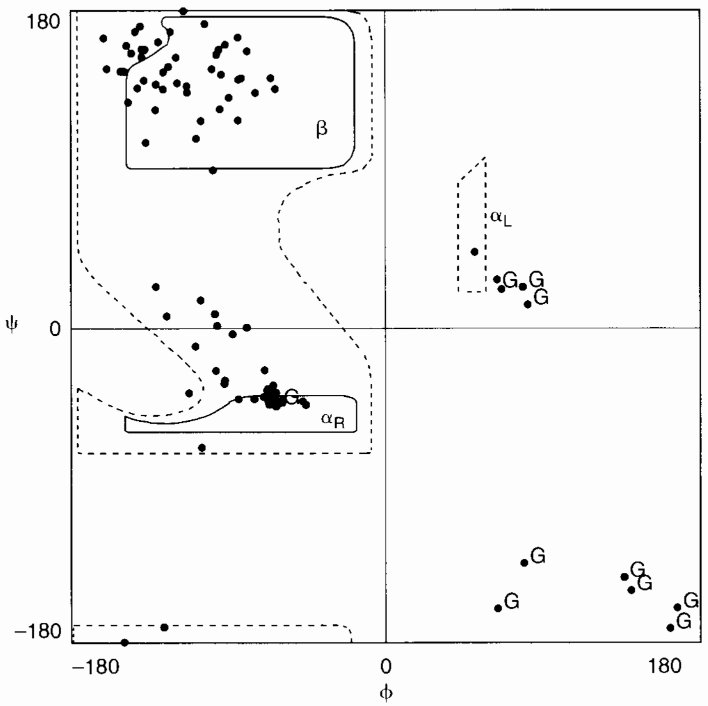


Fig. 5.3 A Sasisekharan-Ramakrishnan-Ramachandran plot of acylphosphatase (PDB code 2ACY). Note the clustering of residues in the α and β regions, and that most of the exceptions occur in Glycine residues (labelled G).

defined regions in a graph called a 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.

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. In some fibrous proteins virtually 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. Steric interactions permit a stretch of consecutive residues to be all in the α_R conformation or all in the β conformation, but disallow a helix followed by a strand and vice versa.

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: The hydrophobic effect).
- **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 to 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.

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? It is the sparing solubility of non-polar solutes in water, arising from the microscopic structure of liquid water around such solutes. 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, as 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 nonpolar 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 nonpolar 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 better 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.

The backbone must traverse the interiors of the protein, and carries with it the polar N and O atoms of the peptide groups, 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. However, charged residues are almost completely excluded from protein interiors; in rare cases they form internal salt bridges. Conversely, the surface of a protein is not exclusively charged or polar. About half the residues on the surface of a protein are nonpolar.

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 nonpolar 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. (Don't forget: one or two uncompensated hydrogen bonds and you've 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 secondary structure 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 other conformations 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 can't 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. Two circumstances conspire to make the *process* by which proteins fold to their native states mysterious as well as complex.

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 don't 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 don't 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 to 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 refolding, 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, 100, 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.

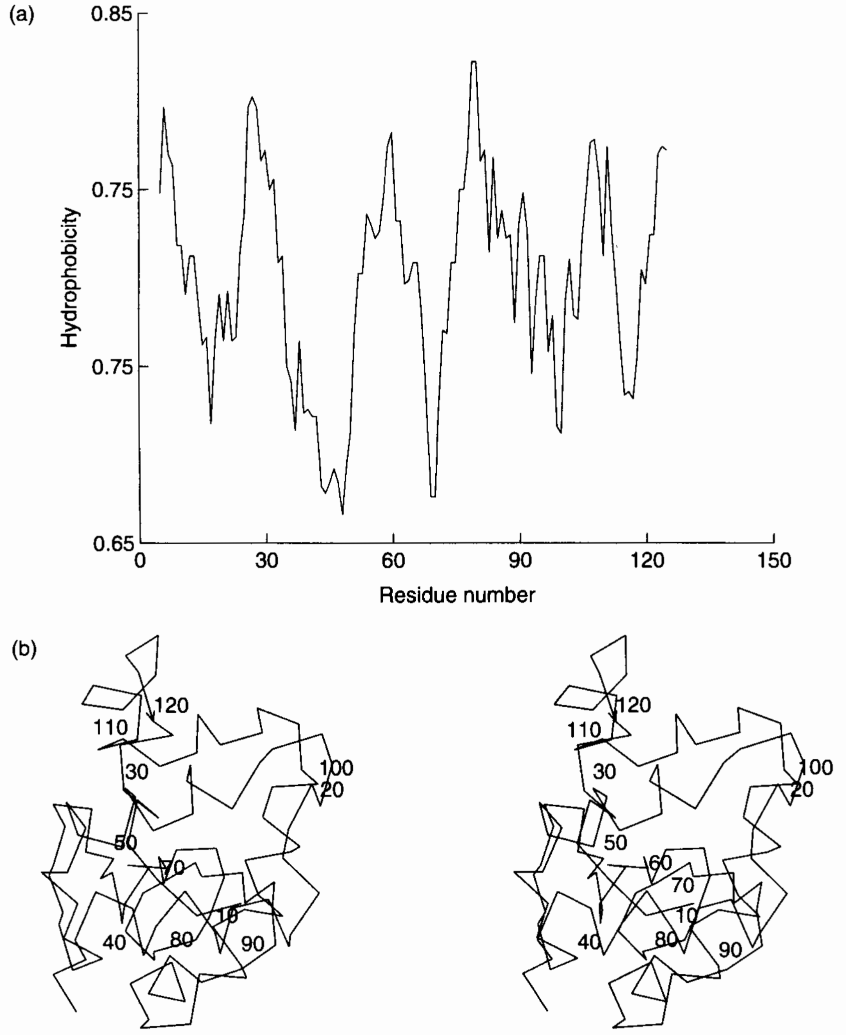


Fig. 5.4 (a) Hydrophobicity profile of hen egg white lysozyme. (Produced using the Primary Structure Analysis tools available through http://www.expasy.org.) (b) Structure of hen egg white lysozyme. Regions corresponding to minima in the hydrophobicity plot are shown in red.

→

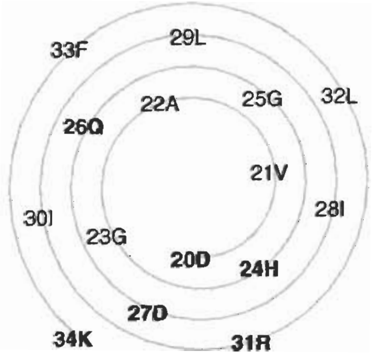
→

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 corresponds to a strand of a β-sheet rather than to 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 α-helix in the native protein structure.

The next box shows a PERL program to draw helical wheels.

PERL Example 5.1 A program to draw helical wheels

```
#!/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
```

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 i^{th} pair of atoms (one atom from each structure) after optimal fitting, and n is the number of points.

This assumes that we have prespecified the correspondence between the points; that is, the alignment.

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, page 235). 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 and measuring the structural similarity of distantly-related proteins. (See Box, page 235.)

Example 5.3 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 5.5 shows a structural superposition of PDB entries 8GCH (γ-chymotrypsin) (black) and 1AGJ (*S. aureus* epidermolytic toxin A) (red). The molecules share the common chymotrypsin-family serine proteinase folding pattern, and the Ser-His-Asp catalytic triad (thicker lines).

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.

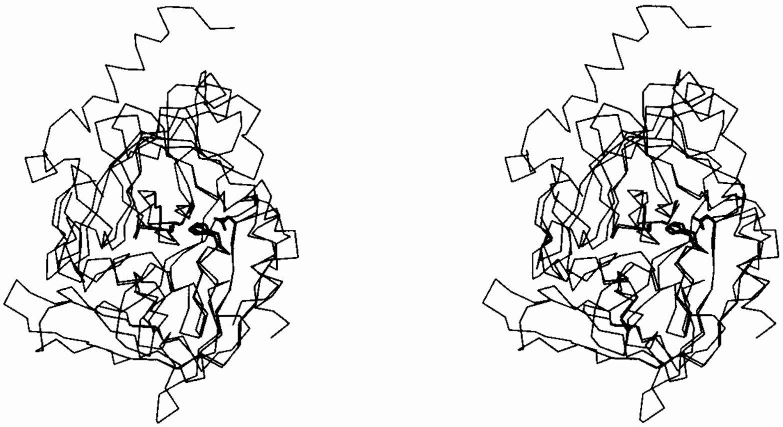


Fig. 5.5 Structural superposition of γ-chymotrypsin [8GCH] (black) and S. *aureus* epidermolytic toxin A [1AGJ] (red). The sidechains of the catalytic triads are shown. Observe that the region around the active site is the best-conserved part of the protein.

Determination of similarity and alignment in computational chemistry

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

$$p_i \longleftrightarrow q_i,\ i = 1, \ldots N.$$

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 possibilities. In the case of proteins or nucleic acids we are limited to correspondences in which we retain the order along the chain:

$$p_{i(k)} \longleftrightarrow q_{j(k)},\ k = 1, \ldots K \leq N, M$$

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 analogous to the Levenshtein distance, or to sequence alignment with gaps. The result of such a calculation is an alignment of parts or all of the sequences.

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

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

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, within two or more molecules, the maximal subsets of atoms that have a similar structure. (See Case Study 5.1.)

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 cause 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 interresidue 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 residue-residue contact patterns in two proteins (this is very easy), and then seeks the maximal matching submatrices

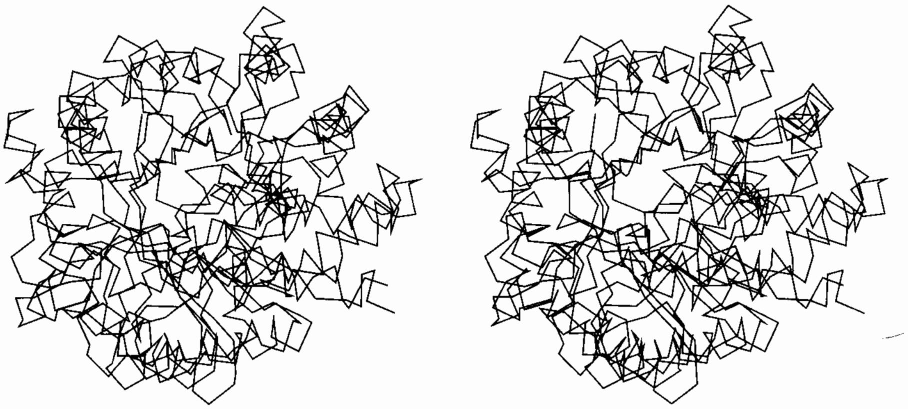


Fig. 5.6 The regions of common fold, as determined by the program DALI by L. Holm and C. Sander, in the TIM-barrel proteins mouse adenosine deaminase [1FKX] (black) and *Pseudomonas diminuta* phosphotriesterase [1PTA] (red). In the alignment shown in this figure, the sequences have only 13% identical residues—closer to midnight than to the twilight zone.

(this is hard). Using carefully chosen approximations, Holm and Sander wrote an efficient program called DALI (for Distance-matrix ALIgnment) 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. 5.6).

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.

Evolution of protein structures

Included in the 30 000 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. 5.5) and the adenosine deaminase–phosphotriesterase family, (Fig. 5.6) 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 determine conserved and variable features, and thereby distinguish that which is crucial to structure and function (and therefore conserved) from that which can survive change (and therefore available to vary).

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 refolding. 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; particular 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 qualitative 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 average no more than 1.0 Å. With increasing sequence divergence, some regions refold 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 5.7a 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 5.7b 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.

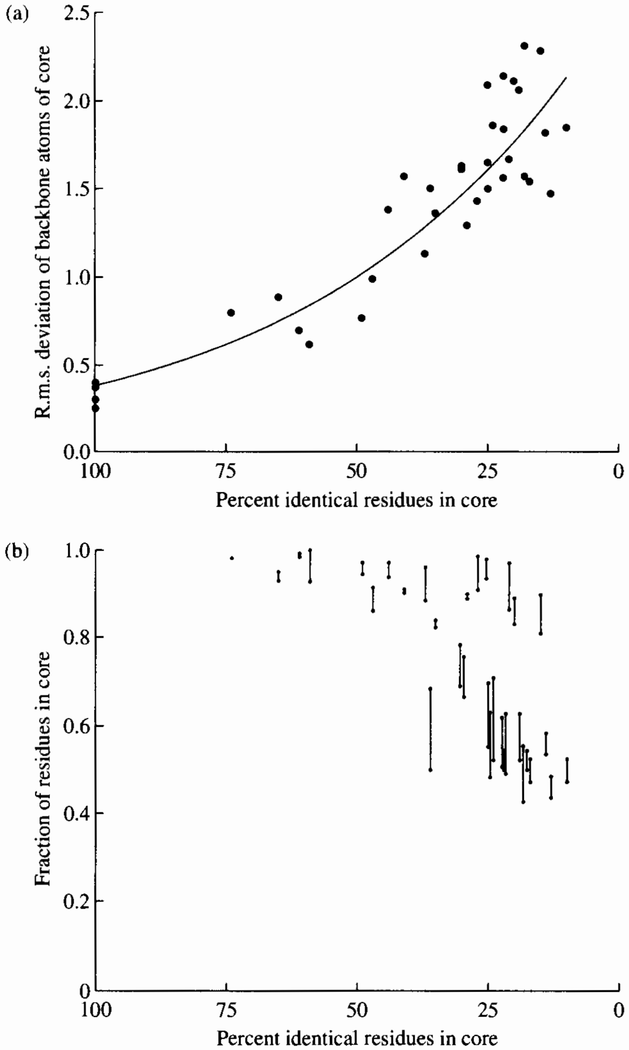


Fig. 5.7 Relationships between divergence of amino acid sequence and three-dimensional structure of the core, in evolving proteins. (a) Variation of r.m.s. deviation of the core with the per cent identical residues in the core. (b) Variation of size of the core with the per cent identical residues in the core. This figure shows results calculated for 32 pairs of homologous proteins of a variety of structural types. (Adapted from Chothia, C. & Lesk, A. M. (1986), Relationship between the divergence of sequence and structure in proteins, *The EMBO Journal*, **5**, 823–826.)

Classifications of protein structures

Being able to measure protein structural differences quantitatively allows us to cluster and classify protein folding patterns. Organization of protein structures according to folding pattern imposes a very useful logical structure on the entries in

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, by I. N. Shindyalov and P. Bourne).

SCOP

SCOP, by A. G. Murzin, L. Lo Conte, B. G. Ailey, S. E. Brenner, T. J. P. Hubbard, and C. Chothia, organizes protein structures in a hierarchy according to evolutionary origin and structural similarity. At the lowest level of the SCOP hierarchy are individual **domains** (see page 42), 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, $\alpha + \beta$, 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 V). For illustrations of the degree of similarities 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.

SCOP classification of Flavodoxin from *Clostridium beijerinckii*

1. **Root** SCOP
2. **Class** α and β proteins (α/β)
   Mainly parallel β-sheets (β-α-β units)
3. **Fold** Flavodoxin-like
   3 layers, $\alpha/\beta/\alpha$; parallel β-sheet of 5 strands, order 21345
4. **Superfamily** Flavoproteins
5. **Family** Flavodoxin-related binds FMN
6. **Protein** Flavodoxin
7. **Species** *Clostridium beijerinckii*

The SCOP release of January 2004 contained 13 220 PDB entries, split into 31 474 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
α + β 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 .

Protein structure prediction and modelling

The observation that each protein folds spontaneously into a unique three-dimensional native conformation implies that nature has an algorithm for predicting protein structure from amino acid sequence. Some attempts to understand this algorithm are based solely on general physical principles; others appeal to known amino acid sequences and protein structures. A proof of our understanding would be the ability to reproduce the algorithm in a computer program that could predict protein structure from amino acid sequence.

Most attempts to predict protein structure from basic physical principles alone try to reproduce the interatomic interactions in proteins, to define a computable energy associated with any conformation. Computationally, the problem of protein structure prediction then becomes a task of finding the global minimum of this conformational energy function. So far this approach has not succeeded, partly because of the inadequacy of the energy function and partly because the minimization algorithms tend to get trapped in local minima.

Other a priori approaches to structure prediction are based on attempts to simplify the problem, to capture somehow the essentials.

The alternative to a priori methods are approaches based on assembling clues to the structure of a target sequence by finding similarities to known structures. These empirical or 'knowledge-based' methods are becoming very powerful.

We are coming closer and closer to saturating the set of possible folds with known structures. This is the stated goal of **structural genomics** projects (see Box). Once we have a complete set of folds and sequences, and powerful methods for relating them, empirical methods will provide pragmatic solutions of many

Structural genomics

In analogy with full-genome sequencing projects, structural genomics has the commitment to deliver the structures of the complete protein repertoire. X-ray crystallographic and NMR experiments will solve a 'dense set' of proteins, such that all proteins are within modelling range of one or more known experimental structures. More so than genomic sequencing projects, structural genomics projects combine results from different organisms. The human proteome is of course of special interest, as are proteins unique to infectious micro-organisms.

