

# Protein Classification and Structure Prediction

## INTRODUCTION, 382

### Protein structure prediction, 382

### Review of protein structure and terminology, 386

$\alpha$  Helix, 387

$\beta$  Sheet, 388

Loop, 389

Coil, 389

### Protein classification, 389

Terms used for classifying protein structures and sequences, 390

Classes of protein structure, 393

Protein databases, 394

## METHODS, 398

### Viewing protein structures, 400

### Protein structure classification databases, 402

### Alignment of protein structures, 403

Dynamic programming, 419

Distance matrix, 421

Fast structural similarity search based on secondary structure analysis, 423

Significance of alignments of secondary structure, 426

Displaying protein structural alignments, 427

### Structural prediction, 427

Use of sequence patterns for protein structure prediction, 427

Prediction of protein secondary structure from the amino acid sequence, 440

Prediction of three-dimensional protein structure, 460

### Evaluating the success of structure predictions, 471

### Structural modeling, 472

### Summary and future prospects, 473

## REFERENCES, 473

## INTRODUCTION

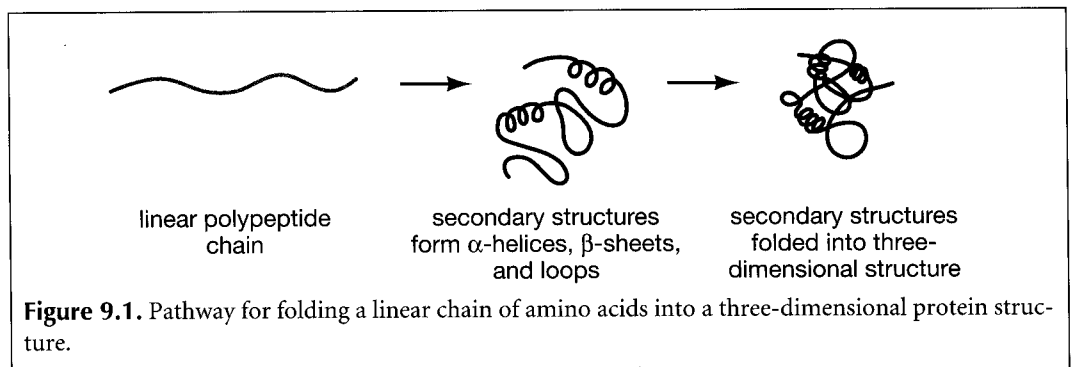
ONE OF THE MAJOR GOALS OF BIOINFORMATICS is to understand the relationship between amino acid sequence and three-dimensional structure in proteins. If this relationship were known, then the structure of a protein could be reliably predicted from the amino acid sequence. Unfortunately, the relationship between sequence and structure is not that simple. Much progress has been made in categorizing proteins on the basis of structure or sequence, and this type of information is very useful for protein modeling. A review of protein synthesis and structure is therefore in order.

### PROTEIN STRUCTURE PREDICTION

The polypeptide chain is first assembled on the ribosome using the codon sequence on mRNA as a template, as illustrated in Figure 9.1. The resulting linear chain forms secondary structures through the formation of hydrogen bonds between amino acids in the chain. Through further interactions among amino acid side groups, these secondary structures then fold into a three-dimensional structure. Chaperone proteins and membranes may assist with this process. For the protein to have biological activity, processing of the protein by cleavage or chemical modification may also be necessary. Therefore, protein structure is largely specified by amino acid sequence, but how one set of interactions of the many possible occurs is not yet fully understood (Branden and Tooze 1991).

Some protein sequences have distinct amino acid motifs that always form a characteristic structure. Prediction of these structures from sequence is quite achievable using presently available methods and information. For most proteins, however, the accuracy of secondary structure prediction is approximately 70–75%. Methods for matching sequence to three-dimensional structure have been formulated, but they are not yet very reliable. However, great forward strides have been made, and there is a very active community of structural biochemists and bioinformaticists working on improvements. The need for such an effort is revealed by the rapid increases in the number of protein sequences and structures.

As of June 2000, more than 12,500 protein structures had been deposited in the Brookhaven Protein Data Bank (PDB), and 86,500 protein sequence entries were in the SwissProt protein sequence database, a ratio of approximately 1 structure to 7 sequences. The number of protein sequences can be expected to increase dramatically as more sequences are produced by research laboratories and the genome sequencing projects. As more and more sequences and structures have been found, there have been some quite



remarkable revelations that make the goals of reliable structure prediction more within reach.

It has first been estimated that there are approximately 1,000 protein families composed of members that share detectable sequence similarity (Dayhoff et al. 1978; Chothia 1992). Thus, as new protein sequences are obtained, they will be found to be similar to other sequences already in the databases and can be expected to share structural features with these proteins. Whether this low number represents physical restraints in folding the polypeptide chain into a three-dimensional structure or merely the selection of certain classes of three-dimensional structure by evolution has yet to be discovered (Gibrat et al. 1996). The sequence alignment, motif-finding, block-finding, and database similarity search methods described in Chapters 3, 4, and 7 may be used to discover these familial relationships. Understanding these relationships is fundamentally important because this information can greatly assist with structural predictions. As discussed below, information from amino acid substitutions at a particular sequence position as obtained from a multiple sequence alignment has been found to increase significantly the prediction of secondary structures from protein sequences. A second major advance in protein structure analysis has been the revelation that proteins adopt a limited number of three-dimensional configurations.

Protein structures include a core region comprising secondary structural elements packed in close proximity in a hydrophobic environment. Specific interactions between the amino acid side chains occur within this core structure. At a given amino acid position in a given core, the amino acids that can substitute are limited by space and available contacts with other nearby amino acids. Outside of the core are loops and structural elements in contact with water, other proteins, and other structures. Amino acid substitutions in these regions are not as restricted as in the core. Through a close comparison of a newly generated three-dimensional structure with previously found structures, the new structure has often been found to fold into  $\alpha$ -helical and  $\beta$ -sheet structural elements in the same order and spatial configuration as one or more structures already in the structural database. Proteins that show such structural similarities often do not share any detectable sequence similarity in these same regions. Hence, entirely different sequences can fold into similar three-dimensional configurations. Databases of these common structural features have been prepared and are available on Web sites described later in this chapter.

The finding that only certain amino acids can be substituted at each position in a particular protein core underscores two difficulties in using sequence alignments to make structural predictions. First, because a different set of substitutions apply to each position in each protein core, standard amino acid substitution matrices such as the Dayhoff PAM matrices and the BLOSUM matrices, described in Chapter 3, may not provide an alignment that has structural significance. The substitutions used to produce these tables are averaged over many sequence alignments, representing observed substitutions in both core regions and loops of sequence families.

Scoring matrices that represent a conserved region in the multiple sequence alignment of a set of similar proteins may also be produced, as described in Chapters 4 and 7. These matrices store information on the amino acid variation found in each column of the multiple sequence alignment. They are powerful tools for searching a new protein sequence for the presence of a sequence pattern that is similar to those in the original set of proteins. These scoring matrices include the profile, which represents gapped alignments, and the position-specific scoring matrix (PSSM), which represents ungapped alignments. A conserved region with gaps in a multiple sequence alignment may also be represented by a profile hidden Markov model (profile HMM), which provides a probability-based model of the multiple sequence alignment. Like the scoring matrices, the profile HMM representa-

tion of a sequence alignment can be used to identify related sequences. These methods are discussed in detail in Chapters 4 and 7.