The goals of structural genomics have become feasible partly by advances in experimental techniques, which make high-throughput structure determination possible; and partly by advances in our understanding of protein structures, which define reasonable general goals for the experimental work, and suggest specific targets.

The theory and practice of homology modelling suggests that at least 30% sequence identity between target and some experimental structure is necessary. This means that experimental structure determinations will be required for an exemplar of every sequence family, including many that share the same basic folding pattern. Experiment will have to deliver the structures of something like 10 000 domains. In the year 2004, approximately 5000 structures were deposited in the PDB, so the throughput rate is not far from what is required.

Methods of bioinformatics can help select targets for experimental structure determination that offer the highest pay-off in terms of useful information. Goals of target selection include:

- elimination of redundant targets—proteins too similar to known structures.
- identification of sequences with undetectable similarity to proteins of known structure.
- identification of sequences with similarity only to proteins of unknown function, or
- proteins of unknown structure with 'interesting' functions; for example, human proteins implicated in disease, or bacterial proteins implicated in antibiotic resistance.
- proteins with properties favourable for structure dermination—likely to be soluble, contain methonine (which facilitates solving the phase problem of X-ray crystallography).

The machinery for carrying out the modelling is already up and running. MODBASE (http://alto.compbio.ucsf.edu/modbase.cgi/index.cgi) and 3DCrunch (http://www.expasy.org/swissmod/SWISS-MODEL.html) collect homology models of proteins of known sequence.

Structural genomics projects are supported by large-scale initiatives from the US National Institutes of Health and private industry.

problems. What will be the effect of this on attempts to predict protein structure a priori? The intellectual appeal of the problem will still be there. After all, nature folds proteins without searching databases. But it is unlikely that the problem will continue to command interest of the same intensity, and support of the same largesse, once a pragmatic solution has been found.

However, there is a paradox: The methods being developed for identifying folding patterns in sequences are more than exercises in tuning parameters in scoring functions. They are experiments that explore and expose the essential features of amino acid sequences that determine protein structures. When they succeed, we will have a far sounder basis for understanding sequence-structure relationships than we do now. It may be that a posteriori understanding will provide the clues that will make a priori prediction possible.

Methods for prediction of protein structure from amino acid sequence include:

- **Attempts to predict secondary structure** without attempting to assemble these regions in three-dimensions. The results are lists of regions of the sequence predicted to form α-helices and regions predicted to form strands of β-sheet.
- **Homology modelling:** prediction of the three-dimensional structure of a protein from the known structures of one or more related proteins. The results are a complete coordinate set for mainchain and sidechains, intended to be a high-quality model of the structure, comparable to at least a low-resolution experimental structure.
- **Fold recognition:** given a library of known structures, determine which of them shares a folding pattern with a query protein of known sequence but unknown structure. If the folding pattern of the target protein does not occur in the library, such a method should recognize this. The results are a nomination of a known structure that has the same fold as the query protein. or a statement that no protein in the library has the same fold as the query protein.
- **Prediction of novel folds**, either by a priori or knowledge-based methods. The results are a complete coordinate set for at least the mainchain and sometimes the sidechain also. The model is intended to have the correct folding pattern. but would not be expected to be comparable in quality to an experimental structure. D. Jones has likened the distinction between fold recognition and a priori modelling to the difference between a multiple-choice question on an exam and an essay question.

Critical Assessment of Structure Prediction (CASP)

The CASP programmes were introduced briefly in Chapter 1. CASP organizes blind tests of protein structure predictions, in which participating crystallographers and NMR spectroscopists make public the amino acid sequences of the proteins they are investigating, and agree to keep the experimental structures secret until predictors have had a chance to submit their models. CASP runs on a two-year cycle. At the end of the year a gala meeting brings the predictors together to discuss the current results and to gauge progress.

Predictions in CASP have traditionally fallen into three main categories: (1) comparative modelling—in effect homology modelling, (2) fold recognition, and (3) modelling of novel folds:

CASP Category	**Nature of target**
Comparative modelling	Close homologues of known structure are available; homology modelling methods are applicable.
Fold recognition	Structures with similar folds are available, but no sufficiently close relative for homology modelling; the challenge is to identify structures with similar topology.
New Fold	No structure with same folding pattern known; requires either a genuine a priori method or a knowledge-based method that can combine features of several known structures.

Assessors, one for each category, compare the predicted and experimental structures, and judge the predictions. Speakers at the end-of-year meeting include the organizers, the assessors, and selected predictors, including those who have been particularly successful, or who have an interesting novel method to present.

The latest CASP programme took place in 2004. Departures from past practice include: (1) secondary structure prediction is no longer separarately assessed, and (2) a new category, prediction of function, was introduced. There were 87 targets. In all categories, 201 groups of predictors submitted a total of 28 965 models. This was approximately equal to the number of entries in the PDB at the time!

Many predictions are prepared by groups of researchers who inspect the results generated by their computer programs, and select and edit them before submission. In addition, the target sequences are sent to web servers that return predictions without human intervention. The CAFASP: Critical Assessment of Fully Automated Structure Prediction programme monitors the quality of these predictions. It is thereby possible to determine to what extent successful procedures could be made fully automatic. CASP thus comprises three challenges:

Human against protein	CASP
Computer against protein	CAFASP
Human against computer	CASP *v.* CAFASP

A separate programme of blind tests of prediction evaluates methods for predicting protein-protein interactions, or 'docking'. This is CAPRI—Critical Assessment of PRedicted Interactions. Both CASP and CAPRI held assessment meetings in December 2004.

Structure predictions of the sixth CASP programme showed continued improvements. For the most part progress has been incremental rather than spectacular, with one notable exception: David Baker's group predicted and redifined the structure of a small (70-residue) protien from *Thermus thermophilus*, producing a model that deviated by 1.59 Å from the X-ray structure! Indeed, improvements in knowledge-based methods originally developed for novel folds threatens to supersede traditional methods for fold recognition, such as threading, that make explicit reference to libraries of complete structures.

Results at CAPRI show that complexes between partners that do not undergo major conformational changes can now be predicted accurately from the structures of the components. Large conformational changes upon complex formation still present difficulties. However, progress could be seen in at least one case, the trimeric TBE envelope protien.

For both CASP and CAPRI, the best results ae very impressive. One observer commented that the current state of protien structure prediction is that 'failure can no longer be guaranteed.' Consistency is the challenge.

Secondary structure prediction

It seems obvious that (1) it should be easier to predict secondary structure than tertiary structure, and (2) to predict tertiary structure, a sensible way to proceed would be first to predict the helices and strands of sheet and then to assemble them. Whether or not these propositions are correct, many people have believed and acted upon them. Given the amino acid sequence of a protein of unknown structure, they produce **secondary structure predictions**, the assignment of regions in the sequence as helices or strands of sheet.

To assess the quality of a secondary structure prediction, classify the residues in the experimental three-dimensional structure into three categories (helix = H, strand = E (extended), and other = -). The per cent of residues predicted correctly is denoted Q3. At the 2000 CASP programme, the PROF server by B. Rost achieved a good prediction of a domain from the *Thermus aquaticus* mismatch protein MutS. The value of Q3 for Rost's prediction is 81%:

```
                            10        20        30        40        50
                             |         |         |         |         |
Amino acid sequence ALVEDPPLKVSEGGLIREGYDPDLDALRAAHREGVAYFLELEERERERTG
Prediction          HH------------EEE------HHHHHHHHHH-HHHHHHHHHHHHHHH-
Experiment          -E-------------E-----HHHHHHHHHHHHHHHHHHHHHHHHHHHH-

                            60        70        80        90       100
                             |         |         |         |         |
Amino acid sequence IPTLKVGYNAVFGYYLEVTRPYYERVPKEYRPVQTLKDRQRYTLPEMKEK
Prediction          --EEEEEEEEEEEEEEEE-----------EEEEEEEE--EEEE-HHHHHH
Experiment          ----EEEEE---EEEEEEEHHHHHH-----EEEEE---EEEEE-HHHHHH

                           110       120
                             |         |
Amino acid sequence EREVYRLEALIRRREEEVFLEVRERAKRQ
Prediction          HHHHHHHHHHHHHHHHHHHHHHHHHHHH-
Experiment          HHHHHHHHHHHHHHHHHHHHHHHHHHH--
```

Figure 5.8 shows the experimental structure, with the *predicted* secondary structures distinguished. Except for a short 3_{10} helix, the secondary structural elements are predicted correctly except for some minor discrepancies in the positions of

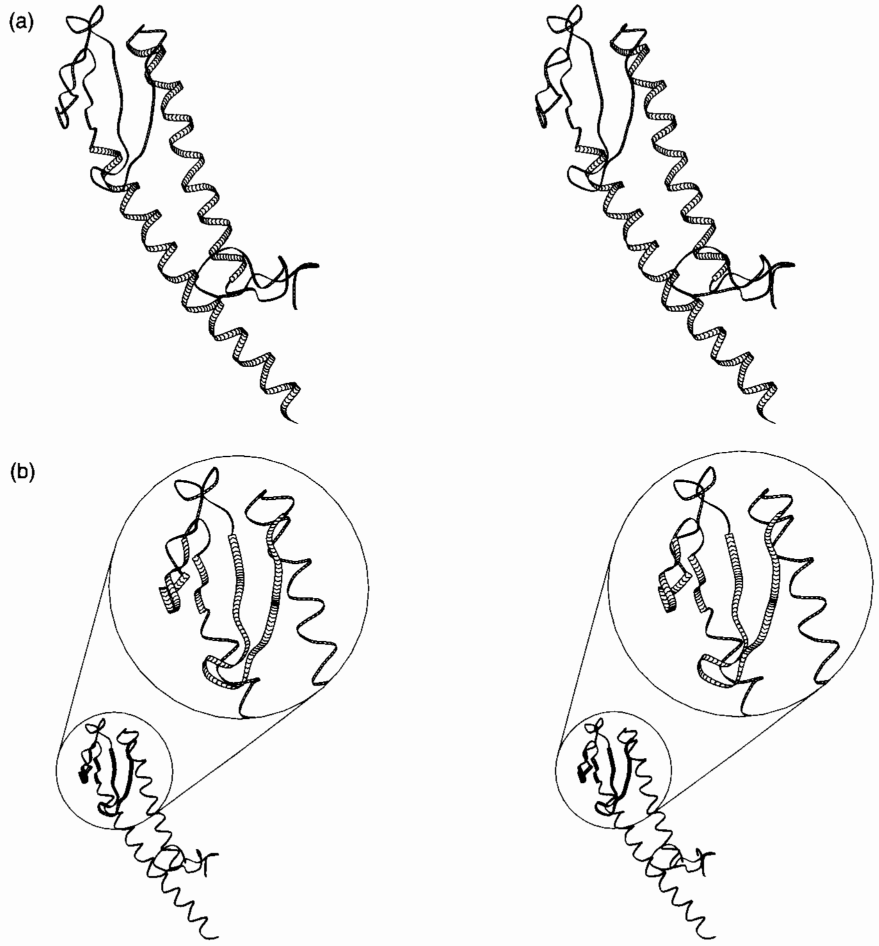


Fig. 5.8 The structure from the *Thermus aquaticus* mismatch repair protein MutS [1EWQ]. (a) The regions predicted by the PROF server of Rost to be helical are shown as wider ribbons. The prediction missed only a short 3_{10} helix, at the top left of the picture. (b) The regions predicted to be in strands are shown as wider ribbons.

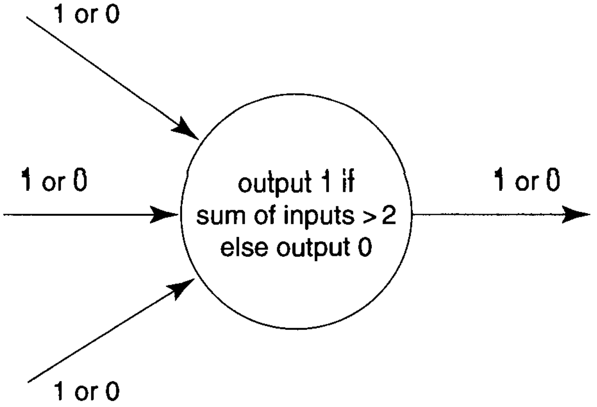
which they start and end. (Other scoring schemes that check for segment overlap are less sensitive to end effects.) The quality of this result is very high but not exceptionally rare. This target was classified as being of *medium* difficulty by the CASP4 assessors. At present, PROF is running at an average accuracy of Q3 ~ 77%. Other secondary structure prediction methods are also doing comparably well.

The most powerful methods of secondary structure prediction are based on **neural networks**.

Neural networks

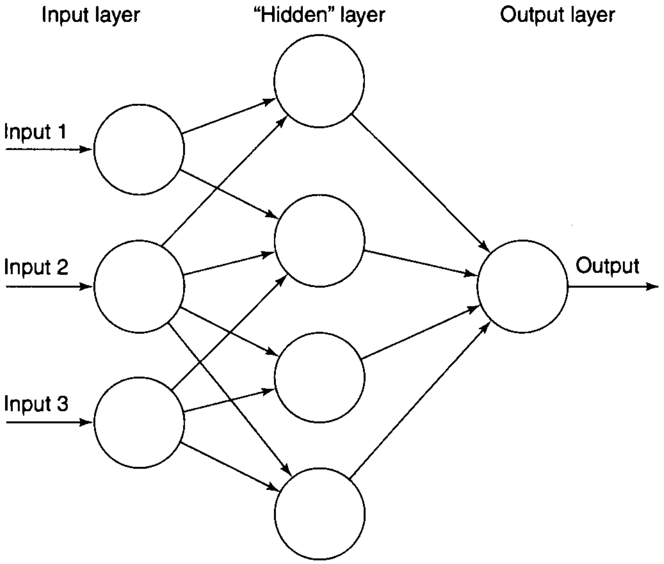
Neural networks are a class of general computational structures based loosely on the anatomy and physiology of biological nervous systems. They have been applied successfully to a wide variety of pattern recognition, classification, and decision problems.

A single neuron, in the computational scheme, is a node in a directed graph, with one or more entering connections designated as input, and a single leaving connection called the output:



In the physiological metaphor, one says that the neuron 'fired' if the output is 1, and that the neuron 'didn't fire' if the output is 0. Simulated neurons can differ in the number of input and output connections, and in the formula for deciding whether to fire (see Box).

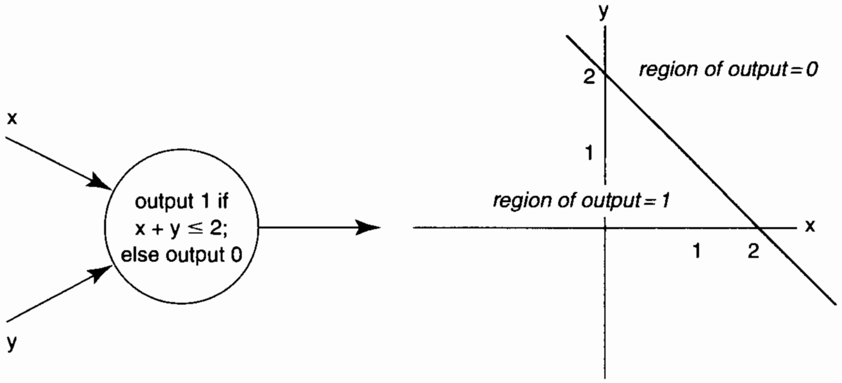
To form a network, assemble several neurons and connect the outputs of some to the inputs of others. Some nodes contain connections that provide input to the entire network; some deliver output information from the network to the outside world; and others, that do not interact directly with the outside, are called **hidden layers**.



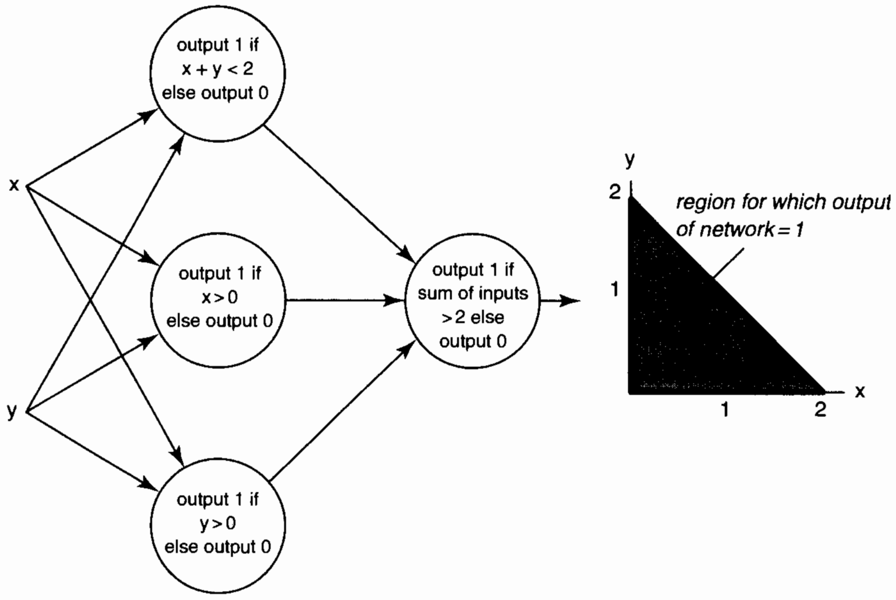
An unlimited degree of complexity is available by assembling and connecting neurons, and by varying the strengths of the connections. That is, instead of taking a simple sum of inputs $i_1 + i_2 + i_3$, take a weighted sum—for instance, $10i_1 + 5i_2 - i_3$, which would make the neuron most sensitive to input 1 and least sensitive to input 3. Biologically, this corresponds to changing the strengths of synapses.