Scoring matrices and profile HMMs can provide a direct link between sequence and structure. If one of the sequences that is represented by the matrix or profile model has a known three-dimensional structure, then any other sequences that match the model are also predicted to have the same structure. Conversely, if the model can be shown to match a protein of known structure, a sequence–structure link may be made. A related method is to produce a HMM (also called a discrete state-space model in the protein structure literature) for a set of proteins that belong to a structural family. These models include information on amino acid preference for positions in secondary structures. A query sequence can then be searched by a set of such models to determine whether the sequence has sequence patterns that represent the structure. A range of Web sites provide a variety of these types of analyses (see Fig. 9.30).

A second difficulty in making sequence alignments reflect structural similarity is that gaps in the alignment should be confined to regions not in the core. Alignments that reflect structures in core regions should have few if any gaps. Some multiple sequence alignment programs such as CLUSTALW (see Chapter 4, p. 153) and the Bayes block aligner (Chapter 3, p. 124) do provide for such variation in gap placement. These programs place alignment gaps where the alignment scores are low and, from a structural viewpoint, represent variable loops. The profile models described above also accommodate such variations of placement.

In addition to sequence-by-sequence alignment and sequence-by-structure alignment, it is also possible to perform a structure-by-structure alignment. In this type of alignment, sequential positions of the backbone carbon atoms for each amino acid in the two sequences are compared to determine whether the chain of atoms is tracing the same path in space. If two or more similar paths are found in the same relative positions and orientations, the structures corresponding to those paths are similar. From these methods, discussed below, databases of structural elements have been made and are available to the laboratory.

What is a reasonable goal for protein structure prediction from the perspective of a molecular biologist? The most satisfying result is to find sequence and structural alignments of a newly identified protein with a protein of known three-dimensional structure. Even if such a prediction can be made, the positions of individual amino acids will probably not be accurately known. If the sequence identity is 50% or better, one sequence can be superimposed on the structure of the other sequence and the predicted structure will be quite accurate. If the sequence identity is greater than 30%, it may be possible to identify common structural features, but it will become more difficult to identify the precise positions of the amino acids in the structure as sequence identity decreases.

The prediction of protein structure is an active and promising area of research. As more three-dimensional structures are found and the computational tools for predicting structure are improved, structural predictions will undoubtedly improve. The existence of new groups of proteins for structural analysis is suggested by the existence of genome “ORFans” that may represent new sets of families (superfamilies) with a unique structure and function (Fischer and Eisenberg 1999). One group of investigators that works on protein classification has developed a protein structure initiative to identify new protein targets for structural analysis (<http://www.structuralgenomics.org/main.html>). A method for estimating the probability for a protein to have a new fold has been described previously (Portugaly and Linial 2000). The Human Proteome/Structural Genomics Pilot Project (<http://proteome.bnl.gov>), a consortium of Brookhaven National Laboratory, the Rockefeller University, and Albert Einstein College of Medicine, is examining the feasibility of

*A special case is the use of the term “structural homology,” meaning as it did with sequence homology that the sequences were derived from a common ancestor, as evidenced by their having significant sequence similarity. As described above, two proteins may have significant structural similarity but no detectable sequence similarity. Therefore, it may be incorrect to refer to these proteins as homologous in the absence of evidence that they are derived from a common ancestor.*

high-throughput determination of three-dimensional structures of proteins starting with genomic sequences.

With a larger set of protein models, the usefulness of structure prediction is increased even further (Pennisi 1998). Many additional methods for structural classification of proteins and for displaying the structures have meanwhile been devised, and the Web has provided these resources to the research community. Formerly, special software and hardware were required to view structures. Now, there are a variety of visualization tools that work with a Web browser and allow one to view a molecule in three dimensions, to compare structures, and to perform other useful procedures. A representation of several useful Web sites for protein structure analysis is given in Table 9.1.

In this chapter, basic features of protein structure and structural terminology and the terms describing them are first reviewed. Some terms refer to sequence similarity, some to structural similarity, and some to both sequence and structure, and it is important not to confuse them.

**Table 9.1.** *Main Web sites for protein structural analysis*

Name of resource	Resources available	Internet address
Protein data bank (PDB) at the State University of New Jersey (Rutgers) <sup>a</sup>	atomic coordinates of structures as PDB files, models, viewers, links to many other Web sites for structural analysis and classification	<a href="http://www.rcsb.org/pdb">http://www.rcsb.org/pdb</a> ; also at mirror Web sites (Berman et al. 2000)
National Center for Biotechnology Information Structure Group	Molecular Modelling Database (MMDB), Vector Alignment Search Tool (VAST) for structural comparisons, viewers, threader software	<a href="http://www.ncbi.nlm.nih.gov/Structure/">http://www.ncbi.nlm.nih.gov/Structure/</a>
Structural Classification of Proteins at Cambridge University	SCOP database of structural relationships among known protein structures classified by superfamily, family, and fold	<a href="http://scop.mrc-lmb.cam.ac.uk/scop">http://scop.mrc-lmb.cam.ac.uk/scop</a> ; also at Web mirror sites
Biomolecular Structure and Modelling group at the University College, London	CATH database, a hierarchical domain classification of protein structures by class, architecture, fold family and superfamily, other databases and structural analyses, threader software	<a href="http://www.biochem.ucl.ac.uk/bsm">http://www.biochem.ucl.ac.uk/bsm</a> ; also at Web mirror sites
European Bioinformatics Institute, Hinxton, Cambridge	databases, TOPS protein structural topology cartoons, Dali domain server, and FSSP database <sup>b</sup>	<a href="http://www2.ebi.ac.uk/">http://www2.ebi.ac.uk/</a>
The PredictProtein server at the European Molecular Biology Laboratory at Heidelberg, Germany	important site for secondary structure prediction by PHD, predator, TOPITS, threader	<a href="http://cubic.bioc.columbia.edu/predictprotein">http://cubic.bioc.columbia.edu/predictprotein</a> ; also at Web mirror sites <sup>c</sup>
Swiss Institute of Bioinformatics, Geneva	basic types of protein analysis <sup>d</sup> databases, the Swiss-Model resource for prediction of protein models, Swiss-PdbViewer	<a href="http://www.expasy.ch/">http://www.expasy.ch/</a>

Additional sites are listed in the text. In addition to these sites, there are a number of Web sites and courses that discuss protein structure. The Swiss Institute for Bioinformatics (ISREC server) provides a tutorial on Principles of Protein Structure, Comparative Protein Modelling, and Visualisation at <http://www.expasy.ch/swissmod/course/course-index.htm>. There is also a Web course in protein structure at Birkbeck College <http://www.cryst.bbk.ac.uk/teaching/>.

<sup>a</sup> A summary of the PDB entries is provided at <http://www.biochem.ucl.ac.uk/bsm/pdbsum/> (Laskowski et al. 1997).

<sup>b</sup> 3Dee database of protein domains at <http://barton.ebi.ac.uk/servers/3Dee.html>. Dali domain server is at <http://www2.embl-ebi.ac.uk/dali/domain/> and FSSP database at <http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html>.

<sup>c</sup> Also performed at the structure prediction server at <http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html>.

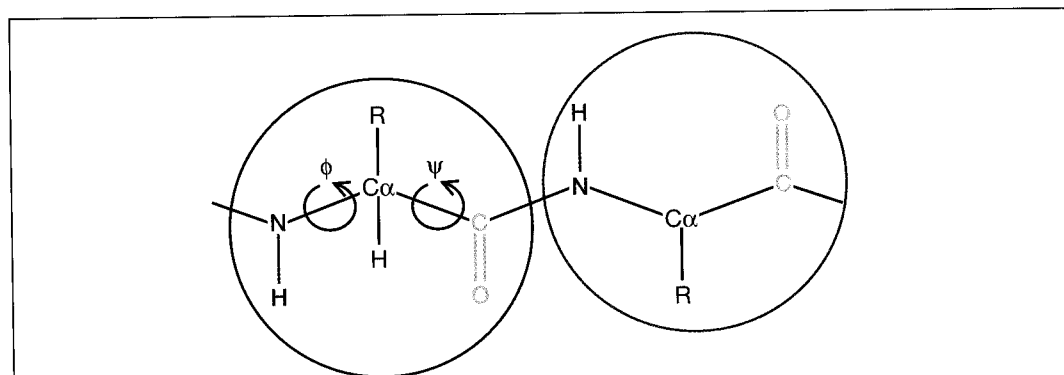
<sup>d</sup> This site offers a series of basic types of protein analysis to assist with protein identification, including identification by amino acid composition, charge, size, and sequence fingerprint. Predictions of posttranslational modifications and oligosaccharide structures are also available.

Other terms that are used to describe protein structure and the methods for displaying and comparing protein structures are described. Some of the more easily found structures and then the methods used to predict secondary and three-dimensional structures are discussed. A flowchart showing the steps to be followed to analyze a new protein sequence is included at the beginning of the Methods section (p. 399). The chapter concludes with a discussion of methods used to evaluate the success of these predictions.

## REVIEW OF PROTEIN STRUCTURE AND TERMINOLOGY

Proteins are chains of amino acids joined by peptide bonds, as illustrated in Figure 9.2. Many conformations of the chain are possible due to the rotation of the chain about each  $C_{\alpha}$  atom. It is these conformational variations that are responsible for differences in the three-dimensional structures of proteins. Each amino acid in the chain is polar, i.e., it has separated positive and negatively charged regions with a chemically free  $C=O$  group, which can act as a hydrogen bond acceptor, and an  $NH$  group, which can act as a hydrogen bond donor. These groups interact in protein structures. The 20 amino acids found in proteins can be grouped according to the chemistry of their  $R$  groups, as depicted in Table 9.2. The  $R$  side chains also play an important structural role. Special roles are played by glycine, which does not have a side chain and can therefore increase local flexibility in structures, and cysteine, which can react with another cysteine to form a cross-link that can stabilize the protein structure.

Much of the protein core comprises regular secondary structures,  $\alpha$  helices and  $\beta$  sheets, folded into a three-dimensional configuration. In these secondary structures, regular patterns of  $H$  bonds are formed between neighboring amino acids, and the amino acids



**Figure 9.2.** The structure of two amino acids in a polypeptide chain. Each amino acid is encircled by a different color ring. The  $R$  group is different for each of the 20 amino acids. Neighboring amino acids are joined by a peptide bond between the  $C=O$  and  $NH$  groups. The  $N-C_{\alpha}-C$  sequence is repeated throughout the protein, forming the backbone of the three-dimensional structure. The amino acid at one end of the chain has a free  $NH_2$  group (chain beginning) and the amino acid at the other end has a free  $COOH$  group (chain end). The bonds on each side of the  $C_{\alpha}$  atom are quite free to rotate, but many combinations of angles are not possible for most amino acids due to spatial constraints from the  $R$  group and neighboring positions in the chain. The conformation of the protein backbone in space is determined by the angles of these bonds,  $\Phi$  of the bond between the  $N$  and  $C_{\alpha}$  atoms and  $\Psi$  of the bond between the  $C_{\alpha}$  and  $C$  of the  $C=O$  group, also named  $C_{\beta}$ . The distribution of these two angles for the amino acids in a particular protein is often plotted on a graph called a Ramachandran plot. The angle  $\Omega$  of the peptide bond joining the  $C=O$  and  $NH$  groups (not shown) is nearly always  $180^\circ$ .

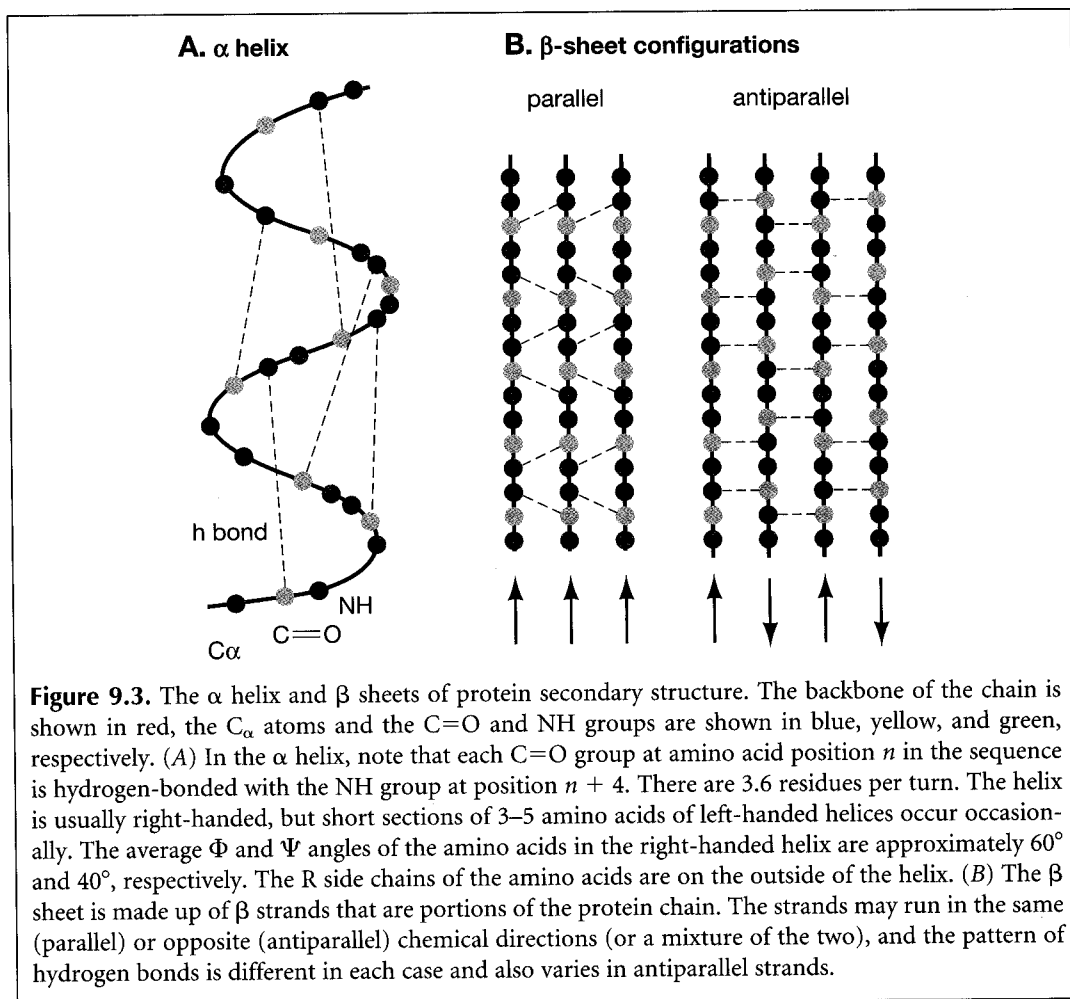
**Table 9.2.** *Chemical properties of the 20 amino acids*