Logic of neural networks

For a single neuron, a linear decision process governing the output has a geometric interpretation in terms of lines and planes. The neuron in the following figure has two inputs. If we interpret the inputs as the coordinates of a point (x, y) in the plane, the neuron 'decides' on which side of a line the input point lies. The output will be 1 if and only if $x + y \leq 2$;that is, if the point is below and to the left of the line $x + y = 2$.



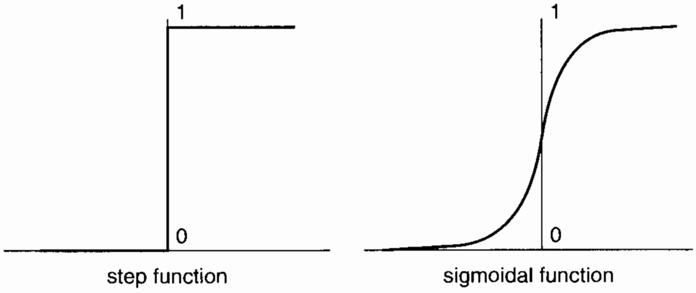
A *neural network* is specified by the topology of its connections, and the weights and decision formulas of its nodes. A network can make more complex decisions than a single neuron. Thus, if one neuron with two inputs can decide on which side of a line a point lies, three neurons can select points that lie within a triangle:



→

Logic of neural networks (*continued*)

Neural networks are more powerful and robust if the output is a smoothly-varying function of the inputs. Such networks can perform more general kinds of computations and are better at pattern recognition. Also, for training the network it is useful if the output is a differentiable function of the parameters. To this end, a sharp threshold function for the output of a neuron is replaced by a smoothed-out step, or sigmoidal, function:



A property of a neural network that gives it great power is that the weights may be regarded as *variables*, and a calculation or learning process may determine the weights appropriate for a particular decision or pattern identifier. To train a network, feed the system sets of sample input for which the desired output is known, and compare the output with the correct answer. If the observed output differs from the desired one, adjust the parameters. The topology of the network remains invariant during the training process, although of course setting a weight to 0 has the effect of detaching an input.

The type of neural network that has been applied to secondary structure prediction is shown in Fig. 5.9.

A major advance in secondary structure prediction occurred with the application of evolutionary information, the recognition that multiple sequence alignment tables contain much more information than individual sequences. The conservation of secondary structure among related proteins means that the sequence-structure correlations are much more robust when a family as a whole is taken into account. Most neural network-based methods for secondary structure prediction now feed the input layer not simply the identities of the amino acid at successive positions, but a profile derived from a multiple sequence alignment.

It has also proved useful to run two neural networks in tandem, to make use of observed correlations among conformations of residues at neighbouring positions. Predictions of the states of several successive residues, by a network similar to the one shown in Fig. 5.9, are combined by a second network into a final prediction.

A test of the maturity of a prediction method is whether it can be made fully automatic. Some computational methods produce only rough drafts of a protein

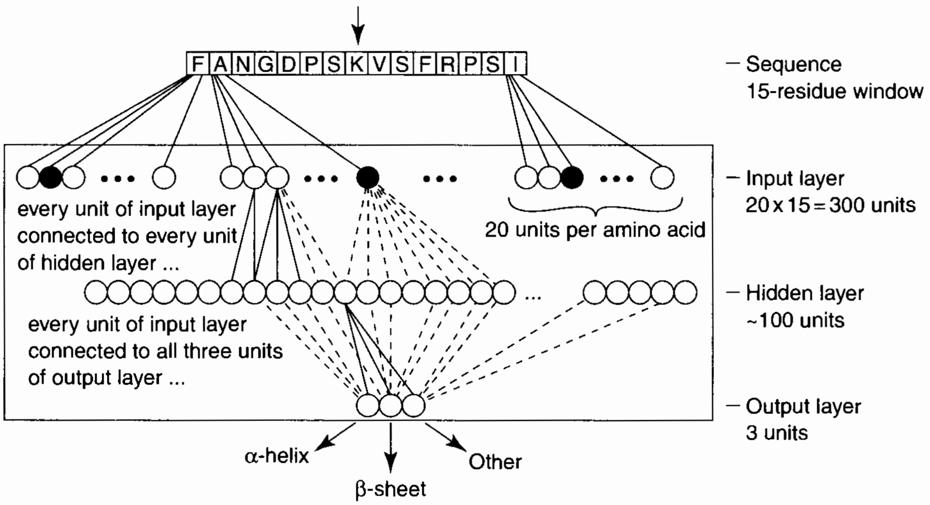

F A N G D P S K V S F R P S I
– Sequence
15-residue window
– Input layer
20 x 15 = 300 units
20 units per amino acid
every unit of input layer connected to every unit of hidden layer ...
– Hidden layer
~100 units
every unit of input layer connected to all three units of output layer ...
– Output layer
3 units
α-helix
β-sheet
Other

Fig. 5.9 A neural network applicable to secondary structure prediction contains three layers:

- The input layer sees a sliding 15-residue window in the sequence. That is, it treats a 15-residue region, predicts the secondary structure of the central residue (marked by an arrow, at the top) and then moves the window one residue along the amino acid sequence and repeats the process. To each of the 15 residues in the current window there correspond 20 nodes in the input layer of the network, one of which will be triggered according to the amino acid in that position.
- A hidden layer of ~100 units connects the input with the output. Each node of the hidden layer is connected to *all* input and output units; not all the connections are shown.
- The output layer consists of only three nodes, that signify prediction that the central residue in the window be in a helix, strand, or other conformation.

structure prediction, requiring human intervention to bring them to final form. Others are automatic, and there are many web servers that will accept sequences and return the predictions. PROF, the system that predicted the secondary structure of MutS, is one of these.

A continuous, fully-automatic, analysis of protein structure prediction web servers, including but not limited to secondary structure predictions, is called EVA. It is a collaboration among groups in New York, San Francisco, and Madrid. The Protein Data Bank supplies sequences shortly in advance of release of the corresponding structures, and the software implementing EVA submits them to prediction servers and analyses the results. It can be thought of as a continuous CASP programme, restricted to methods that can be both applied *and* judged automatically. See: http://cubic.bioc.columbia.edu/eva .

The goals of EVA are to monitor progress in the field, and to indicate to users the best protein structure prediction servers in different categories. EVA has access to much more data than has been available in the CASP programmes themselves. Therefore its conclusions are in principle less vulnerable to statistical fluctuations in the nature and difficulty of the targets than the tests reported in CASP programmes.

Homology modelling

Model-building by homology is a useful technique when one wants to predict the structure of a target protein of known sequence, when the target protein is related to at least one other protein of known sequence *and* structure. If the proteins are closely related, the known protein structures—called the parents—can serve as the basis for a model of the target. Although the quality of the model will depend on the degree of similarity of the sequences, it is possible to specify this quality before experimental testing (see Fig. 5.7). In consequence, knowing the quality of the model required for the intended application permits intelligent prediction of the probable success of the exercise.

Steps in homology modelling are:

1. Align the amino acid sequences of the target and the protein or proteins of known structure. It will generally be observed that insertions and deletions lie in the loop regions between helices and sheets.

2. Determine mainchain segments to represent the regions containing insertions or deletions. Stitching these regions into the mainchain of the known protein creates a model for the complete mainchain of the target protein.

3. Replace the sidechains of residues that have been mutated. For residues that have not mutated, retain the sidechain conformation. Residues that have mutated tend to keep the same sidechain conformational angles, and could be modelled on this basis. However, computational methods are now available to search over possible combinations of sidechain conformations.

4. Examine the model—both by eye and by programs—to detect any serious collisions between atoms. Relieve these collisions, as far as possible, by manual manipulations.

5. Refine the model by limited energy minimization. The role of this step is to fix up the exact geometrical relationships at places where regions of mainchain have been joined together, and to allow the sidechains to wriggle around a bit to place themselves in comfortable positions. The effect is really only cosmetic—energy refinement will not fix serious errors in such a model.

To a great extent, this procedure produces 'what you get for free' in that it defines the model of the protein of unknown structure by making minimal changes to its known relative. Unfortunately it is not easy to make substantial improvements. A rule of thumb (referring again to Fig. 5.8) is that if the two sequences have at least 40–50% identical amino acids in an optimal alignment of their sequences, the procedure described will produce a model of sufficient accuracy to be useful for many applications. For very distantly-related proteins, neither the procedure described nor any other currently available method will produce a model, correct in detail, of the target protein from the structure of its relative.

In most families of proteins the structures contain relatively constant regions and more variable ones. The core of the structure of the family retains the folding topology, although it may be distorted, but the periphery can entirely refold. A single parent structure will permit reasonable modelling of the conserved portion of the target protein, but will fail to produce a satisfactory model of the variable portion. Moreover, it will not be easy to predict which are the variable and constant regions. A more favourable situation occurs when several related proteins of known structure can serve as parents for modelling a target protein. These reveal the regions of constant and variable structure in the family. The observed distribution of structural variability among the parents dictates an appropriate distribution of constraints to be applied to the model.

Mature software for homology modelling is available. SWISS-MODEL is a web site that will accept the amino acid sequence of a target protein, determine whether a suitable parent or parents for homology modelling exist, and, if so, deliver a set of coordinates for the target. SWISS-MODEL was developed by T. Schwede, M. C. Peitsch and N. Guex, now at The Geneva Biomedical Research Institute. Another program in widespread use, MODELLER, was originally developed by A. Šali.

COMPANION WEB SITE

Web resources: Homology modelling

SWISS-MODEL (homology modelling server):
http://www.expasy.ch/swissmod/SWISS-MODEL.html

Results of application of SWISS-MODEL to proteins of known sequence are available through 3DCrunch:
http://www.expasy.org/swissmod/SWISS-MODEL.html

MODELLER (homology modelling software):
http://salilab.org/modeller/modeller.html

Results of application of MODELLER to proteins of known sequence are available through MODBASE:
http://alto.compbio.ucsf.edu/modbase.cgi/index.cgi

For a description of web sites in structural genomics: Wixon, J. (2001), Structural genomics on the web, *Comp. Funct. Genomics*, 2, 103–113.

An example of the automatic prediction by SWISS-MODEL is the prediction of the structure of a neurotoxin from red scorpion (*Buthus tamulus*) from the known structure of the neurotoxin from the related North African yellow scorpion (*Androctonus australis hector*). These two proteins have 52% identical residues in their sequence alignment. With such a close degree of similarity it is not surprising that the model fits the experimental result very closely, even with respect to the sidechain conformation (Fig. 5.10).

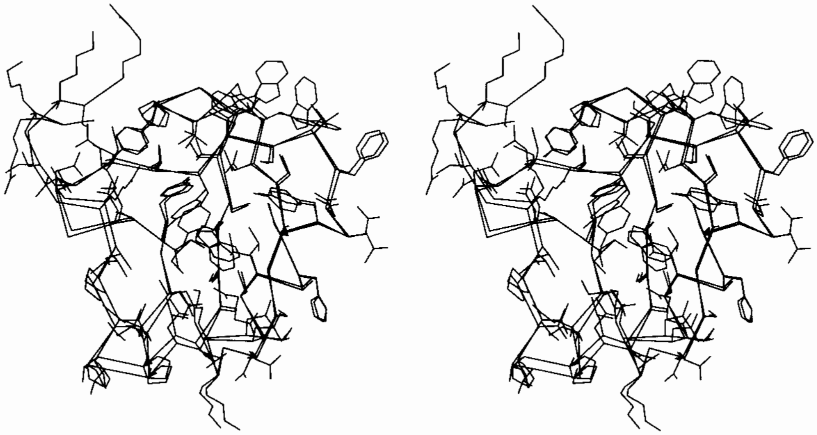


Fig. 5.10 SWISS-MODEL predicts the structure of red scorpion neurotoxin [1DQ7] (red) from a closely-related protein [1PTX] (black). The prediction was done *automatically*. Observe that most of the buried sidechains have not mutated, and have very similar conformations. Some sidechains on the surface have different conformations, and the mainchain of the C-terminus is in a different position (upper left). Not shown is a network of disulphide bridges, which constrain the structure. However, a model of this high quality would be expected for two such closely-related proteins, even without the extra constraints.

Fold recognition

Searching a sequence database for a probe sequence and searching a structure database with a probe structure are problems with known solutions. The mixed problems—probing a sequence database with a structure, or a structure database with a sequence, are less straightforward. They require a method for evaluating the compatibility of a given sequence with a given folding pattern.

The goal is to abstract the essence of a set of sequences or structures. Other proteins that share the pattern are expected to adopt similar structures.

3D profiles

We have discussed patterns and profiles derived from multiple sequence alignments and their application to detection of distant homologues. One way to take advantage of available structural information to improve the power of these methods is a type of profile derived from the available sequences *and* structures of a family of proteins.

J. U. Bowie, R. Lüthy and D. Eisenberg analysed the *environments* of each position in known protein structures and related them to a set of preferences of the 20 amino acids for these structural contexts.

Given a protein structure, they classified the environment of each amino acid in three separate categories:

1. Its mainchain hydrogen-bonding interactions, that is, its secondary structure.

2. The extent to which it is buried within or on the surface of the protein structure.

3. The polar/nonpolar nature of its environment.

The secondary structure may be one of three possibilities: *helix*, *sheet* and *other*. A sidechain is considered buried if the accessible surface area is less than 40 Å^2, partially buried if the accessible surface area is between 40 and 114 Å^2, and exposed if the accessible surface area is greater than 114 Å^2. The fraction of sidechain area covered by polar atoms is also measured. The authors define six classes on the basis of accessibility and polarity of the surroundings. Sidechains in each of these six classes may be in any of three classes according to the secondary structures. This gives a total of 18 classes.

Assigning each sidechain to one of 18 class means that it is possible to write a coded description of a protein structure as a message in an alphabet of 18 letters, called a **3D structure profile**. Algorithms developed for sequence searches can thereby be applied to 'sequences' of encoded structures. For example, one could try to align two distantly-related sequences by aligning their 3D structure profiles rather than their amino acid sequences. The 3D profile method translates protein structures into one-dimensional probe (or probe-able) objects that do not explicitly retain either the sequence or structure of the molecules from which they were derived.

Next, how can one relate the 3D structure profile to the corpus of known sequences and structures? It is clear that some amino acids will be unhappy in certain kinds of sites; for example, a charged sidechain would not be buried in an entirely nonpolar environment. Other preferences are not so clear-cut, and it is necessary to derive a preference table from a statistical survey of a library of well-refined protein structures.

Suppose now that we are given a sequence and want to evaluate the likelihood that it takes up, say, the globin fold. From the 3D structure profile of the known sperm whale myoglobin structure we know the environment class of each position of the sequence. Consider a particular alignment of the unknown sequence with sperm whale myoglobin, and suppose that the residue in the unknown sequence that corresponds to the first residue of myoglobin is phenylalanine. The environment class in the 3D structure profile of the first residue of sperm whale myoglobin is: exposed, no secondary structure. One can score the probability of finding phenylalanine in this structural environment class from the table of preferences of particular amino acids for this 3D structure profile class. (The fact that the first residue of the sperm whale myoglobin sequence is actually valine is not used, and in fact that information is not directly accessible to the algorithm. Sperm whale myoglobin is represented only by the sequence of environment classes of its residues, and the preference table is averaged over proteins with many different folding patterns.) Extension of this calculation to all positions and to all possible alignments not allowing gaps within regions of secondary structure gives a score that measures how well the given unknown sequence fits the sperm whale myoglobin profile.

A particular advantage of this method is that it can be automated, with a new sequence being scored against every 3D profile in the library of known folds, in essentially the same way as a new sequence is routinely screened against a library of known sequences.

Use of 3D profiles to assess the quality of structures

The 3D profile derived from a structure depends only very indirectly on the amino acid sequence. It is therefore meaningful to ask, not only whether it is possible to identify other amino acid sequences compatible with the given fold, but whether the score of a 3D profile for its own parent sequence is a measure of the compatibility of the sequence with the structure. Naturally, if real sequences did not generally appear to be compatible with their own structures, one would be forced to conclude that a useful method for examining the relationship between sequence and structure had not been achieved. Two interesting results are observed: (1) Protein structures determined correctly do fit their own profiles well, although other, related, proteins may give *higher* scores. The profile is abstracting properties of the family, not of individual sequences. (2) When a sequence does *not* match a profile computed from an experimental structure of that protein, there is likely to have been an error in the structure determination. The positions in the profile that do not match can identify the regions of error.

Threading

Threading is a method for fold recognition. Given a library of known structures, and a sequence of a query protein of unknown structure, does the query protein share a folding pattern with any of the known structures? The fold library could include some or all of the Protein Data Bank, or even hypothetical folds.