Chemical group	Amino acid (one-letter code)	Name
Hydrophobic	A	alanine
	V	valine
	Y	phenylalanine
	P	proline
	M	methionine
	I	isoleucine
Charged	L	leucine
	D	aspartic acid
	E	glutamic acid
	K	lysine
	R	arginine
Polar	S	serine
	T	threonine
	Y	tyrosine
	H	histidine
	C	cysteine
	N	asparagine
	Q	glutamine
Glycine	W	tryptophan
	G	glycine
Cross-linking	—	cysteine + cysteine

have similar  $\Phi$  and  $\Psi$  angles, as depicted in Figure 9.3. The formation of these structures neutralizes the polar groups on each amino acid. The secondary structures are tightly packed in the protein core in a hydrophobic environment. Each amino acid side group has a limited volume to occupy and a limited number of possible interactions with other nearby side chains, a situation that must be taken into account in molecular modeling and alignments.

## $\alpha$ Helix

The  $\alpha$  helix depicted in Figure 9.3A is the most abundant type of secondary structure in proteins. The helix has 3.6 amino acids per turn with an H bond formed between every fourth residue; the average length is 10 amino acids (3 turns) or 10 Å but varies from 5 to 40 (1.5 to 11 turns). The alignment of the H bonds creates a dipole moment for the helix with a resulting partial positive charge at the amino end of the helix. Because this region has free  $\text{NH}_2$  groups, it will interact with negatively charged groups such as phosphates. The commonest location of  $\alpha$  helices is at the surface of protein cores, where they provide an interface with the aqueous environment. The inner-facing side of the helix tends to have hydrophobic amino acids and the outer-facing side hydrophilic amino acids. Thus, every third of four amino acids along the chain will tend to be hydrophobic, a pattern that can be quite readily detected. In the leucine zipper motif, a repeating pattern of leucines on the facing sides of two adjacent helices is highly predictive of the motif. A helical-wheel plot can be used to show this repeated pattern (see below). Other  $\alpha$  helices buried in the protein core or in cellular membranes have a higher and more regular distribution of



hydrophobic amino acids, and are highly predictive of such structures. Helices exposed on the surface have a lower proportion of hydrophobic amino acids. Amino acid content can be predictive of an  $\alpha$ -helical region. Regions richer in alanine (A), glutamic acid (E), leucine (L), and methionine (M) and poorer in proline (P), glycine (G), tyrosine (Y), and serine (S) tend to form an  $\alpha$  helix. Proline destabilizes or breaks an  $\alpha$  helix but can be present in longer helices, forming a bend. There are computer programs for predicting quite reliably the general location of  $\alpha$  helices in a new protein sequence.

## $\beta$ Sheet

$\beta$  Sheets are formed by H bonds between an average of 5–10 consecutive amino acids in one portion of the chain with another 5–10 farther down the chain, as shown in Figure 9.3B. The interacting regions may be adjacent, with a short loop in between, or far apart, with other structures in between. Every chain may run in the same direction to form a parallel sheet, every other chain may run in the reverse chemical direction to form an antiparallel sheet, or the chains may be parallel and antiparallel to form a mixed sheet. As illustrated in Figure 9.3, the pattern of H bonding is different in the parallel and antiparallel configurations. Each amino acid in the interior strands of the sheet forms two H bonds with neighboring amino acids, whereas each amino acid on the outside strands forms only



one bond with an interior strand. Looking across the sheet at right angles to the strands, more distant strands are rotated slightly counterclockwise to form a left-handed twist, which is apparent in some of the structures shown below. The  $C_{\alpha}$  atoms alternate above and below the sheet in a pleated structure, and the R side groups of the amino acids alternate above and below the pleats. The  $\Phi$  and  $\Psi$  angles of the amino acids in  $\beta$  sheets vary considerably in one region of the Ramachandran plot (see Fig. 9.2 legend). It is more difficult to predict the location of  $\beta$  sheets than of  $\alpha$  helices. The situation improves somewhat when the amino acid variation in multiple sequence alignments is taken into account.

## Loop

Loops are regions of a protein chain that are (1) between  $\alpha$  helices and  $\beta$  sheets, (2) of various lengths and three-dimensional configurations, and (3) on the surface of the structure. Hairpin loops that represent a complete turn in the polypeptide chain joining two antiparallel  $\beta$  strands may be as short as two amino acids in length. Loops interact with the surrounding aqueous environment and other proteins. Because amino acids in loops are not constrained by space and environment as are amino acids in the core region, and do not have an effect on the arrangement of secondary structures in the core, more substitutions, insertions, and deletions may occur. Thus, in a sequence alignment, the presence of these features may be an indication of a loop. The positions of introns in genomic DNA sometimes correspond to the locations of loops in the encoded protein. Loops also tend to have charged and polar amino acids and are frequently a component of active sites. A detailed examination of loop structures has shown that they fall into distinct families.

## Coil

A region of secondary structure that is not a helix, a sheet, or a recognizable turn is commonly referred to as a coil.

## PROTEIN CLASSIFICATION

---

Proteins may be classified according to both structural and sequence similarity. For structural classification, the sizes and spatial arrangements of secondary structures described in the above section are compared in known three-dimensional structures. For classification by sequence similarity, alignments of protein sequences are made using the methods described in Chapters 3 and 4. Classification based on sequence similarity was historically the first to be used. Initially, similarity based on alignments of whole sequences was performed. Later, proteins were classified on the basis of the occurrence of conserved amino acid patterns. Databases that classify proteins by one or more of these schemes are available.