The basic idea of threading is to build many rough models of the query protein, based on each of the known structures and using different possible alignments of the sequences of the known and unknown proteins. This systematic exploration of the many possible alignments gives threading its name: Imagine trying out all alignments by pulling the query sequence gently through the three-dimensional framework of any known structure. Gaps must be allowed in the alignments, but if the thread is thought of as being sufficiently elastic the metaphor of threading survives.

Both threading and homology modelling deal with the three-dimensional structure induced by an alignment of the query sequence with known structures of homologues. Homology modelling focuses on one set of alignments and the goal is a very detailed model. Threading explores many alignments and deals with only rough models usually not even constructed explicitly:

Homology modelling	**Threading**
First, identify homologues	Try all possible parents
Then, determine optimal alignment	Try many possible alignments
Optimize one model	Evaluate many rough models

Successful fold recognition by threading requires:

1. A method to score the models, so that we can select the best one.
2. A method for calibrating the scores, so that we can decide whether the best-scoring model is likely to be correct.

Several approaches to scoring have been tried. One of the most effective is based on empirical patterns of residue neighbours, as derived from known structures. Observe the distribution of interresidue distances in known protein structures, for all 20 × 20 pairs of residue types. For each pair, derive a probability distribution, as a function of the separation in space and in the amino acid sequence. For instance, for the pair Leu–Ile, consider every Leu and Ile residue in known structures, and, for each Leu–Ile pair, record the distance between their Cβ atoms, and the difference in their positions in the sequence. Collecting these statistics permits estimation of how well the distributions observed in a model agree with the distributions in known structures.

The Boltzmann equation relates probabilities and energies. Usual applications of the Boltzmann equation start from an energy function and predict a probability distribution. (A standard example is the prediction of the density of the atmosphere as a function of altitude, from the gravitational potential energy function of the air molecules.) For threading, one turns this on its head, and *derives* an energy function *from* the probability distribution. This energy function is then used to score threading models.

For each structure in the fold library, the procedure finds the assignment of residues that produces the lowest energy score. Although this is an alignment problem, the nonlocal interactions mean that it can't be solved by dynamic programming.

Fold recognition at CASP

The best methods for fold recognition are consistently effective. These include but are not limited to methods based on threading.

Figures 5.11 and 5.12 show a prediction by A. G. Murzin, and another prediction by Bonneau, Tsai, Ruczinski and Baker, of targets from the 2000 CASP programme, both proteins of unknown function from *H. influenzae*.

Conformational energy calculations and molecular dynamics

A protein is a collection of atoms. The interactions between the atoms create a unique state of maximum stability. Find it, that's all!

The computational difficulties in this approach arise because (a) the model of the interatomic interactions is not complete or exact, and (b) even if the model were exact we should face an optimization problem in a large number of variables, involving nonlinearities in the objective function and the constraints, creating a very rough energy surface with many local minima. Like a golf course with many bunkers, such problems are very difficult.

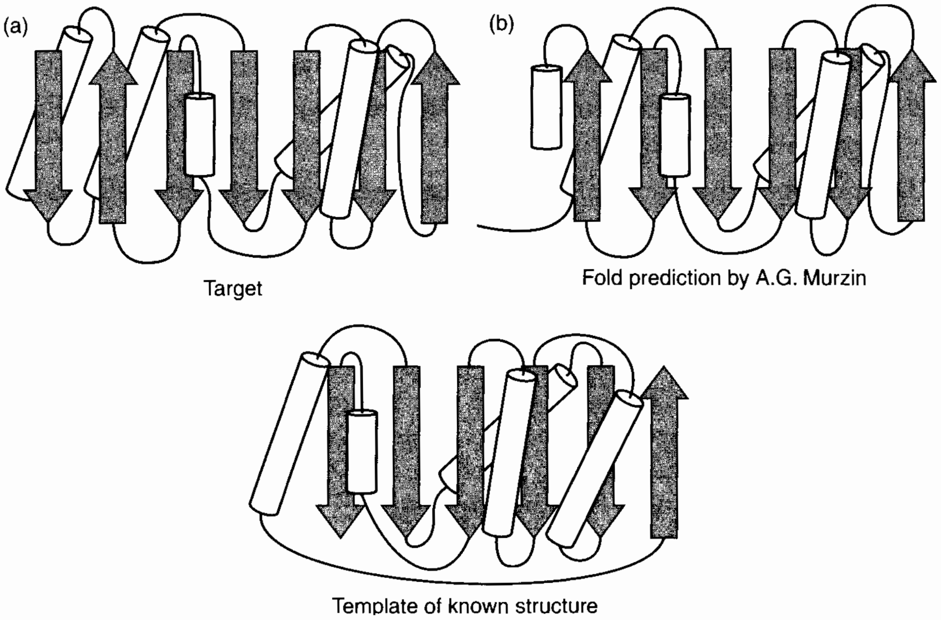


Fig. 5.11 Prediction of structure of *H. influenzae*, hypothetical protein. (a) The folding pattern of the target. (b) Prediction by A. G. Murzin. (c) Folding pattern of the closest homologue of known structure: an N-ethylmaleimide-sensitive fusion protein involved in vesicular transport (PDB entry 1NSF). The topology of Murzin's prediction is closer to the target than that of the closest single parent.

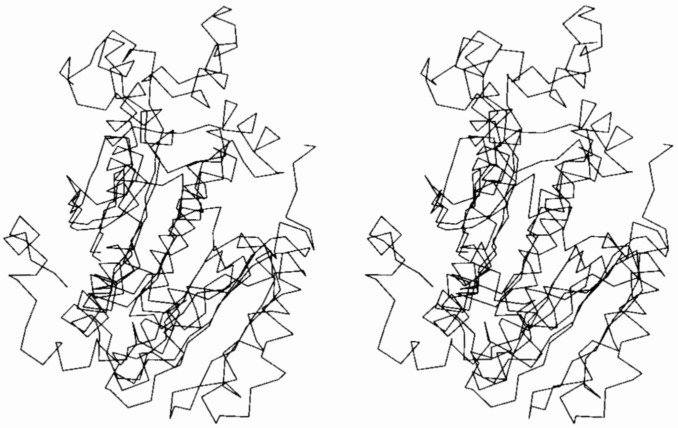


Fig. 5.12 Prediction by Bonneau, Tsai, Ruczinski and Baker of another hypothetical protein from *H. influenzae*, based on glycine N-methyltransferase [1XVA]. Experimental structure, black; prediction, red. Note that much of the prediction superposes well on the experimental structure and that the parts that do not superpose well have similar local structures but improper orientation and packing against the main body of the protein.

The interactions between atoms in a molecule can be divided into:

(a) Primary chemical bonds—strong interactions between atoms that must be close together in space. These are regarded as a fixed set of interactions that are not broken or formed when the conformation of a protein changes, but which, however, are equally consistent with a large number of conformations.

(b) Weaker interactions that depend on the conformation of the chain. These can be significant in some conformations and not in others—they affect sets of atoms that are brought into proximity by different folds of the chain.

The conformation of a protein can be specified by giving the list of atoms in the structure, their coordinates, and the set of primary chemical bonds between them (this can be read off, with only slight ambiguity, from the amino acid sequence). Terms used in the evaluation of the energy of a conformation typically include:

- **Bond stretching** $\Sigma_{bonds}\, K_r(r - r_0)^2$. Here r_0 is the equilibrium interatomic separation and K_r is the force constant for stretching the bond. r_0 and K_r depend on the type of chemical bond.
- **Bond angle bend** $\Sigma_{angles}\, K_\theta(\theta - \theta_0)^2$. For any atom *i* that is chemically bonded to two (or more) other atoms *j* and *k*, the angle *i*–*j*–*k* has an equilibrium value θ_0 and a force constant for bending K_θ.
- Other terms to enforce proper stereochemistry penalize deviations from planarity of certain groups, or enforce correct chirality (handedness) at certain centres.
- **Torsion angle** $\Sigma_{dihedrals}\, \frac{1}{2}V_n[1+\cos n\theta]$. For any four connected atoms: *i* bonded to *j* bonded to *k* bonded to *l*, the energy barrier to rotation by an angle θ of atom *l* with respect to atom *i* around the *j*–*k* bond is given by a periodic potential. V_n is the height of the barrier to internal rotation; *n* barriers are encountered during a full 360° rotation. The mainchain conformational angles ϕ, ψ and ω are examples of torsional rotations (see Fig. 5.2).
- **Van der Waals interactions** $\Sigma_i\, \Sigma_{j<i}\, (A_{ij}R_{ij}^{-12} - B_{ij}R_{ij}^{-6})$. For each pair of non-bonded atoms *i* and *j*, the first term accounts for a short-range repulsion and the second term for a long-range attraction between them. R_{ij} is the distance between atoms *i* and *j*. The parameters *A* and *B* depend on atom type.
- **Hydrogen bond** $\Sigma_i\, \Sigma_{j<i}\, (C_{ij}R_{ij}^{-12} - D_{ij}R_{ij}^{-10})$. The hydrogen bond is a weak chemical/electrostatic interaction between two polar atoms. Its strength depends on distance and also on the bond angle. This approximate hydrogen bond potential does not explicitly reflect the angular dependence of hydrogen bond strength; other potentials attempt to account for hydrogen bond geometry more accurately.

- **Electrostatics** $\Sigma_i \Sigma_{j<i} Q_iQ_j/(\epsilon R_{ij})$. Q_i and Q_j are the effective charges on atoms i and j, R_{ij} is the distance between them, and ϵ is the dielectric 'constant'. This formula applies only approximately to media that are not infinite and isotropic, including proteins.
- **Solvent** Interactions with the solvent, water, and cosolutes such as salts and sugars, are crucial for the thermodynamics of protein structures. Attempts to model the solvent as a continuous medium, characterized primarily by a dielectric constant, are approximations. With the increase in available computer power, it is now possible to include solvent explicitly, simulating the motion of a protein in a box of water molecules.

There are numerous sets of conformational energy potentials of this or closely-related forms, and a great deal of effort has gone into the tuning of parameter sets. The energy of a conformation is computed by summing these terms over all appropriate sets of interacting atoms.

The potential functions satisfy necessary but not sufficient conditions for successful structure prediction. One test is to take the right answer—an experimentally determined protein structure—as a starting conformation, and minimize the energy starting from there. In general most energy functions produce a minimized conformation that is about 1 Å (root-mean-square deviation) away from the starting model. This can be thought of as a measure of the resolution of the force field. Another test has been to take deliberately misfolded proteins and minimize their conformational energies, to see whether the energy value of the local minimum in the vicinity of the correct fold is significantly lower than that of the local minimum in the vicinity of an incorrect fold. Such tests reveal that multiple local minima cannot be reliably distinguished from the correct one on the basis of calculated conformational energies.

Attempts to predict the conformation of a protein by minimization of the conformational energy have so far not provided a method for predicting protein structure from amino acid sequence. In order to overcome the problems both of getting trapped in local minima, and of the absence of a good model for protein-solvent interactions, molecular dynamics models have been developed. The protein plus explicit solvent molecules are treated—via the force field—by classical Newtonian mechanics. It is true that this permits exploration of a much larger segment of phase space. However, as an a priori method of structure prediction, it has still not succeeded consistently. However, these are calculations that are extremely computationally intensive and here, perhaps more than anywhere else in this field, advances deriving from the increasing 'brute force' power of processors will have an effect.

In the meantime, molecular dynamics, if supplemented by experimental data, regularly makes extremely important contributions to structure determinations by both X-ray crystallography (usually) and nuclear magnetic resonance (always). How is molecular dynamics integrated into the process of structure determination? For any conformation, one can measure the consistency of the model with the experimental data. In crystallography, the experimental data are the absolute values of the Fourier coefficients of the electron density of the molecule. In nuclear magnetic resonance, the experimental data provide constraints on the distances between certain pairs of residues. But in both X-ray crystallography (almost always) and nuclear magnetic resonance, the experimental data underdetermine the protein structure. To solve a structure one must seek a set of coordinates that minimizes a combination of the deviation from the experimental data and the conformational energy. Molecular dynamics is successful at determining such coordinate sets: the dynamics provides adequate coverage of conformation space, and the bias derived from the experimental data channels the calculation towards the correct structure.

ROSETTA

ROSETTA is a program by D. Baker and colleagues that predicts protein structure from amino acid sequence by assimilating information from known structures. In recent CASP programmes, ROSETTA showed consistent success on targets in the Novel Fold categories. At present, it leads the field.

ROSETTA predicts a protein structure by first generating structures of fragments using known structures, and then combining them. First, for each contiguous region of 3 and 9 residues, instances of that sequence and related sequences are identified in proteins of known structure. For fragments this small, there is no assumption of homology to the target protein. The distribution of conformations of the fragments serves as a model for the distribution of possible conformations of the corresponding fragments of the target structure.

ROSETTA explores the possible combinations of fragments using Monte Carlo calculations (see Box). The energy function has terms reflecting compactness, paired β-sheets and burial of hydrophobic residues. The procedure carries out 1000 independent simulations, with starting structures chosen from the fragment conformation distribution pattern generated previously. The structures that result from these simulations are clustered, and the centres of the largest clusters presented as predictions of the target structure. The idea is that a structure that emerges many times from independent simulations it is likely to have favourable features.

Figure 5.13 shows successful predictions by ROSETTA of two targets from the 2000 CASP programme.

ROSETTA is available by License or as a webserver.*

LINUS

LINUS (Local Independently Nucleated Units of Structure) is a program for prediction of protein structure from amino acid sequence, by G. D. Rose and R. Srinivasan. It is a completely a priori procedure, making no explicit reference to any known structures or sequence-structure relationships. LINUS folds the

* http://depts.washington.edu/bakerpg

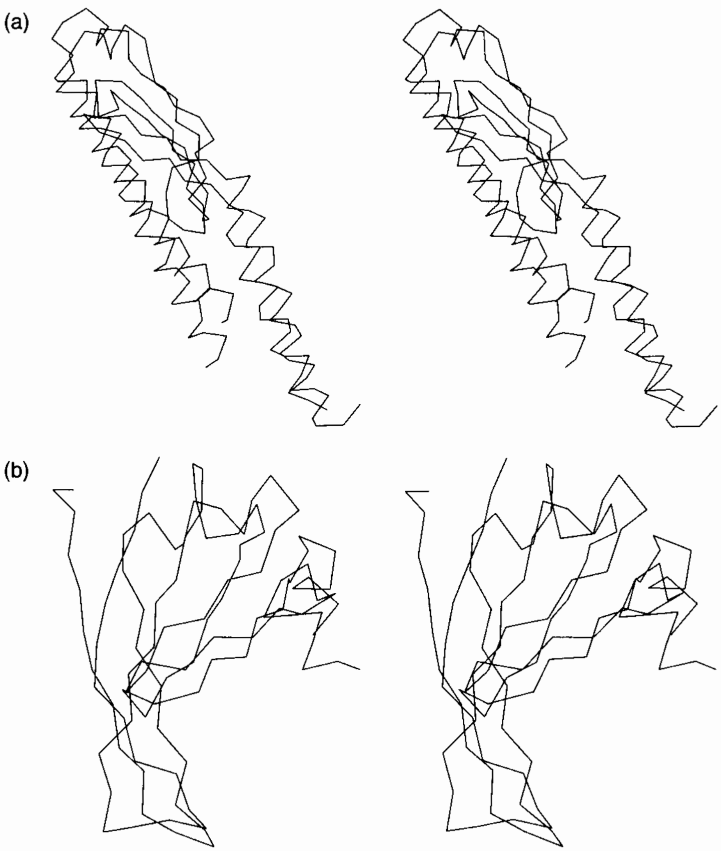


Fig. 5.13 Predictions by ROSETTA of (a) *H. influenzae*, hypothetical protein, (b) The N-terminal half of domain 1 of human DNA repair protein XRCC4. Part b shows a selected substructure containing a 55-residue N-terminal segment (out of a total of 116 residues). Experimental structures, black; predicted structures, red.

polypeptide chain in a *hierarchical* fashion, first producing structures of short segments, and then assembling them into progressively larger fragments.*

An insight underlying LINUS is that the structures of local regions of a protein—short segments of residues consecutive in the sequence—are controlled by local interactions within these segments. During natural protein folding, each segment will preferentially sample its most favourable conformations. However, these preferred conformations of local regions, even the one that will ultimately be adopted in the native state, are below the threshold of stability. Local structure will form transiently and break up many times before a suitable interacting partner stabilizes it. But in the computer one is free to anticipate the results. In a LINUS simulation, favourable structures of local fragments, as determined by their frequent recurrence during the simulation, transmit their

* LINUS is freely available from www.roselab.jhu.edu

Monte Carlo algorithms

Monte Carlo algorithms are used very widely in protein structure calculations, to explore conformations efficiently, and also in many other optimization problems, to search for the minimum of a complicated function. Simple minimization methods based on moving 'downhill' in energy fail because the calculation gets trapped in a local minimum far from the native state.

In general, Monte Carlo methods make use of random numbers to solve problems for which it is difficult to calculate the answer exactly. The name was invented by J. von Neumann, referring to the applications of random number generators in the famous gambling casino.

To apply Monte Carlo techniques to find the minimum of a function of many variables—for instance, the minimum energy of a protein as a function of the variables that define its conformation—suppose that the configuration of the system is specified by the variables x, and that for any values of these variables, we can calculate the energy of the conformation, $\varepsilon(x)$. (x stands for a whole set of variables—perhaps the set of atomic coordinates of a protein, or the mainchain and sidechain torsion angles.)

Then the Metropolis procedure (invented in 1953, allegedly at a dinner party in Los Alamos) prescribes:

1. Generate a random set of values of x, to provide a starting conformation. Calculate the energy of this conformation, $\varepsilon = \varepsilon(x)$.
2. Perturb the variables: $x \rightarrow x'$, to generate a neighbouring conformation.
3. Calculate the energy of the new conformation, $\varepsilon(x')$
4. Decide whether to *accept* the step: to move $x \rightarrow x'$, or to stay at x and try a different perturbation:
   (a) If the energy has decreased; that is, $\varepsilon = \varepsilon(x) > \varepsilon(x')$—in other words the step went *downhill*—always accept it. The perturbed conformation becomes the new current conformation: set $x' \rightarrow x$ and $\varepsilon = \varepsilon(x')$.
   (b) If the energy has increased or stayed the same; that is $\varepsilon(x) \leq \varepsilon(x')$—in other words the step goes *uphill—sometimes* accept the new conformation. If $\Delta = \varepsilon(x') - \varepsilon(x)$, accept the step with a probability $\exp[-\Delta/(kT)]$, where k is Boltzmann's constant, and T is an effective temperature.
5. Return to step 2.

It is step 4b that is the ingenious one. It has the potential to get over barriers, out of traps in local minima. The effective temperature, T, controls the chance that an uphill move will be accepted. T is not the physical temperature at which we wish to predict the protein conformation, but simply a numerical parameter that controls the calculation. For any temperature, the

→

→

higher the uphill energy difference, the less likely that the step will be accepted. For any value of ε, if T is low, then $\varepsilon(x)/(kT)$ will be high, and $\exp[-\varepsilon(x)/(kT)]$ will be relatively low. If T is high, then $\varepsilon(x)/(kT)$ will be low, and $\exp[-\varepsilon(x)/(kT)]$ will be relatively high. The higher the temperature, the more probable the acceptance of an uphill move.

This relatively simple idea has proved extremely effective, with successful applications including but by no means limited to protein structure calculations.

Simulated annealing is a development of Monte Carlo calculations in which T varies—first it is set high to allow efficient exploration of conformations, then it is reduced to drop the system into a low-energy state.

preferred conformations as biases that influence subsequent steps. The procedure applies the principle of a rachet to direct the calculation along productive lines.

LINUS begins by building the polypeptide from the sequence as an extended chain. The simulation proceeds by perturbing the conformations of a succession of randomly-chosen three-residue segments, and evaluating the energies of the results. Structures with steric clashes are rejected out of hand; other energetic contributions are evaluated only in terms of local interactions. A Monte Carlo procedure (see Box) is used to decide whether to accept a perturbed structure or revert to its predecessor. LINUS performs a large number of such steps. It periodically samples the conformations of the residues, to accumulate statistics of structural preferences.

Subsequent stages in the simulation assemble local regions into larger fragments, using the conformational biases of the smaller regions to guide the process. The window within the sequence controlling the range of interactions is progressively opened, from short local regions, to larger ones, and ultimately to the entire protein.

The LINUS representation of the protein folding process is realistic in essential respects, although approximate. All non-hydrogen atoms of a protein are modelled, but the energy function is approximate and the dynamics simplified. The energy function captures the ideas of: (1) steric repulsion preventing overlap of atoms, (2) clustering of buried hydrophobic residues, (3) hydrogen bonding, and (4) salt bridges.

Currently LINUS is generally successful in getting correct structures of small fragments (size between supersecondary structure and domain), and in some cases can assemble them into the right global structure. Figure 5.14 shows the LINUS prediction of the C-terminal domain of rat endoplasmic reticulum protein ERp29.

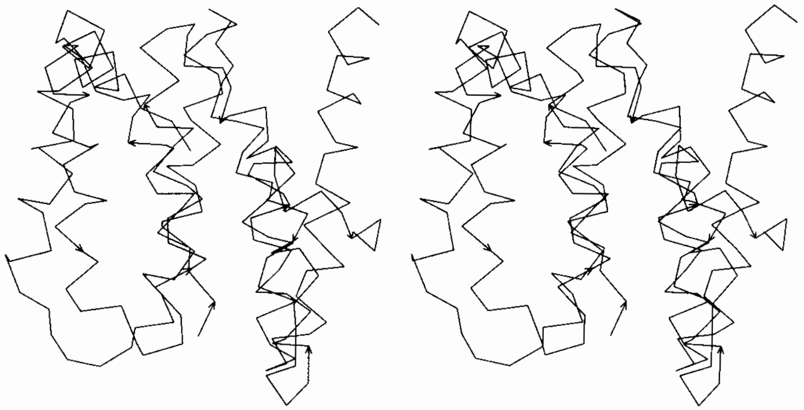


Fig. 5.14 A LINUS prediction of the C-terminal domain of rat endoplasmic reticulum protein ERp29. Experimental structure, black; prediction, red.

Assignment of protein structures to genomes

A genome sequence is the complete statement of a potential life. Assignment of structures to gene products is a first step in understanding how organisms implement their genomic information.

We want to understand the structures of the molecules encoded in a genome, their individual activities and interactions, and the organization of these activities and interactions in space and time during the lifetime of the organism. We want to understand the relationships among the molecules encoded in the genome of one individual, and their relationships to those of other individuals and other species.

For individual proteins, knowing their structure is essential for understanding the mechanism of their function and interactions. For entire organisms, knowing the structures tells us how the repertoire of possible protein folds is used, and how it is distributed among different functional categories in different species. For interspecies comparisons, protein structures can reveal relationships invisible in highly-diverged sequences.

Several methods have been applied to structure assignment:

- **Experimental structure determination.** The best way of all!
- **Detection of homology in sequences.** Sophisticated sequence comparison methods such as PSI-BLAST or Hidden Markov Models can identify relationships between proteins, both within an organism and between species. If the structure of any homologue is known experimentally, at least the general fold of the family can be inferred.
- **Fold-recognition methods** can assign folds to some proteins even in the absence of evidence for homology.
- Specialized techniques detect **membrane proteins**, and **coiled-coils**.

The results of structure assignments provide partial inventories of proteins in the different genomes, and, for the subset of proteins with sufficiently close relatives of known structure, detailed three-dimensional models. The degree of coverage of assignments is changing very fast, primarily because of the rapid growth of sequence and structural data. The table contains a current scorecard.

Species	Number of sequences	Structures assigned	%
E. coli	4289	916	21
M. jannaschi	1773	262	15
S. cerevisiae	6289	1109	18
D. melanogaster	13687	2990	22

(From: GeneQuiz, http://jura.ebi.ac.uk:8765/ext-genequiz/)

What do these results tell us about the usage of the potential protein repertoire? At present, proteins of known structure fall into approximately 750 fold classes, out of an estimated total of 1000. A comparison of folds deduced from the genomes of an archaeon *Methanococcus jannaschii*, a bacterium *Haemophilus influenzae*, and a eukaryote *Saccharomyces cerevisiae*, revealed that out of a total of 148 folds, 45 were common to all three species—and by implication, probably common to most forms of life. The archaeon, *M. jannaschii*, had the fewest unshared folds (see Fig. 5.15).

An inventory of the structures common to all three species showed that the five most common folding patterns of domains are: (1) the P-loop-containing NTP hydrolase fold, (2) the NAD-binding domain, (3) the TIM-barrel fold, (4) the flavodoxin fold,

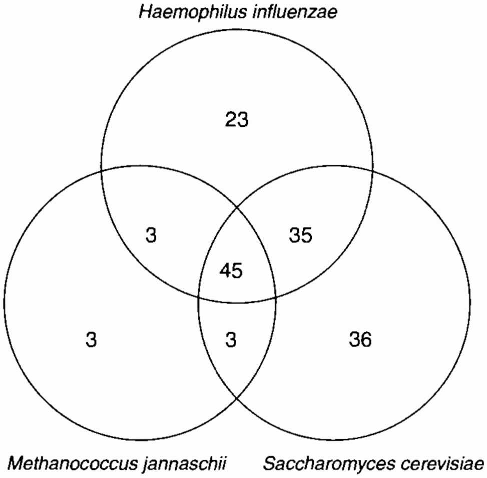


Fig. 5.15 Shared protein folds in an archaeon *Methanococcus jannaschii*, a bacterium *Haemophilus influenzae*, and a eukaryote *Saccharomyces cerevisiae*. (From: Gerstein, M. (1997), A structural census of genomes: comparing bacterial, eukaryotic, and archaeal genomes in terms of protein structure, *J. Mol. Biol.*, **274**, 562–576.)

and (5) the thiamin-binding fold. Plate VI shows the structure and a simplified schematic diagram of the topology of the last of these (see also Weblems 5.3 and 5.4). All are of the α/β type.

Prediction of protein function

The cascade of inference should ideally flow from sequence → structure → function. However, although we can be confident that similar amino acid sequences will produce similar protein structures, the relation between structure and function is more complex. Proteins of similar structure and even of similar sequence can be recruited for very different functions. Conversely, very widely diverged proteins may retain similar functions. Moreover, just as many different sequences are compatible with the same structure, proteins with different folds can carry out the same function.

As proteins evolve they may:

1. retain function and specificity,
2. retain function but alter specificity,
3. change to a related function, or a similar function in a different metabolic context,
4. change to a completely unrelated function.

People often ask: How much must a protein sequence or structure change before the function changes? The answer is: Some proteins have multiple functions, so they need not change at all!

- In the duck, an active lactate dehydrogenase and an enolase serve as crystallins in the eye lens, although they do not encounter the substrates *in situ*. In other cases crystallins are closely related to enzymes, but some divergence has already occurred, with loss of catalytic activity. (This proves that the enzymatic activity is not necessary in the eye lens.)
- A protein from *E. coli*, called Do or DegP or HtrA, acts as a chaperone (catalysing protein folding) at low temperatures, but at 42 °C turns into a proteinase. The rationale seems to be: under normal conditions or moderate heat stress the goal is to salvage proteins that are having difficulty folding; under more severe heat stress, when salvage is impossible, to recycle them.
- We have mentioned already the *E. coli* enzyme lipoate dehydrogenase that is *also* an essential subunit of pyruvate dehydrogenase, 2-oxoglutarate dehydrogenase and the glycine cleavage complex.

These examples of structure-function relationships are at the extreme end of a spectrum that can include a wide range of behaviour.

One problem is that it is not easy to define the idea of difference in function quantitatively. When are two different functions more similar to each other than two other different functions? In some cases altered function may conceal similarity of mechanism. For example, the enolase superfamily contains several homologous enzymes that catalyse different reactions with shared mechanistic features. This group includes enolase itself, mandelate racemase, muconate lactonizing enzyme I, and D-glucarate dehydratase. Each acts by abstracting an α proton from a carboxylic acid to form an enolate intermediate. The subsequent reaction pathway, and the nature of the product, vary from enzyme to enzyme. These proteins have very similar overall structures, a variant of the TIM-barrel fold. Different residues in the active site produce enzymes that catalyse different reactions.

Divergence of function: orthologues and paralogues

The family of chymotrypsin-like serine proteinases includes closely-related enzymes in which function is conserved, and widely diverged homologues that have developed novel functions. Trypsin, a digestive enzyme in mammals, catalyses the hydrolysis of peptide bonds adjacent to a positively-charged residue, Arg or Lys. (A **specificity pocket**, a surface cleft in the active site, is complementary in shape and charge distribution to the sidechain of the residue adjacent to the scissile bond.) Enzymes with similar sequence, structure, function, and specificity exist in many species, including human, cow, Atlantic salmon, and even *Streptomyces griseus* (Fig. 5.16). The similarity of the *S. griseus* enzyme to vertebrate trypsins suggests a lateral gene transfer. For the three vertebrate enzymes, each pair of sequences has ≥64% identical residues in the alignment, and the bacterial homologue has ≥30% identical residues with the others; all have very similar structures. These enzymes are called **orthologues**—homologous proteins in different species. (Other bacterial homologues are very different in sequence.)

Evolution has also created related enzymes in the *same* species with different specificities. Chymotrypsin and pancreatic elastase are other digestive enzymes that, like trypsin, cleave peptide bonds, but next to different residues: Chymotrypsin cleaves adjacent to large flat hydrophobic residues (Phe, Trp) and elastase cleaves adjacent to small residues (Ala). The change in specificity is effected by mutations of residues in the specificity pocket. Another homologue, leukocyte elastase (the object of database searching in Chapter 3) is essential for phagocytosis and defence against infection. Under certain conditions it is responsible for lung damage leading to emphysema.

Homologous proteins in the same species are called **paralogues**. Trypsin, chymotrypsin and pancreatic elastase function in digestion of food. Another set of paralogues mediates the blood coagulation cascade. Although all are proteinases, the requirements for activation and control are very different for digestion and blood coagulation, and the families have diverged and become specialized for these respective roles.

Some homologues of trypsin have developed entirely new functions:

- Haptoglobin has lost its proteolytic activity. It acts as a chaperone, preventing unwanted aggregation of proteins. Haptoglobin forms a tight complex with

```
                           10        20        30        40        50        60        70        80
                            |         |         |         |         |         |         |         |
Human              IVGGYNCEENSVPYQVSLNSGYHFCGGSLINEQWVVSAGHCYKSR---IQVRLGEHNIEVLEGNEQFINAAKIIRHPQYD
Cow                IVGGYTCGANTVPYQVSLNSGYHFCGGSLINSQWVVSAAHCYKSG---IQVRLGEDNINVVEGNEQFISASKSIVHPSYN
Atlantic salmon    IVGGYECKAYSQAHQVSLNSGYHFCGGSLVNENWVVSAAHCYKSR---VEVRLGEHNIKVTEGSEQFISSSRVIRHPNYS
S. griseus         VVGGTRAAQGEFPFMVRLSMG---CGGALYAQDIVLTAAHCVSGSGNNTSITATGGVVDLQSGAAVKVRSTKVLQAPGYN
                   iVGGy c     p qVsLnsGyhfCGGsL n  wVvsA HCyks      vrlge ni v eG eqfi   k i hP Y

                           90       100       110       120       130       140       150       160
                            |         |         |         |         |         |         |         |
Human              RKTLNNDIMLIKLSSRAVINARVSTISLPTAPPATGTKCLISGWGNTASSGADYPDELQCLDAPVLSQAKCEASYP-GKI
Cow                SNTLNNDIMLIKLKSAASLNSRVASISLPTSCASAGTQCLISGWGNTKSSGTSYPDVLKCLKAPILSDSSCKSAYP-GQI
Atlantic salmon    SYNIDNDIMLIKLSKPATLNTYVQPVALPTSCAPAGTMCTVSGWGNTMSSTADS-NKLQCLNIPILSYSDCNNSYP-GMI
S. griseus         --GTGKDWALIKLAQPINQ----PTLKIATTTAYNQGTFTVAGWGANREGGSQQRYLLKAN-VPFVSDAACRSAYGNELV
                        nDimLIKL   a  n  v    lpT     gt c  sGWGnt ssg      L cl  P lS   C   YP g i

                          170       180       190       200       210       220       230
                            |         |         |         |         |         |         |
Human              TSNMFCVGFLE-GGKDSCQGDSGGPVVCNG-----QLQGVVSWGDGCAQKNKPGVYTKVYNYVKWIKNTIAANS
Cow                TSNMFCAGYLE-GGKDSCQGDSGGPVVCSG-----KLQGIVSWGSGCAQKNKPGVYTKVCNYVSWIKQTIASN-
Atlantic salmon    TNAMFCAGYLE-GGKDSCQGDSGGPVVCNG-----ELQGVVSWGYGCAEPGNPGVYAKVCIFNDWLTSTMASY-
S. griseus         ANEEICAGYPDTGGVDTCQGDSGGPMFRKDNADEWIQVGIVSWGYGCARPGYPGVYTEVSTFASAIASAARTL-
                   t  mfC G le GGkDsCQGDSGGPvvc g      lqG VSWG GCA    PGVYtkV     wi  t a
```

Fig. 5.16 Alignment of sequences of trypsins from human, cow, Atlantic salmon and *Streptomyces griseus*. In the lines under the blocks, uppercase letters indicate absolutely conserved residues and lowercase letters indicate residues conserved in three of the four sequences (in most but not all cases the *S. griseus* sequence is the exception).

haemoglobin fragments released from erythrocytes, with several useful effects including preventing the loss of iron.

- The serine proteinase of rhinovirus has developed a separate, independent function, of forming the initiation complex in RNA synthesis, using residues on the opposite side of the molecule from the active site for proteolysis. This is not a modification of an active site—it is the creation of a new one.
- Subunits homologous to serine proteinases appear in plasminogen-related growth factors. The role of these subunits in growth factor activity is not yet known, but it cannot be a proteolytic function because essential catalytic residues have been lost.
- An antifreeze glycoprotein in Antarctic fish is homologous to trypsinogen.
- The insect 'immune' protein scolexin is a distant homologue of serine proteinases that induces coagulation of haemolymph in response to infection.