In considering protein classification schemes, it is important to keep several observations in mind. First, two entirely different protein sequences from different evolutionary origins may fold into a similar structure. Conversely, the sequence of an ancient gene for a given structure may have diverged considerably in different species while at the same time maintaining the same basic structural features. Recognizing any remaining sequence similarity in such cases may be a very difficult task. Second, two proteins that share a significant degree of sequence similarity either with each other or with a third sequence also share an evolutionary origin and should share some structural features also. However, gene duplication and genetic rearrangements during evolution may give rise to new gene copies,

which can then evolve into proteins with new function and structure. Examples of these events are discussed at the beginning of Chapter 2 and in Chapters 6 and 10. To make assessments of protein structure, a number of terms that describe protein similarity and structural relationships are used.

## Terms Used for Classifying Protein Structures and Sequences

The more commonly used terms for describing evolutionary and structural relationships among proteins are listed below. Many additional terms are used to describe various kinds of structural features found in proteins. Descriptions of such terms may be found at the CATH Web site (<http://www.biochem.ucl.ac.uk/bsm/cath/lex/glossary.html>), the Structural Classification of Proteins (SCOP) Web site (<http://pdb.wehi.edu.au/scop/gloss.html> and Web mirror sites), and a Glaxo-Wellcome tutorial on the Swiss bioinformatics Expasy Web site (<http://www.expasy.ch/swissmod/course/course-index.htm>).

**Active site** is a localized combination of amino acid side groups within the tertiary (three-dimensional) or quaternary (protein subunit) structure that can interact with a chemically specific substrate and that provides the protein with biological activity. Proteins of very different amino acid sequences may fold into a structure that produces the same active site.

**Architecture** describes the relative orientations of secondary structures in a three-dimensional structure without regard to whether or not they share a similar loop structure. In contrast, a fold is a type of architecture that also has a conserved loop structure. Architecture is a classification term used by the CATH database (<http://www.biochem.ucl.ac.uk/bsm/cath/>).

**Blocks** is a term used to describe a conserved amino acid sequence pattern in a family of proteins. The pattern includes a series of possible matches at each position in the represented sequences, but there are not any inserted or deleted positions in the pattern or in the sequences. By way of contrast, sequence profiles are a type of scoring matrix that represents a similar set of patterns that includes insertions and deletions. Profile HMMs are hidden Markov models of such gapped patterns (see Chapters 4 and 7). There are 2,290 HMM profile models in Pfam release 5.4 described below.

**Class** is a term used to classify protein domains according to their secondary structural content and organization. Four classes were originally recognized by Levitt and Chothia (1976), and several others have been added in the SCOP database described below. Three classes are given in the CATH database: mainly- $\alpha$ , mainly- $\beta$ , and  $\alpha$ - $\beta$ , with the  $\alpha$ - $\beta$  class including both alternating  $\alpha/\beta$  and  $\alpha + \beta$  structures. Thus, class 4 of the SCOP database is included in class 3 of the CATH database.

**Core** is the portion of a folded protein molecule that comprises the hydrophobic interior of  $\alpha$  helices and  $\beta$  sheets. The compact structure brings together side groups of amino acids into close enough proximity so that they can interact. When comparing protein structures, as in the SCOP database, core refers to the region common to most of the structures that share a common fold or that are in the same superfamily. In structure prediction, core is sometimes defined as the arrangement of secondary structures that is likely to be conserved during evolutionary change (Madej et al. 1995). A library of protein cores designated LPFC is maintained at Stanford University at <http://www-camis.stanford.edu/projects/helix/LPFC/> and is based on multiple sequence alignments using amino acid scoring matrices based on structural substitutions.

**Domain** (sequence context). See Homologous domain.

**Domain** (structural context; also see Homologous domain entry) refers to a segment of a polypeptide chain that can fold into a three-dimensional structure irrespective of the presence of other segments of the chain. The separate domains of a given protein may interact extensively or may be joined only by a length of polypeptide chain. A protein with several domains may use these domains for functional interactions with different molecules. 3Dee, a database of protein domain definitions, is provided at <http://barton.ebi.ac.uk/servers/3Dee.html/>. A structural classification of protein domains is maintained at <http://www2.embl-ebi.ac.uk/dali/domain/> (Holm and Sander 1998), and ddbase, a database of protein domains, may be found at <http://www-cryst.bioc.cam.ac.uk/~ddbbase>. Another domain database may be found at <http://www3.icgeb.trieste.it/> (Pongor et al. 1993).

**Family** (sequence context), as defined originally by Dayhoff et al. (1978), is a group of proteins of similar biochemical function that are more than 50% identical when aligned. This same cutoff is still used by the Protein Information Resource (PIR). A protein family comprises proteins with the same function in different organisms (orthologous sequences) but may also include proteins in the same organism (paralogous sequences) derived from gene duplication and rearrangements (Henikoff et al. 1997). If a multiple sequence alignment of a protein family reveals a common level of similarity throughout the lengths of the proteins, PIR refers to the family as a homeomorphic family. The aligned region is referred to as a homeomorphic domain, and this region may comprise several smaller homology domains that are shared with other families. Families may be further subdivided into subfamilies or grouped into superfamilies based on respective higher or lower levels of sequence similarity (Barker et al. 1995; <http://www-nbrf.georgetown.edu/>). The SCOP database described below (release 1.50) reports 1296 families and the CATH database (version 1.7 beta), also described below, reports 1846 families.

When the sequences of proteins with the same function are examined in greater detail, some are found to share high sequence similarity. They are obviously members of the same family by the above criteria. However, others are found that have very little, or even insignificant, sequence similarity with other family members. In such cases, the family relationship between two distant family members A and C can often be demonstrated by finding an additional family member B that shares significant similarity with both A and C (Pearson 1996; Park et al. 1997). Thus, B provides a connecting link between A and C. Another approach is to examine distant alignments for highly conserved matches (Patthy 1987, 1996).

At a level of identity of >50%, proteins are likely to have the same three-dimensional structure, and the identical atoms in the sequence alignment will also superimpose within approximately 1 Å in the structural model (Holm and Sander 1994). Thus, if the structure of one member of a family is known, a reliable prediction may be made for a second member of the family, and the higher the identity level, the more reliable the prediction. Protein structural modeling can be performed by examining how well the amino acid substitutions fit into the core of the three-dimensional structure.

**Family** (structural context), as used in the FSSP database (Holm and Sander 1998) and the DALI/FSSP Web site (see below), refers to two structures that have a significant level of structural similarity but not necessarily significant sequence similarity.

**Fold** is a term with similar meaning to structural motif, but in general refers to a somewhat larger combination of secondary structural units in the same configuration. Thus, proteins sharing the same fold have the same combination of secondary structures that are connected by similar loops. An example is the Rossman fold comprising several

alternating  $\alpha$  helices and parallel  $\beta$  strands. In the SCOP, CATH, and FSSP databases described below, the known protein structures have been classified into hierarchical levels of structural complexity with the fold as a basic level of classification. From a survey of the currently known protein structures in the Brookhaven Protein Data Bank (Holm and Sander 1998), approximately 500 independent folds have been identified. The number of distinct folds in the SCOP database is 548 (release 1.50) and the number of the equivalent topological families in the CATH database is 580 (version 1.70 beta release). These databases are described below. **Foldon** is a related term that has been used to describe an independently folding unit (Panchenko et al. 1996, 1997).

**Homologous domain** (sequence context, also see Domain, structural context) refers to an extended sequence pattern, generally found by sequence alignment methods, that indicates a common evolutionary origin among the aligned sequences. A homology domain is generally longer than motifs. The domain may include all of a given protein sequence or only a portion of the sequence. Some domains are complex and made up of several smaller homology domains that became joined to form a larger one during evolution. A domain that covers an entire sequence is called the homeomorphic domain by PIR (Barker et al. 1996; see <http://www-nbrf.georgetown.edu/>).

**Module** is a region of conserved amino acid patterns comprising one or more motifs and considered to be a fundamental unit of structure or function. The presence of a module has also been used to classify proteins into families.

**Motif** (sequence context) refers to a conserved pattern of amino acids that is found in two or more proteins. In the Prosite catalog, a motif is an amino acid pattern that is found in a group of proteins that have a similar biochemical activity, and that often is near the active site of the protein. Examples of sequence motif databases are the Prosite catalog (<http://www.expasy.ch/prosite>) and the Stanford Motifs Database (<http://dna.stanford.edu/emotif/>).

**Motif** (structural context) refers to a combination of several secondary structural elements produced by the folding of adjacent sections of the polypeptide chain into a specific three-dimensional configuration. An example is the helix-loop-helix motif. Structural motifs are also referred to as supersecondary structures and folds.

**Position-specific scoring matrix** (sequence context, also known as weight or scoring matrix) represents a conserved region in a multiple sequence alignment with no gaps. Each matrix column represents the variation found in one column of the multiple sequence alignment.

**Position-specific scoring matrix—3D** (structural context) represents the amino acid variation found in an alignment of proteins that fall into the same structural class. Matrix columns represent the amino acid variation found at one amino acid position in the aligned structures (Kelley et al. 2000).

**Primary structure** refers to the linear amino acid sequence of a protein, which chemically is a polypeptide chain composed of amino acids joined by peptide bonds.

**Profile** (sequence context) is a scoring matrix that represents a multiple sequence alignment of a protein family. The profile is usually obtained from a well-conserved region in a multiple sequence alignment. The profile is in the form of a matrix with each column representing a position in the alignment and each row one of the amino acids. Matrix values give the likelihood of each amino acid at the corresponding position in the alignment. The profile is moved along the target sequence to locate the best scoring regions by a dynamic programming algorithm. Gaps are allowed during matching and a gap penalty is included in this case as a negative score when no amino acid is matched. A sequence profile may also be represented by a hidden Markov model, referred to as a profile HMM.

**Profile** (structural context) is a scoring matrix that represents which amino acids should fit well and which should fit poorly at sequential positions in a known protein structure. Profile columns represent sequential positions in the structure, and profile rows represent the 20 amino acids. As with a sequence profile, the structural profile is moved along a target sequence to find the highest possible alignment score by a dynamic programming algorithm. Gaps may be included and receive a penalty. The resulting score provides an indication as to whether or not the target protein might adopt such a structure.

**Quaternary** structure is the three-dimensional configuration of a protein molecule comprising several independent polypeptide chains. A Web site for predicting quaternary structure is described at <http://msd.ebi.ac.uk/Services/Quaternary/quaternary.html>. A database of experimentally identified interacting domains of protein subunits (DIP) is maintained at <http://dip.doe-mbi.ucla.edu>; Xenarios et al. 2000; also see Table 9.5).

**Secondary** structure refers to the interactions that occur between the C=O and NH groups on amino acids in a polypeptide chain to form  $\alpha$  helices,  $\beta$  sheets, turns, loops, and other forms, and that facilitate the folding into a three-dimensional structure.

**Superfamily** is a group of protein families of the same or different lengths that are related by distant yet detectable sequence similarity. Members of a given superfamily thus have a common evolutionary origin. Originally, Dayhoff defined the cutoff for superfamily status as being the chance that the sequences are not related of  $<10^{-6}$ , on the basis of an alignment score (Dayhoff et al. 1978). Proteins with few identities in an alignment of the sequences but with a convincingly common number of structural and functional features are placed in the same superfamily. At the level of three-dimensional structure, superfamily proteins will share common structural features such as a common fold, but there may also be differences in the number and arrangement of secondary structures. The PIR resource uses the term homeomorphic superfamilies to refer to superfamilies that are composed of sequences that can be aligned from end to end, representing a sharing of single sequence homology domain, a region of similarity that extends throughout the alignment. This domain may also comprise smaller homology domains that are shared with other protein families and superfamilies. Although a given protein sequence may contain domains found in several superfamilies, thus indicating a complex evolutionary history, sequences will be assigned to only one homeomorphic superfamily based on the presence of similarity throughout a multiple sequence alignment. The superfamily alignment may also include regions that do not align either within or at the ends of the alignment (Barker et al. 1995, 1996; <http://www-nbrf.georgetown.edu/>). In contrast, sequences in the same family align well throughout the alignment. The SCOP Web site reports 820 superfamilies (release 1.50), and the CATH Web site (version 1.7 beta) reports 900 superfamilies (sites described below).

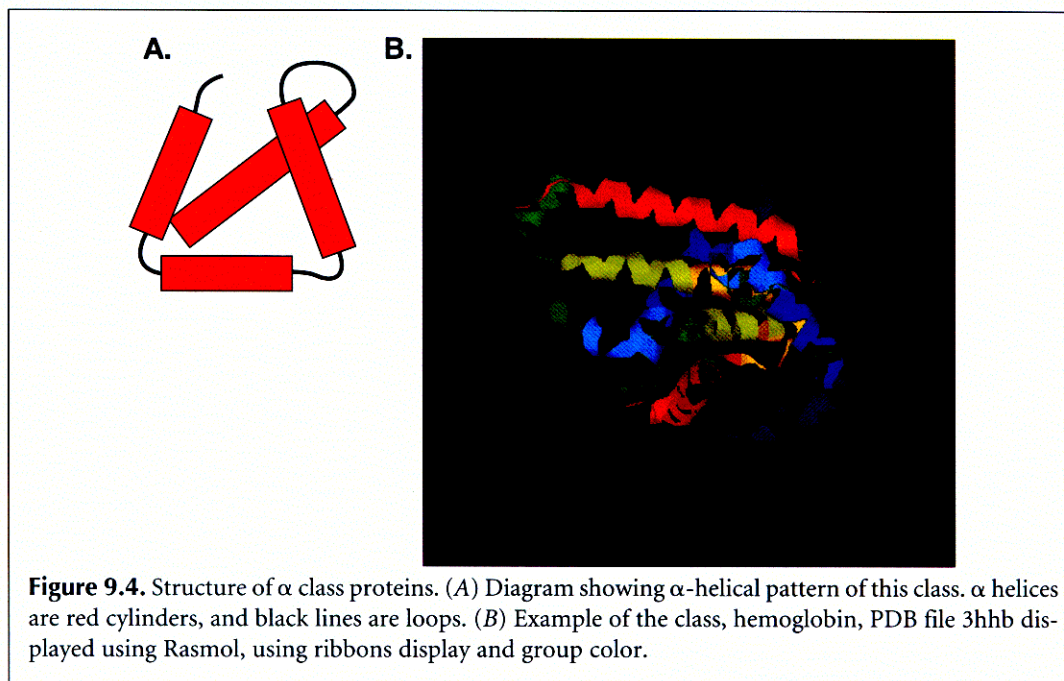
**Supersecondary** structure is a term with similar meaning to a structural motif.