In the chymotrypsin family we see a retention of structure with similar functions in closely-related proteins, and progressive divergence of function in some but not all distantly-related ones.

The message is that the overall folding pattern of a protein is an unreliable guide to predicting function, especially for very distant homologues. For correct prediction of function in distantly-related proteins it is necessary to focus on the active site. For example:

- J. F. Bazan and R. Fletterick, and, independently, P. Argos, G. Kamer, M. J. Nicklin, and E. Wimmer, recognized that viral 3C proteinases are homologues of chymotrypsin, despite the fact that the serine of the catalytic triad is changed to cysteine.
- W. R. Taylor and L. Pearl recognized the distant homology between retroviral and aspartic proteinases from conserved Asp, Thr, and Gly residues.

Like motif libraries such as PROSITE, such approaches go directly from signature patterns of active-site residues in the sequence to conserved function, even in the absence of an experimental structure.

In focussing on the active site, there is opportunity to use methods similar to those used in drug design in designing ligands, to predict ligands that might bind to the proteins. It is important to make use of other experimental information available, such as tissue distribution patterns of expression, and catalogues of proteins that interact. Attempts to measure function directly, for instance by means of 'gene knockouts', will sometimes provide an answer, but are unproductive if the knockout phenotype is lethal or if there are multiple proteins that share a function.

It seems likely that the contribution of bioinformatics to prediction of protein function from sequence and structure will not be a simple algorithm that provides an unambiguous answer (as there is hope will someday be the case for prediction of structure from sequence). More reasonable aims are to suggest productive experiments and to contribute to the interpretation of the results. These are not unworthy goals.

Drug discovery and development

It is a sobering experience to ask a classroom full of students how many would be alive today without at least one course of drug therapy during a serious illness. (This ignores diseases escaped through vaccination.) Or to ask the students how many of their surviving grandparents would be leading lives of greatly reduced quality without regular treatment with drugs. The answers are eloquent. They engender fear of the new antibiotic-resistant strains of infectious micro-organisms. Other drugs target human proteins, to deal with protein dysfunction or to adjust regulatory controls. It is necessary to develop new drugs, which, in combination with genomic information that can improve their specificity, will extend and improve our lives.

However, it is not easy to be a drug. For a chemical compound to qualify as a drug, it must be:

1. safe
2. effective
3. stable—both chemically and metabolically
4. deliverable—the drug must be absorbed and make its way to its site of action
5. available—by isolation from natural sources or by synthesis
6. novel, that is, patentable

Steps in the development of new drug are summarized in the Box. The process involves scientific research, clinical testing to prove safety and efficacy, and very important economic and legal aspects involving patent protection and estimation of returns on the very high required investment.

Steps in the development of a new drug

1. Understanding the biological nature and symptoms of a disease. Is it caused by:
   - an infectious agent—bacterium, virus, other?
   - a poison of nonbiological origin?
   - a mutant protein in the patient?
2. Developing an assay. Given a candidate drug, can you test it by:
   - its effect on the growth of a micro-organism?
   - its effect on cells grown in tissue culture?
   - its effect on animals that suffer the disease or an analogue?
   - its binding to a known protein target?

→

→

Steps in the development of a new drug (*continued*)

3. Is an effective agent from a natural source known from folklore/tradition? If so, go to 6.
4. Identify a specific molecular target, usually a protein. Determine its structure experimentally or by model-building.
5. Get a general idea of what kind of molecule would fit the site on the target. Is there a known substrate or inhibitor?
6. Identification of a *lead compound:* a chemical that shows the desired biological activity to *any* measurable extent. A lead compound is a bridgehead; finding lead compounds and subsequently modifying them are quite different kinds of activities.
7. Development of the lead compound: Extensive study of variants of the compound, with the goal of building in all the desired properties and enhancing the biological activity.
8. Preclinical testing, *in vitro* and with animals, to prove effectiveness and safety. At this point the drug may be patented. (In principle, one wants to delay patenting as long as possible because of finite lifetime of the patent, and many lengthy steps still remain before the drug can be sold.)
9. In the USA: submission of an Investigational New Drug Application to the Federal Drug Administration (FDA). This is followed by three phases of clinical trials.
10. Phase I clinical trials. Test the compound for safety on healthy volunteers. Determine how the body deals with the drug—how it is absorbed, distributed, metabolized, excreted. The results suggest a safe dosage range.
11. Phase II clinical trials. Test the compound for efficacy against a disease on approximately 200 volunteer patients. Does it cure the disease or alleviate symptoms? Calibrate the dosage.
12. Phase III clinical trials. Test approximately 2000 patients, to demonstrate conclusively that the compound is better than the best known treatment. These are randomized double-blind tests, either against a placebo or against a currently-used drug. These trials are very expensive; it is not uncommon to kill a project before embarking on this step, if the phase II trials expose side effects or unsatisfactory efficacy.
13. File a New Drug Application with the FDA, containing supporting data proving safety and efficacy. FDA approval allows selling the drug. Only now can the drug generate income.
14. Phase IV studies, subsequent to FDA approval and marketing, involve continued monitoring of the effects of the drug, reflecting the wider experience in its use. New side effects may turn up in some classes of patients, leading to restrictions on the use of the drug, or even possibly its recall.

To develop a drug, first you must choose a target disease. You will want to study what is known about its possible causes, its symptoms, its genetics, its epidemiology, its relationship to other diseases—human and animal—and all known treatments. Assuming that the potential utility of a drug justifies the major time, expense, and effort required to develop one, you are now ready to begin.

From the target disease, you must select a target protein. About 500 proteins are the targets of known drugs.

You must develop a suitable assay with which to detect success in the initial phase. If a known protein is the target, binding can be measured directly. A potential anti-bacterial drug can be tested by its effect on growth of the pathogen. Some compounds might be tested for effects on eukaryotic cells grown in tissue culture. If a laboratory animal is susceptible to the disease, compounds can be tested on animal subjects. However, compounds may have different effects on animals and humans. For example, tamoxifen, now a drug used widely against breast cancer, was originally developed as a birth-control pill. In fact it is a fine contraceptive for rats but *promotes* ovulation in women.

The lead compound

A goal in the early stages of drug development is identification of one or more **lead compounds.** A lead compound is any substance that shows the biological activity you seek. It demonstrates that a compound exists that possesses at least some of the desired properties.

There are a number of ways to find lead compounds:

1. Serendipity: penicillin is the classic example.
2. Survey of natural sources. 'Grind and find' is the medicinal chemist's motto. Sometimes traditional remedies point to a source of active compounds. For example, digitalis was isolated from leaves of the foxglove, which had been used for congestive heart failure. (Why not just continue to use the traditional remedy? Isolation of the active principle makes it possible to regulate dosage, and to explore variants.) Approximately half the drugs in current use are based on natural products.
3. Study what is known about substrates, inhibitors, and the mechanism of action, and select potentially active compounds from these properties.
4. Consider drugs effective against similar diseases.
5. Large-scale screening. Techniques of combinatorial chemistry permit parallel testing of large sets of related compounds. A special technique applicable to polypeptides is phage display.
6. Occasionally, from side effects of existing drugs. Minoxidil (2,4-diamino-6-piperidino-pyrimidine-3-oxide), originally designed as an antihypertensive, was found to induce hair growth. Viagra, originally developed as heart medicine, is another example.

7. Experimental screening. The US National Cancer Institute has screened tens of thousands of compounds. (Screening of variants is also very important *after* a lead compound has been found.)
8. Computer screening and *ab initio* computer design.

Discovery of a lead compound triggers other kinds of research activities. Many variants of the lead compound must be tested to improve its effectiveness, and to build in other essential properties. For instance, a compound that binds to its target *in vitro* is no good as a drug unless it can get to the target *in vivo*. Deliverability of a drug to a target within the body requires the capacity to be absorbed and transported. It requires metabolic stability, and 'shelf' stability. It requires the proper solubility profile—a drug must be sufficiently water-soluble to be absorbed, but not so soluble that it is excreted immediately; it must (in most cases) be sufficiently lipid-soluble to get across membranes, but not so lipid-soluble that it is merely taken up by fat stores. There must be a reasonable synthetic route to produce the compound in quantity.

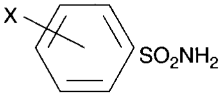
Improving on the lead compound: Quantitative Structure-Activity Relationships (QSAR)

For any compound with pharmacological activity, similar compounds typically exhibit related activity but vary in potency and specificity. Starting with a lead compound, chemists must survey large numbers of related molecules to optimize desired pharmacological properties. To search systematically, it would be very useful to understand how the variation in structural and physicochemical features in the family of molecules is correlated with pharmacological properties. The problem is that there are very many possible descriptors for characterizing molecules. These include structural features such as the nature and distribution of substituents; experimental features such as solubility in aqueous and organic solvents, or dipole moments; and computed features such as charges on individual atoms.

Quantitative Structure-Activity Relationships (QSAR) provide methods for predicting the pharmacological activity of a set of compounds from the relationship between molecular features and pharmacological activity, based on test cases. The method was developed by C. Hansch and colleagues in the 1960s, and has been of very widespread use.

C. Hansch, J. McClarin, T. Klein and R. Langridge applied QSAR methods to study inhibitors of carbonic anhydrase. Carbonic anhydrase is an enzyme that catalyses the reaction $CO_2 + H_2O \rightleftharpoons H^+ + HCO_3^-$. Clinical applications of carbonic anhydrase inhibitors include diuretics, treatment of high interocular pressure in glaucoma by supressing secretion of aqueous humour (the fluid within the eye), and anti-epileptic agents. High-altitude climbers take carbonic anhydrase inhibitors for relief of symptoms of acute mountain sickness.

Measurements of carbonic anhydrase binding of 29 phenylsulphonamides:



(X stands for a set of substituents on the ring that are variable in both structure and position) showed that the binding constant was related to the Hammett electronic substituent constant σ, a measure of the electron-withdrawing or -donating strength of the substituent; the octanol-water partition coefficient P of the unionized form of the ligand; and the location (*ortho* or *meta*) of the substitution:

$$\log K \approx 1.55\sigma + 0.65 \log P - 2.07I_1 + 3.28I_2 + 6.94$$

in which K = binding constant, $I_1 = 1$ if X is *meta* and 0 otherwise, and $I_2 = 1$ if X is *ortho* and 0 otherwise. The substituents X were of the form -alkyl, or -COO-alkyl, or -CONH-alkyl.

This type of correlation has two implications:

1. A large number of compounds can be screened in the computer and those predicted to be the best tested experimentally.
2. It is possible to visualize the binding site from analysis of the parameters:

   - The positive coefficient of σ, implying that electron-withdrawing substituents are favoured, suggests that the ionized form of the $-SO_2NH_2$ moiety binds to the Zn ion in the carbonic anhydrase active site.
   - The positive coefficient of $\log P$ suggests a hydrophobic interaction between the protein and ligand.
   - The negative coefficients of I_1 and I_2 suggest steric clashes with substituents in the *meta* or *ortho* positions.

Structures of ligated carbonic anhydrase confirm these conclusions (see Weblem 5.9).

Bioinformatics in drug discovery and development

Computing and information retrieval contribute to several steps in drug discovery and development projects. These include: target identification; design, analysis, and enhancement of ligands; and selection and *in silicio* screening of libraries. Information systems are also important in the organization of the theoretical predictions, the experimental designs, and analysis of the data. D. Searls has called the intimate interplay between theory and experiment 'wet-dry cycles'.

Target selection

To develop a drug against a disease, it is necessary to select a protein linked to the disease in a way that suggests that it would be therapeutically useful to affect its function or expression. New high-throughput data sources, particularly of genome sequences and protein expression patterns, provide a rich source of material for identifying potential drug targets. **Differential genomics and proteomics,** the comparisons of healthy and diseased humans or animals, can pinpoint which particular protein is missing, dysfunctional, improperly regulated, or expressed only in affected cells. Information about protein-protein complexes make it possible to target not just a single protein, but a specific protein-protein interaction.

Knowledge of prokaryotic and viral genomes supports identification of targets for drugs against infectious disease. Of particular interest are metabolic pathways specific to micro-organisms, and the proteins that participate in them. A drug affecting such a target is less likely to interact with a human homologue, with consequent side effects. Proteins with sequences similar across bacterial clades offer the possibility of broad-spectrum antibiotics. Conversely, gene duplications warn of potential redundant functions, with concomitant insensitivity to inactivation of the target. Knowledge of the relative speed of evolution of different proteins, including horizontal gene transfer rates, indicates the expected stability of a therapy against development of resistant strains.

Commitment to a target by a large pharmaceutical company involves a very heavy investment of resources. The profit expected to flow from a successful drug exerts a very important influence on the choice of targets actively pursued. Analysis of the history of drugs that currently yield high profits suggests that prediction of economic returns is not a very precise science. Now, even generously-supported bioinformatics efforts are much less expensive than laboratory work. The possibility that calculations will improve predictions and enhance profit is behind the espousal of bioinformatics by the pharmaceutical industry, in addition to the purely scientific contributions of bioinformatics to drug discovery. This contribution to economic forecasting is especially important when a company considers high-risk projects, such as those aimed at developing a drug against a new class of targets. Such projects must compete with lower-risk activities such as trying to improve on a competitor's success.

Prediction of a lead compound

Methods for predicting ligands suitable as lead compounds for drug discovery can be divided into inductive and deductive approaches.

Inductive methods depend on correlations between known affinities of some test set of compounds, and molecular features characterizing entire libraries of potential ligands. These features include structural properties such as size, geometry, charge distributions and specific functional groups including hydrogen-bond donors and acceptors. They include general 'drug-like' qualities such as solubility in aqueous and organic solvents, easy route of administration, appropriate distribution in body tissues and metabolic turnover rate. Medicinal chemists apply an equivalent of the duck test: if it walks like a drug, swims like a drug and quacks like a drug, then maybe it will be a drug. The relevant characteristics of compounds are compiled into a **feature vector** used to compare the overall match between compounds of known affinity and a complete library. The requirements for organization, encoding, storage, and searching of information about small molecules has created a new field, **chemoinformatics**, which complements bioinformatics in applications to drug discovery.

Deductive methods are applicable if the binding site on the target protein is known or can be inferred. However, because binding affinity and specificity are only two requirements for a lead compound—admittedly essential ones—it is necessary to combine deductive methods with the correlation to desirable properties as in the purely inductive approach. Binding assays on purified systems give little idea of the behaviour of a compound as a drug in its biological context. Bioinformatics has a contribution to make in integrating the information available from molecular and cell biology, and physiology and pharmacology, to help bridge the gap between *in vitro* experiments and therapeutic activities.

Molecular modelling in drug discovery

A central problem in drug discovery is the identification of a compound that will bind tightly and specifically to a target protein. Tight binding is necessary for efficacy at low concentrations. Specificity is necessary to minimize side effects.

If the structure of the target is known from experiment, it is possible to apply molecular modelling directly to ligand design. If the structure of the target is unknown, a picture of the binding site must be created from indirect evidence, and ligand design is correspondingly more difficult. Ligand design without the target structure is like trying to catch a bank robber from eyewitness descriptions; ligand design to a target of known structure is like trying to catch the bank robber from a clear image on a CCTV videotape.

Goals of molecular modelling applied to drug design include:

- Ideally: suggestion of a lead compound that already shows reasonable affinity and specificity. This is a rare achievement.
- Analysis of compounds known to bind to the target. Understanding the important interactions serves as a guide to design and testing of potential ligands, and for selecting structural features to build into combinatorial synthesis of libraries. In the case of antibacterial or antiviral projects, a model of the protein-ligand complex can give some idea of how easy it would be for the pathogen to develop resistance by mutations that lower the affinity.
- **Pharmacophore** identification is the extraction of common substructures of many compounds that share a pharmacological activity, or at least that bind to the same site on a protein. The hypothesis is that there is some common constellation of atoms within the structures that is responsible. The computational problem of extracting the pharmacophore from a set of compounds is similar to that of structural alignment of a set of homologous proteins. However, although typical ligands are much smaller than proteins, the combinatorial problems are more severe because one has lost the linear ordering of the residues in proteins. (see page 235.) Inferred pharmacophore properties are integrated with QSAR methods to filter libraries of compounds for candidate ligands.
- *In silicio* screening: predicting of affinities, even qualitatively, suggests candidate ligands from a library of chemical structures. The results can be used either for setting priorities in experimental tests, or be integrated into broader approaches to computer screening of libraries on the basis of features correlated with favourable chemical and pharmacological properties. Many readers will be aware of the harnessing of screensavers worldwide to search for potential drugs.[†] At present, over 2.6 million computers have joined this project. They have contributed a cumulative total of over 320 000 years of CPU power.

see Box, Docking: prediction of ligand geometry and affinity

[†] http://www.chem.ox.ac.uk/curecancer.html

- **Lead compound improvement:** Once a compound is identified that binds to a target protein, albeit with low affinity and specificity, interactive modelling can suggest modifications that are expected to enhance the fit. Synthesis and testing of compounds predicted to show enhanced affinity, and even solution of crystal structures of their complexes, can guide the search for improved compounds. The modelling is usually coupled with combinatorial chemistry and experimental screening of libraries of compounds.