**Tertiary** structure is the three-dimensional or globular structure formed by the packing together or folding of secondary structures of a polypeptide chain.

## Classes of Protein Structure

From the work of Levitt and Chothia (1976), four principal classes of protein structure were recognized based on the types and arrangements of secondary structural elements. These classes are described and illustrated below. In addition, several other classes recognized in the SCOP database discussed below (p. 402) (Murzin et al. 1995) are also included. Examples of this classification are taken from Branden and Tooze (1991).

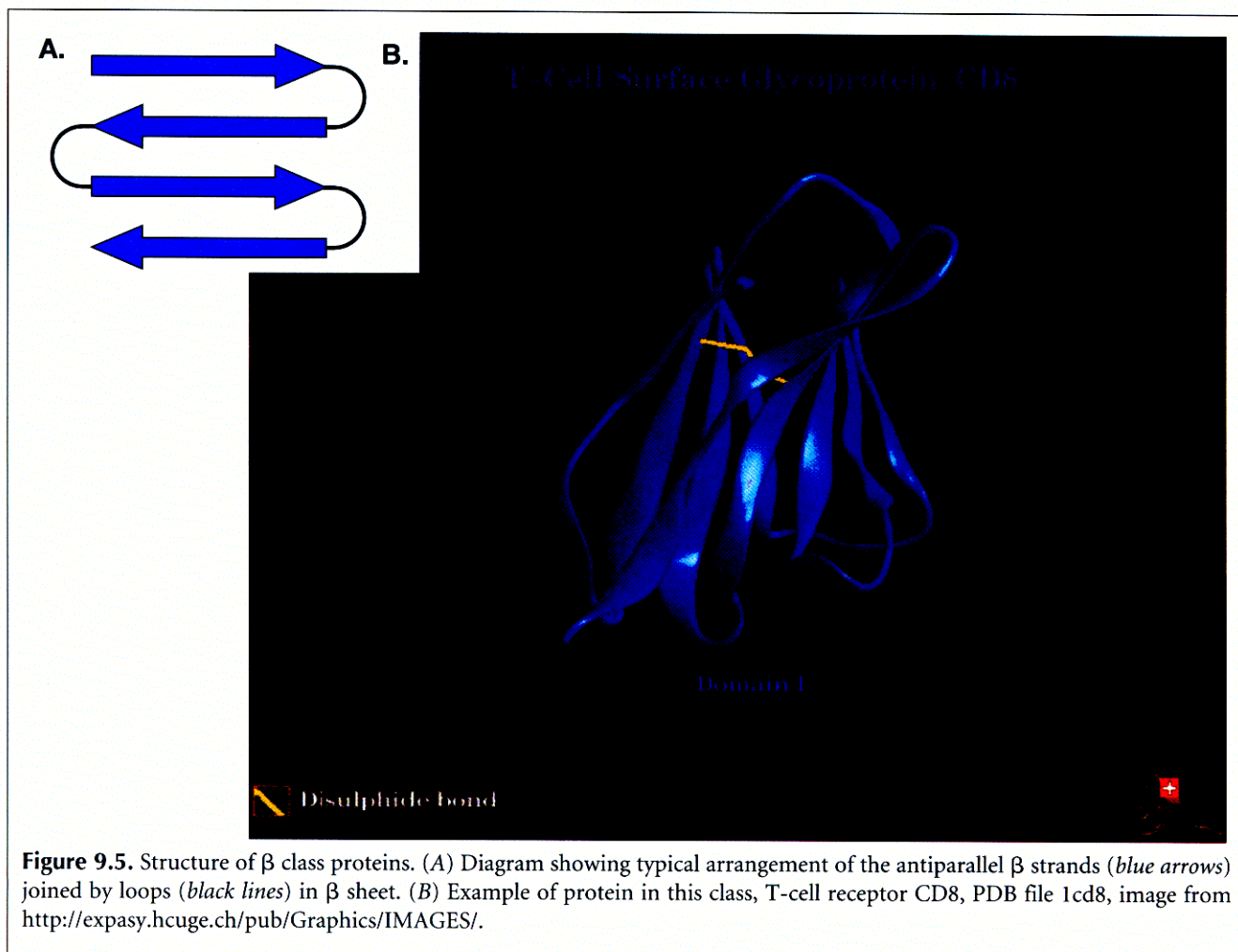
1. Class  $\alpha$  comprises a bundle of  $\alpha$  helices connected by loops on the surface of the proteins (see Fig. 9.4).



2. Class  $\beta$  comprises antiparallel  $\beta$  sheets, usually two sheets in close contact forming a sandwich (see Fig. 9.5). Alternatively, a sheet can twist into a barrel with the first and last strands touching. Examples are enzymes, transport proteins, antibodies, and virus coat proteins such as neuraminidase.
3. Class  $\alpha/\beta$  comprises mainly parallel  $\beta$  sheets with intervening  $\alpha$  helices, but may also have mixed  $\beta$  sheets (see Fig. 9.6). In addition to forming a sheet in some proteins in this class, as illustrated below, in others parallel  $\beta$  strands may form into a barrel structure that is surrounded by  $\alpha$  helices (not shown). This class of proteins includes many metabolic enzymes.
4. Class  $\alpha + \beta$  comprises mainly segregated  $\alpha$  helices and antiparallel  $\beta$  sheets (Fig. 9.7).
5. Multidomain ( $\alpha$  and  $\beta$ ) proteins comprise domains representing more than one of the above four classes.
6. Membrane and cell-surface proteins and peptides excluding proteins of the immune system comprise this class (see Fig. 9.8).