Docking: prediction of ligand geometry and affinity

Docking is prediction of ligand binding. In includes prediction both of binding of small molecules to proteins, and of protein-protein binding. The goals of docking are (1) to identify the binding site on the protein, and determine the position and orientation of the ligand, and (2) to estimate the affinity.

(1) Identification of mode of binding Docking of small molecules to proteins requires matching of the ligand to a site on a protein of known structure. The binding site may be known in advance, or it may be necessary to try many different modes of apposition of the ligand and protein to predict the optimal binding site.

The basis for docking is the identification of complementarity in size, shape, and distribution of charge, polarity, and potential for hydrophobic and hydrogen-bonding interactions. A complication is the possibility of flexibility in both partners. Small organic molecules containing many single bonds have a high degree of conformational flexibility. (Drug designers love structures with rings and bridges.) Many proteins show conformational changes upon binding ligands. Therefore the experimental structure of an unligated protein cannot be assumed to serve as a rigid target for docking. However, allowing for flexibility complicates docking calculations substantially.

Water molecules at interfaces present another difficulty. They can contribute to the surface complementarity, and provide bridging hydrogen bonds.

(2) Estimation of affinity It is difficult to estimate absolute affinities. However, comparative docking can provide useful information about *relative* affinities. A suitable scoring function, that can predict the ranking of different ligands in approximate order of affinity, allows selectivity, and setting of priorities, in experimental testing. Such scoring schemes can be *ab initio*—based on the kinds of force fields described on pages 257–258—or empirical. Conversely, comparative docking of one ligand to many proteins can predict the specificity of the interaction.

→

Compare:

Docking calculation	Information provided
1 ligand –1 protein	mode of binding, estimate of affinity
many ligands –1 protein	ranking of affinities of a series of potential ligands
1 ligand – many proteins	prediction of specificity

Docking and scoring are important steps in the filter between a full potential library and testing at the bench. A typical narrowing of the funnel might run as follows:

overall library size	10^{12} compounds
after general filters	10^5
docking	10^4
scoring	10^3
visual	10–100 for experimental testing

Two case studies illustrate the range of chemical and molecular-biological techniques involved in drug development, and show some interesting similarities and contrasts. They concern two well-known families of analgesic drugs—colloquially, 'pain-killers'—typified by morphine and aspirin. The two groups of compounds have different mechanisms of actions, different potencies, and different spectra of side effects.

Case Study 5.1: Development of analgesic drugs based on morphine[‡]

Morphine and codeine are natural alkaloids contained in the latex of the opium poppy (*Papaver somniferum*) (Fig. 5.17). The pharmacological effects have been known since antiquity. Modern chemistry has explored and developed many variants. Heroin was synthesized in 1874 (Fig. 5.17). More hydrophobic than the natural compounds, heroin traverses the blood-brain barrier more readily, giving it a more rapid onset of action.

Both codeine and heroin are metabolized to produce morphine, the active form. Codeine is therefore a natural example of a **prodrug**, an inactive agent that is converted to an active one. The conversion depends on a cytochrome, CYP2D6, which is absent in 5–10% of Caucasians and 1-3% of Afro-Americans and Asians, in whom codeine is ineffective.

Morphine and codeine have been applied in medicine and surgery as analgesics, drugs to relieve severe pain. Side effects include passivity and euphoria,

[‡] Coop, A. & MacKerell, A. D. Jr. (2000), The future of opioid analgesics, *Amer. J. Pharm. Educ.*, **66**, 153–156.

→

Case Study 5.1 *(continued)*

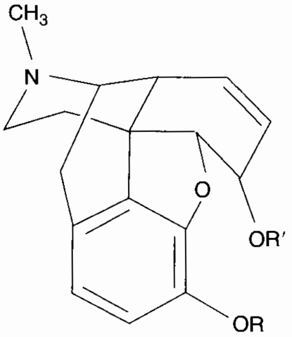


Fig. 5.17 Morphine, codeine and heroin have structures differing only in substituents at two positions:

Compound	**R**	**R′**
Morphine	-H	-H
Codeine	$-CH_3$	-H
Heroin	$-COCH_3$	$-COCH_3$

and physical dependence and addiction. Drug developers have therefore long sought a compound that would relieve pain without the harmful side effects. Of course there was no guarantee that this would be possible.

Synthetic variants of morphine allow correlation of biological effects with chemical structure.

One approach is to try to simplify the structure. The goals are (1) to infer the minimal pharmacophore required for activity, and (2) if possible, to dissect the parts of the structure that relieve pain from those causing addiction. Morphine, codeine and heroin are rigid compounds containing five fused rings. Levorphanol differs from morphine by loss of the bridging oxygen (removal of the tetrahydrofuran ring) and one of the hydroxyl groups (Fig. 5.18). It is a more potent analgesic than morphine but still addictive. Benzomorphan, cyclazocine and pentazocine break the cyclohexene ring (Fig. 5.19). The addictive effects of these compounds are lesser than those of morphine and levorphanol. Demerol, which opens the cyclohexene ring, and methadone, which has *no* fused rings, retain analgesic activity, sharing even smaller common substructures with morphine.

From these structures, one can infer the pharmacophore shown in Fig. 5.20.

In contrast to simplifying the molecule to identify a pharmacophore, attempts to enhance specificity have retained the pharmacophore but made the molecule more complex. Some success has been achieved. Etorphine and buprenorphine, discovered in the 1960s, are far more powerful analgesics

→

than morphine (etorphine is used for sedation of large animals), and have lower addictive potential (see Fig. 5.21). Indeed, the most important clinical use of buprenorphine is in treatment of drug addiction, rather than in analgesia.

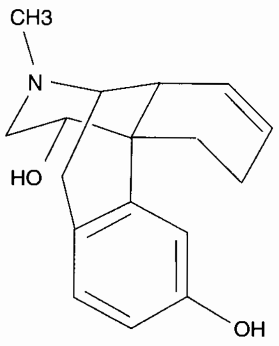


Fig. 5.18 The structure of levorphanol.

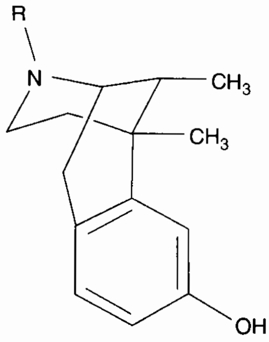


Fig. 5.19 The structures of benzomorphan: R = CH_3; cyclazocine: R = CH_2-cp (cp = cyclopropane); pentazocine: R = $CH_2CH= C(CH_3)_2$.

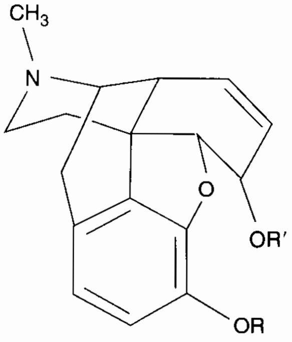


Fig. 5.20 Pharmacophore (red) derived from structural comparisons among morphine derivatives. (After A. D. MacKerell, Jr.)

→

Case Study 5.1 *(continued)*

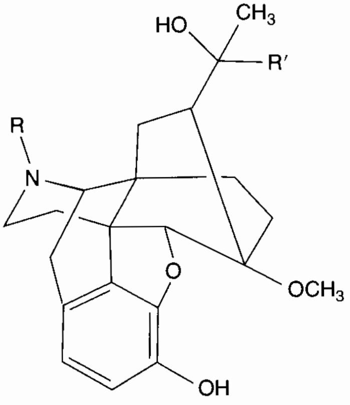


Fig. 5.21 The strucures of etorphine: R = CH_3, R′ = C_3H_7; buprenorphine: R = CH_2-cp (cp = cyclopropane), R′ = *t*-butyl.

This exploration of variants went on before the natural receptors were identified. We now know that the natural targets of action of morphine and related molecules are receptors for endogenous peptides called endorphins. These include:

β-endorphin	YGGFMTSEKSQTPLVTLFKNAIIKNAYKKGE
dynorphin	YGGFLRRIRPKLKWDNQ

and their cleavage products:

Met-enkephalin	YGGFM
Leu-enkephalin	YGGFL

Morphine is therefore a natural **peptidomimetic**, a non-peptide that shares a structure and activity with a peptide.

Several classes of receptors are known, including μ, κ, and δ types, and a recently-discovered fourth type, called ORL-1 (ORL = opiate-receptor like). Their sequences are about 50–70% identical at the residue level. They are G-protein-coupled receptors, similar in structure to bacteriorhodopsin (see Fig. 4.7). Different ligands—natural and synthetic—have different affinity to different receptors, and different kinetics of binding and dissociation. The natural targets of morphine are μ receptors. It is thought that μ receptors tend to be more involved in physical dependence and addiction than κ receptors, although this statement of the situation is extremely oversimplified. Nevertheless, it suggests that to produce a drug that provides analgesia with reduced side effects one should look at the *distribution* of affinities of compound for the different types of receptor.

Case Study 5.2: Computer-aided drug design: specific inhibitors of prostaglandin cyclooxygenase 2

Prostaglandins are a family of natural compounds that mediate a wide variety of physiological processes. Pharmacological applications include the use of prostaglandins themselves, and, conversely, drugs that block prostaglandin synthesis. Prostaglandin E_2 (dinoprostone) is used in obstetrics to induce labour. Aspirin, ibuprofen, acetaminophen (tylenol), and other **nonsteroidal anti-inflammatory drugs (NSAIDs)** are effective against arthritis and related diseases (see Box, page 283). They achieve this effect by inhibiting enzymes in the pathway of prostaglandin synthesis, specifically, prostaglandin cyclooxygenases. A well-known side effect of aspirin is bleeding from the walls of the stomach. This occurs because prostaglandins (the production of which aspirin inhibits) suppress acid secretions by the stomach and promote formation of a mucus coating protecting the stomach lining.

Aspirin and other NSAIDs inhibit *two* closely-related prostaglandin cyclooxygenases, called COX-1 and COX-2. (Unfortunately the same abbreviations are used for cytochrome oxidases 1 and 2.) COX-1 is expressed constitutively in the stomach lining. COX-2 is inducible, and up-regulated in response to inflammation. This suggests that a drug that would inhibit COX-2 but not COX-1 would retain the desired activity of NSAIDs but reduce unwanted side effects. [Note added at time of going to press: some COX-2 inhibitors have recently been implicated in increased risk of cardiovascular disease.]

The amino acid sequences and crystal structures of COX-1 and COX-2 are known. (These proteins have 65% sequence identity.) Figure 5.22 shows part of the structure of COX-1, acetylated by the aspirin analogue 2-bromoacetoxybenzoic acid (aspirin brominated on the methyl group of the acetyl moiety). The salicylate moiety binds nearby. The effect is to block the entrance to the active site. Most NSAIDs bind but do not covalently modify the enzyme.

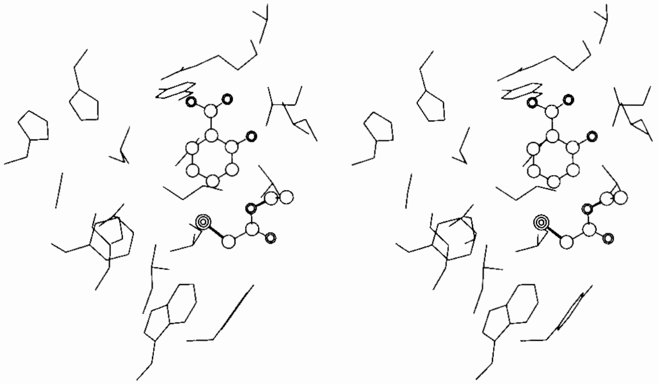


Fig. 5.22 The binding site in COX-1 for an aspirin analogue, 2-bromoacetoxybenzoic acid. The ligand has reacted with the protein, transferring the bromoacetyl group to the sidechain of serine 530. The protein is shown in skeletal representation, in black. The aspirin analogue is shown in ball-and-stick representation, in red.

→

Case Study 5.2 *(continued)*

Figure 5.23 shows the same figure with the corresponding region of COX-2 superposed. Can you see regions of structural difference, that could be clues to the design of selective drugs? Figure 5.24 shows the region of COX-2 with the selective inhibitor SC-558 (1-phenylsulphonamide-3-trifluoromethyl-5-parabromophenylpyrazole, made by Searle). From Fig. 5.25 we can see why SC-558 cannot inhibit COX-1. There would be steric clashes with the isoleucine sidechain, which corresponds to a valine in COX-2.

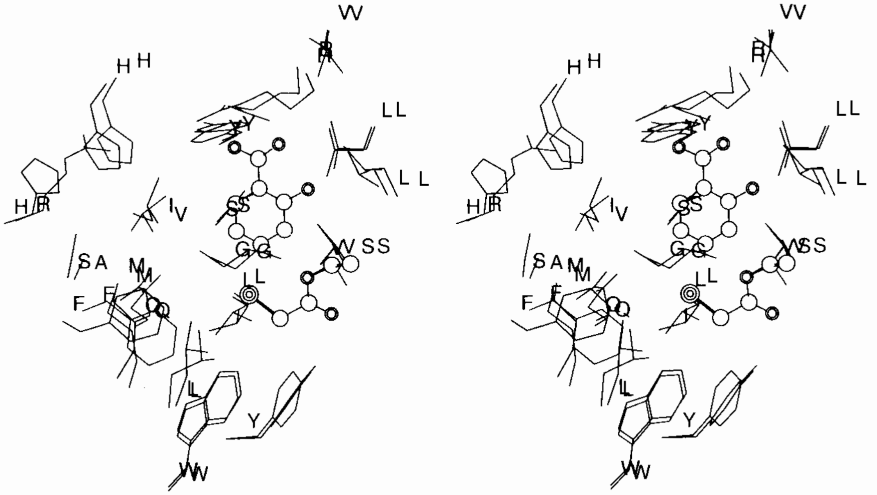


Fig. 5.23 The binding site in COX-1 for an aspirin analogue, 2-bromoacetoxybenzoic acid, in black, and the homologous residues of COX-2, in red. Can you see what unoccupied space exists in the site that could accommodate a larger ligand? Can you see any sequence differences that might be exploited to design an inhibitor that would bind to COX-2 (red) but *not* to COX-1 (black)?

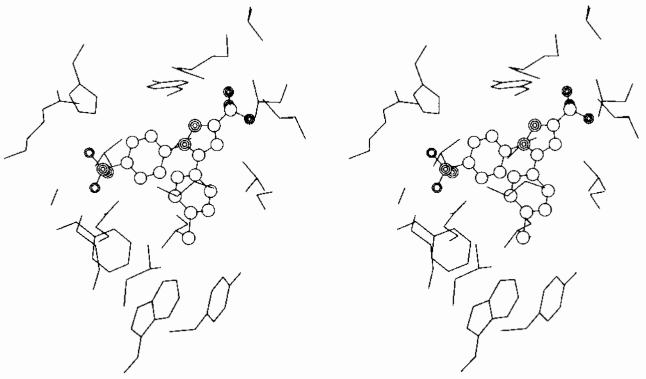


Fig. 5.24 The binding site in COX-2 (black) for a *selective* inhibitor of COX-2, SC-558 (1-phenylsulphonamide-3-trifluoromethyl-5-parabromophenylpyrazole) (red).

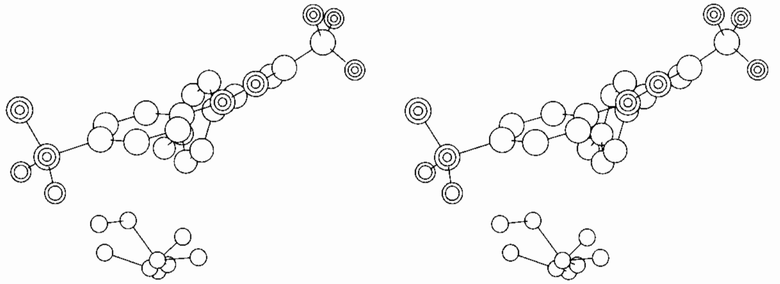


Fig. 5.25 SC-558 and the residue in COX-1 (black, isoleucine) and COX-2 (red, valine) that appears to produce the selectivity. SC-558 cannot bind to COX-1 because there would be steric contacts between it and the isoleucine.

Aspirin

Aspirin is one of the oldest of folk remedies and newest of scientific ones. At least 5000 years ago Hippocrates noted the effectiveness of preparations of willow leaves or bark to assuage pain and reduce fever. The active ingredient, salicin, was purified in 1828, and synthesized in 1859 by Kolbe. The mechanism of its action was unknown, and indeed remained unknown until, in the 1970s, J. Vane and colleagues discovered that aspirin acts by blocking prostaglandin synthesis. Not knowing the mechanism of action was never an impediment to its use.

A century ago, sodium salicylate was used in the treatment of arthritis. Because stomach irritation was a serious side effect, F. Hoffman sought to reduce the compound's acidity by forming acetylsalicylic acid, or aspirin. Aspirin was the first synthetic drug, which started the modern pharmaceutical industry. (The name salicin comes from the Latin name for willow, *salix*, and the name aspirin comes from 'a' for acetyl and 'spir' from the *spirea* plant, another natural source of salicin.)

Aspirin has the effect of reducing fever, and giving relief from aches and pains. In high doses it is effective against arthritis. Aspirin is also used for prevention and treatment of heart attacks and strokes. The applications to cardiovascular disease depend on inhibition of blood clotting by suppressing prostaglandin control over platelet clumping. The many applications of aspirin reflect the many physiological processes that involve prostaglandins.

Aspirin's many uses

Small doses	Medium doses	Large doses
Interferes with blood clotting	Fever/pain	Reduces pain and inflammation of arthritis and related diseases

Recommended reading

Protein folding

Baldwin, R. L. & Rose, G. D. (1999), Is protein folding hierarchic? I. Local structure and peptide folding. II. Folding intermediates and transition states, *Trends Biochem. Sci.*, **24**, 26–32; 77–83. [Introduction to current thinking about protein stability and folding.]

Structure alignment and sequence-structure relationships

Holm, L. & Sander, C. (1995), Dali: a network tool for protein structure comparison, *Trends Biochem Sci.*, **20**, 478–480. [Describes DALI and its applications to structural alignment.]

Smith, T. F. (1999), The art of matchmaking: sequence alignment methods and their structural implications, *Structure Fold. Des.*, **7**, R7–R12. [Description of work melding sequence and structure analysis.]

Connections among sequences, structures and functions

Das, R., Junker, J., Greenbaum, D. & Gerstein, M. B. (2001), Global perspectives on proteins: comparing genomes in terms of folds, pathways and beyond, *The Pharmacogenomics Journal*, **1**, 115–125.

Galperin, M. Y. & Koonin, E. V., Sequence–Evolution–Function / Computational Approaches in Comparative Genomics (Boston: Kluwer 2003).

State of the art in homology modelling and its application in structural genomics

Tramontano, A. (2004), Integral and differential form of the protein folding problem, *Physics of Life Revs.*, **1**, 103–127.

Guex, N., Diemand, A. & Peitsch, M. C. (1999), Protein modelling for all, *Trends. Biochem. Sci.*, **24**, 364–367.

Peitsch, M. C., Schwede, T. & Guex, N. (2000), Automated protein modelling–the proteome in 3D, *Pharmacogenomics*, **1**, 257–266. [What it will take to complete the structural genomics problem.]

Schwede, T., Kopp, J., Guex, N. & Peitsch, M. C. (2003), SWISS-MODEL: An automated protein homology-modeling server, *Nucl. Acids Res.*, **31**, 3381–3385. [Descriptions of SWISS-MODEL.]

Martí-Renom, M. A., Stuart, A. C., Fiser, A., Sánchez, R., Melo, F., Šali, A. (2000), Comparative protein structure modeling of genes and genomes, *Annu. Rev. Biophys. Biomol. Struct.*, **29**, 291–325.

Pieper, U., Eswar, N., Braberg, H., Madhusudhan, M. S., Davis, F. P., Stuart, A. C., Mirkovic, N., Rossi, A., Martí-Renom, M. A., Fiser, A., Webb, B., Greenblatt, D., Huang, C. C., Ferrin, T. E. & Šali, A. (2004), MODBASE, a database of annotated comparative protein structure models, and associated resources, *Nucl. Acids Res.*, **32**, D217–D222.

Other protein structure prediction methods

Bonneau, R. & Baker, D. (2001), Ab initio protein structure prediction: Progress and Prospects, *Annu. Rev. Biophys. Biomol. Struct.*, **30**, 173–189. [Recent review of structure prediction methods by authors of the most successful of them.]

Classics still well worth reading

Kauzmann, W. (1959), Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.*. **14**, 1–63.

Richards, F. M. (1977), Areas, volumes, packing and protein structure, *Annu. Rev. Biophys. Bioeng.*, **6**, 151–176.

Chothia, C. (1984), Principles that determine the structure of proteins, *Annu. Rev. Biochem.*, **53**, 537–572.

Richards, F. M. (1991), The protein folding problem, *Scientific Amer.*, **264(1)**, 54–57, 60–63.

Exercises, Problems, and Weblems

Exercises

5.1 The heat of sublimation of ice = 51 kJ mol^{-1} at the freezing point. In the solid state, each molecule of H_2O makes two hydrogen bonds. What is the energy of a single water-water hydrogen bond?

5.2 Which pairs are orthologues, which are paralogues, and which are neither?

(a) Human haemoglobin α and human haemoglobin β.

(b) Human haemoglobin α and horse haemoglobin α.

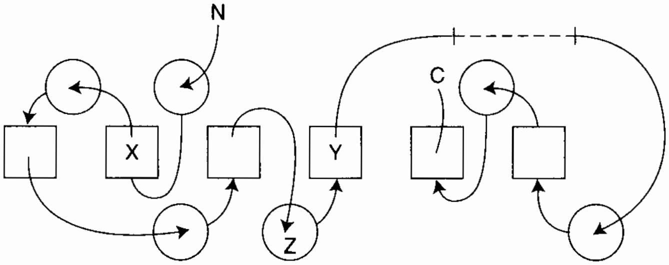
(c) Human haemoglobin α and horse haemoglobin β.

(d) Human haemoglobin α and human haemoglobin ζ

(e) The proteinases human chymotrypsin and human thrombin.

(f) The proteinases human chymotrypsin and kiwi fruit actinidin.

5.3 On a photocopy of Plate VI, indicate the locations in the structure that correspond to X, Y and Z in the following diagram.



5.4 On a photocopy of Fig. 5.8b, highlight the region of 3_{10} helix that was not predicted to be helical.

5.5 Which of the following shows the correct topology—correct strand order in the sequence and orientation—of the β-sheet in Fig. 5.8b?

(a) ↑ ↑ ↑ ↑ (b) ↑ ↓ ↑ ↓ (c) ↑ ↑ ↓ ↑
1 2 3 4 3 4 2 1 1 3 2 4

5.6 In the structure prediction of the *H. influenzae* hypothetical protein, Fig. 5.11: (a) What are the differences in folding pattern between the target protein and the experimental parent or template? (b) What are the differences in folding pattern between the prediction by A. G. Murzin and the target? (c) What are the differences in folding pattern between the prediction by A. G. Murzin and the experimental parent? In what respects is Murzin's prediction a better representation of the folding pattern than the experimental parent?

5.7 Draw the chemical structures of aspirin and 2-bromoacetoxy-benzoic acid.

5.8 Many proteins from pathogens have human homologues. Suppose you had a method for comparing the determinants of specificity in the binding sites of two homologous proteins. How could you use this method to select propitious targets for drug design?

5.9 In the neural network illustrated on page 246 (lower figure), how many parameters—variable weights and thresholds—are available to adjust, assuming a linear decision procedure?

5.10 What is the geometrical interpretation of a neuron that accepts two inputs x and y and 'fires' if and only if $x + 2y \geq 2$?

5.11 Sketch a neuron with two inputs x and y each of which may have any numerical value, that will emit 1 if and only if the value of the first input is greater than or equal to that of the second. What is the geometric interpretation of this neuron?

Problems

5.1 In the table of aligned sequences of ETS domains (see Problem 1.1): (a) Which are the most similar and which the most distant members of the family? (b) Suppose that an experimental structure is known only for the first sequence. For which others would you expect to be able to build a model with an overall r.m.s. deviation of ≤1.0 Å for 90% or more of the residues?

5.2 Sketch a neural network that accepts 8 inputs, each of which has value 0 or 1, with the interpretation that the 8 inputs correspond to the residues in a sequence of 8 amino acids, and that the value of the ith input is 0 if the ith residue is hydrophilic and 1 if the ith residue is hydrophobic. The network should output 1 if the pattern appears helical—for simplicity demand that it be PPHHPPHH where H = hydrophobic (uncharged) and P = polar or charged—and 0 otherwise.

5.3 Write a more reasonable set of patterns to identify helices from the hydrophobic/hydrophilic character of the residues in a 10-residue sequence. Your patterns might include 'wild cards'—positions that could be either hydrophobic or hydrophilic, or correlations between different positions. Generalize the previous problem by sketching neural networks to detect these more complex patterns.

5.4 We, and computers, can do logic with arithmetic. Define: 1 = TRUE and 0 = FALSE. Sketch simulated neurons with two inputs, each of which can have only the values 0 or 1, and a linear decision process for firing, for which (a) the output is the logical AND of the two inputs and (b) the output is the logical OR of the two inputs. (c) What is the simplest neural network, with each neuron having a linear decision process for firing, that produces as its output the EXCLUSIVE OR of the two inputs (the EXCLUSIVE OR is TRUE if either one of the inputs is TRUE, FALSE if neither or both of the inputs are TRUE.) Can this be done with a single layer? If not, what is the minimum number of layers in the network required?

5.5 Modify the PERL program for drawing helical wheels (pages 232–233) so that different amino acids are all represented in the same font, but appear in different colours, as follows: GAST, cyan; CVILFYPMW, green; HNQ, magenta, DE, red; and KR blue.

5.6 Hydrophobic cluster analysis. Suppose a region of a protein forms an α-helix. To represent its surface, imagine winding the sequence into an α-helix (even if in fact it forms a strand of sheet or loop in the native structure). Then 'ink' the surface of this helix, and roll it onto a sheet of paper, to print the names of the residues. By rolling the helix over twice, all faces are simultaneously visible.

From such a diagram, hydrophobic patches on surfaces of helices can be identified. In this way it is possible to try to predict which regions of the sequence actually form helices in the native structure. Comparisons of hydrophobic clusters can also be used to detect distant relationships.

Write a PERL program to produce such diagrams.

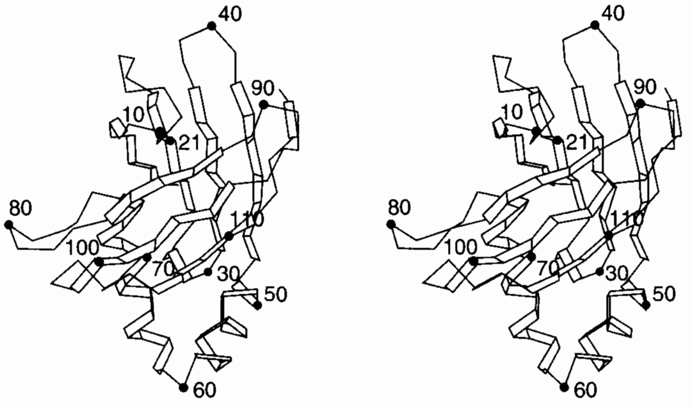
5.7 In the 2000 Critical Assessment of Structure Prediction (CASP4), one of the targets in the category for which no similar fold was known was the N-terminal domain of the human DNA end-joining protein XRCC4, residues 1–116.

The secondary structure prediction by B. Rost, using the method PROF: profile-based neural network prediction, is as follows (An H under a residue means that that residue is predicted to be in a Helix, an E means that that residue is predicted to be in an Extended conformation, or strand, and – means Other):

```
                     1         2         3         4         5         6
                     0         0         0         0         0         0
Sequence    MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Prediction  ---EEEEEEE-----HHHHHH-HHHHHHH--EEEEEE------EE---HHHHHHHHHHHH

                                                   1         1
                     7         8         9         0         1
                     0         0         0         0         0
Sequence    MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Prediction  HHH-HHHHHHHHHHHH-----EEEEEE-----EEEEE------EEEE-----HHHH
```

The experimental structure of this domain, released after the predictions were submitted (PDB entry [1FU1]), is shown here:



HUMAN XRCC4 [1fu1] domain1 HUMAN XRCC4 [1fu1] domain1

The secondary structure assignments from the PDB entry are:

Secondary Structure	Residue ranges
Helix	27–29, 49–59, 62–75
Sheet 1	2–8, 18–24, 31–37, 42–48, 114–115
Sheet 2	84–88, 95–101, 104–111

(a) Calculate the value of Q3, the percentage of residues correctly assigned to helix (H), strand (E) and other (-).

(b) On a photocopy of the picture of XRCC4, highlight, in separate colours, the regions *predicted* to be in helices and strands.

(c) From the result of (b): How many predicted helices overlap with helices in the experimental structure? How many strands overlap with strands in the experimental structure?

5.8 In CASP4, the group of Bonneau, Tsai, Ruczinski and Baker made a prediction of the full three-dimensional structure of protein XRCC4, residues 1-116. The secondary structure prediction derived from their model is as follows (H = helix, E = strand (extended), - = other):

```
                    1         2         3         4         5         6
                    0         0         0         0         0         0
Sequence   MERKISRIHLVSEPSITHFLQVSWEKTLESGFVITLTDGHSAWTGTVSESEISQEADDMA
Prediction ----E--EEEE---EEEE--EHHHHHHHH----EEEE--EEEE-----HHHHHHHHHHHH

                    7         8         9         0         1
                    0         0         0         0         0
Sequence   MEKGKYVGELRKALLSGAGPADVYTFNFSKESCYFFFEKNLKDVSFRLGSFNLEKV
Prediction HHH---HHHHHHHHHHH-----EEEEEEE--EEEEEEE------HHHH----HHHH
```

(a) What is the value of Q3 for this prediction? (b) In this case, which method gives the better results, as measured by Q3, for the prediction of secondary structure: the neural network that produces only a secondary structure prediction, or a prediction of the full three-dimensional structure?

5.9 Write and test PERL programs that implements the neural networks shown on page 247.

5.10 Suppose that you are trying to evaluate, using a threading approach, whether a sequence of length M is likely to have the folding pattern of a protein of known structure of length $N > M$. (a) How many different alignments of the sequences are possible. (b) Suppose that half the residues of the known protein form α-helices, and no gaps within helical regions are permitted. How many different alignments of the sequences are now possible? (c) How many alignments are there, under each of these assumptions, if $N = 200$, and $M = 150$?

5.11 Write a PERL program to calculate approximate values of π by a Monte Carlo method, as follows: The square in the plane with corners at (0, 0), (1,0), (0, 1), and (1,1) has area 1. Compute a series of *pairs* of random numbers (x, y) in the range [0, 1] to generate points distributed at random in this square. Count the number of points that lie within a circle of radius 0.5 inscribed in the square. The ratio of the number of points that fall within the circle to the total number of points = the ratio of the area of the circle to the area of the square = $\pi/4$.

Determine the average relationship between the number of points chosen and the number of correct digits in the calculated value of π. Estimate the number of points required to determine π correctly to 50 decimal places.

5.12 To convert the output of a neuron from a step function to a smooth function (see page 248), one can replace a statement of the form 'Let X be some weighted sum of the inputs; then output 1 if $X > 0$ else output 0' with 'Let X be some weighted sum of the inputs; then output $1/(1 + e^{-X})$.' (a) Verify that as $X \to -\infty$, $1/(1 + e^{-X}) \to 0$, as $X \to +\infty$, $1/(1 + e^{-X}) \to 1$, and that if $X = 0$, $1/(1 + e^{-X}) = 0.5$. (b) Suppose the network for determining whether a point lies within a triangle (page 247, bottom) is so altered, so that the output of each neuron is described by the smooth function $1/(1 + e^{-X})$ rather than a step function, and that a point is considered inside the accepted area if the output of the network is >0.5. Write a PERL program to determine what area is then defined.

Weblems

5.1 The bacterium *Pseudomonas fluorescens* and the fungus *Curvularia inaequalis* each possess a chloroperoxidase, an enzyme that catalyses halogenation reactions. Do these enzymes have the same folding pattern?

5.2 Check the prediction of transmembrane helical segments in bacteriorhodopsin from the secondary structure assignments in the experimental structure in the Protein Data Bank.

5.3 Plate VI showed the structure of a thiamin-binding domain, identified by M. Gerstein as one of the five most common folding patterns appearing in archaea, bacteria and eukarya. Using facilities available in SCOP, draw pictures of the four other structures.

5.4 Using either the results of Weblem 5.3, or pictures available in *Introduction to Protein Architecture: The Structural Biology of Proteins*, draw simplified topology diagrams analogous to the one in Plate VI for the other four structures.

5.5 Does the human θ_1 globin gene encode an active globin? Or is it really a pseudogene? Send the amino acid sequence of human θ_1 globin to SWISS-MODEL, including a request for a WhatCheck report on the result. What can you conclude from the result about the status of human θ_1 globin?

5.6 Compare the number of entries in SCOP, in different categories, listed on page 240, with the number that SCOP contains now.

5.7 Align the sequences of γ-chymotrypsin and *S. aureus* epidermolytic toxin A using pairwise sequence alignment methods. Compare the result with the structural alignment shown in the text.

5.8 Align the sequences of human neutrophil elastase and *C. elegans* elastase. (a) In the optimal alignment, how many identical resdues are there? (b) Would it be reasonable to build a model of *C. elegans* elastase starting from the structure of human neutrophil elastase?

5.9 S. Chakravarty and K. K. Kannan solved the structures of carbonic anhydrase with a benzenesulphonamide ligand (Protein Data Bank entry 1CZM.) Draw pictures of the binding site showing the nature of the interactions between the protein and ligands. Describe the nature of the interactions in terms of the conclusions drawn from the QSAR analysis.