## Protein Databases

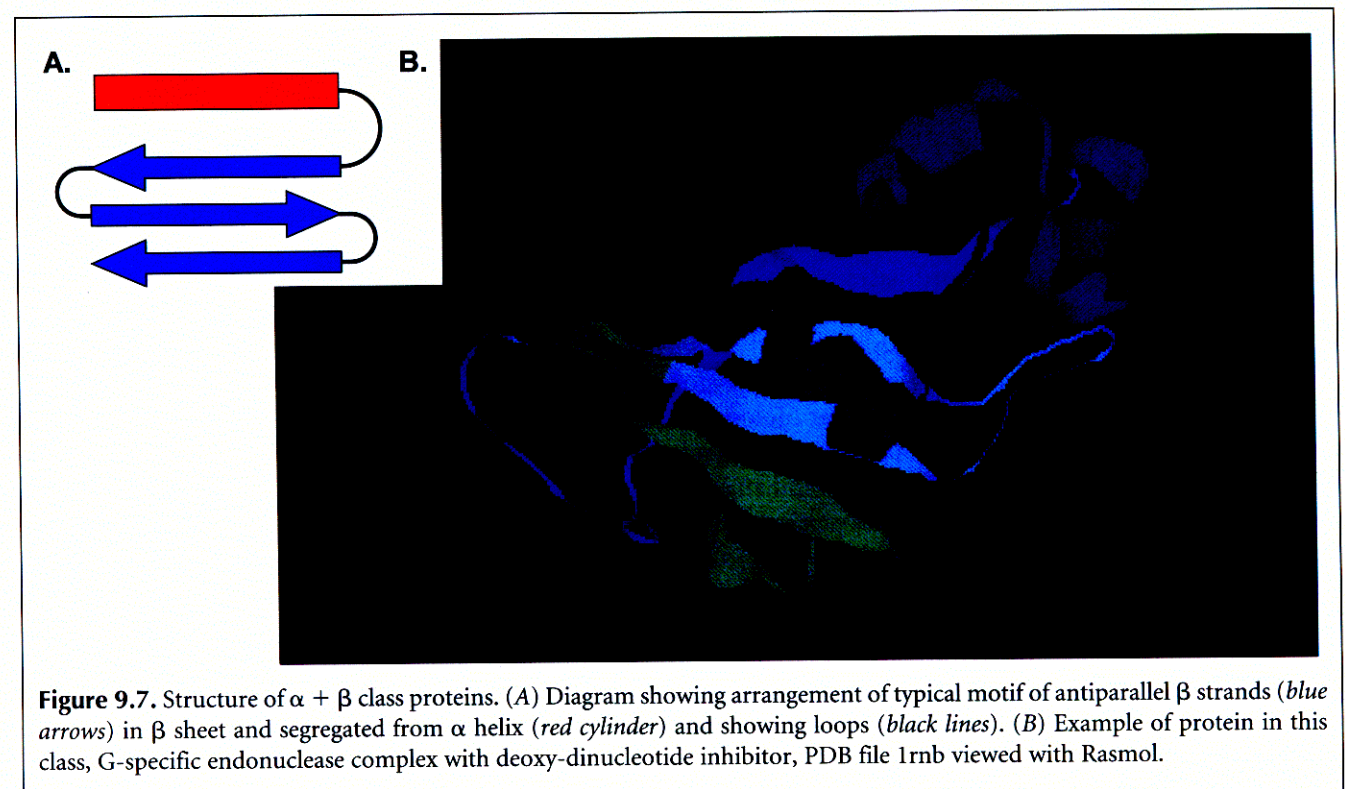
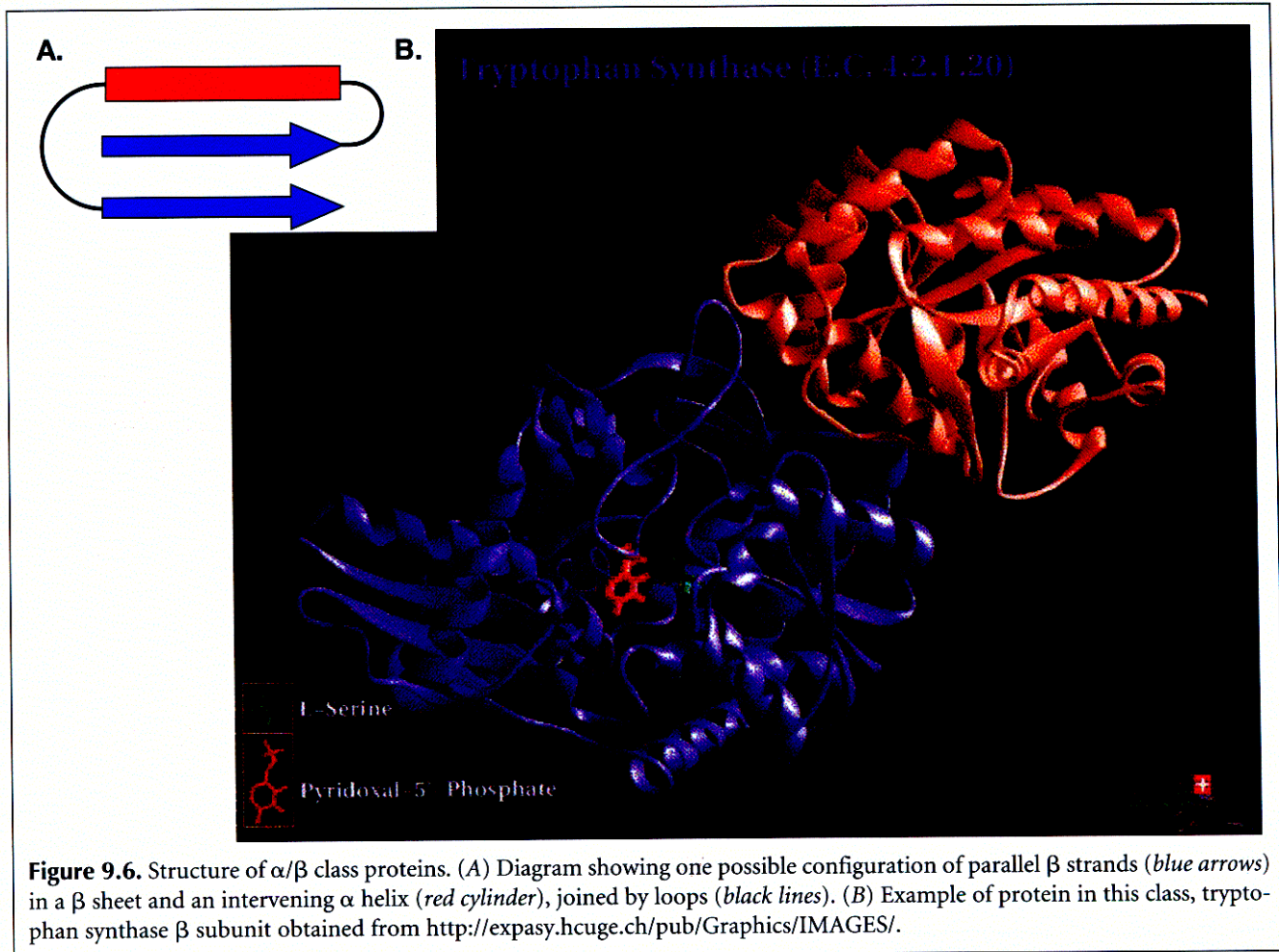
A protein can be analyzed in the laboratory at the levels of sequence and structure. The amino acid sequence and the atomic coordinates of each atom in the structure are unique to each protein. The sequence is obtained in the molecular biology laboratory as a DNA sequence and translated into the amino acid sequence of the encoded protein (see Chapter 8). DNA sequences are deposited in the DNA sequence databases such as GenBank and EMBL, where they are automatically translated to produce the Genpept and TrEMBL protein databases, respectively. Sometimes protein fragments are also sequenced, and matches with DNA sequence databases are used to identify the encoding gene (Chapter 8). The encoded proteins are additionally annotated in databases such as SwissProt and PIR as described in Chapter 2.



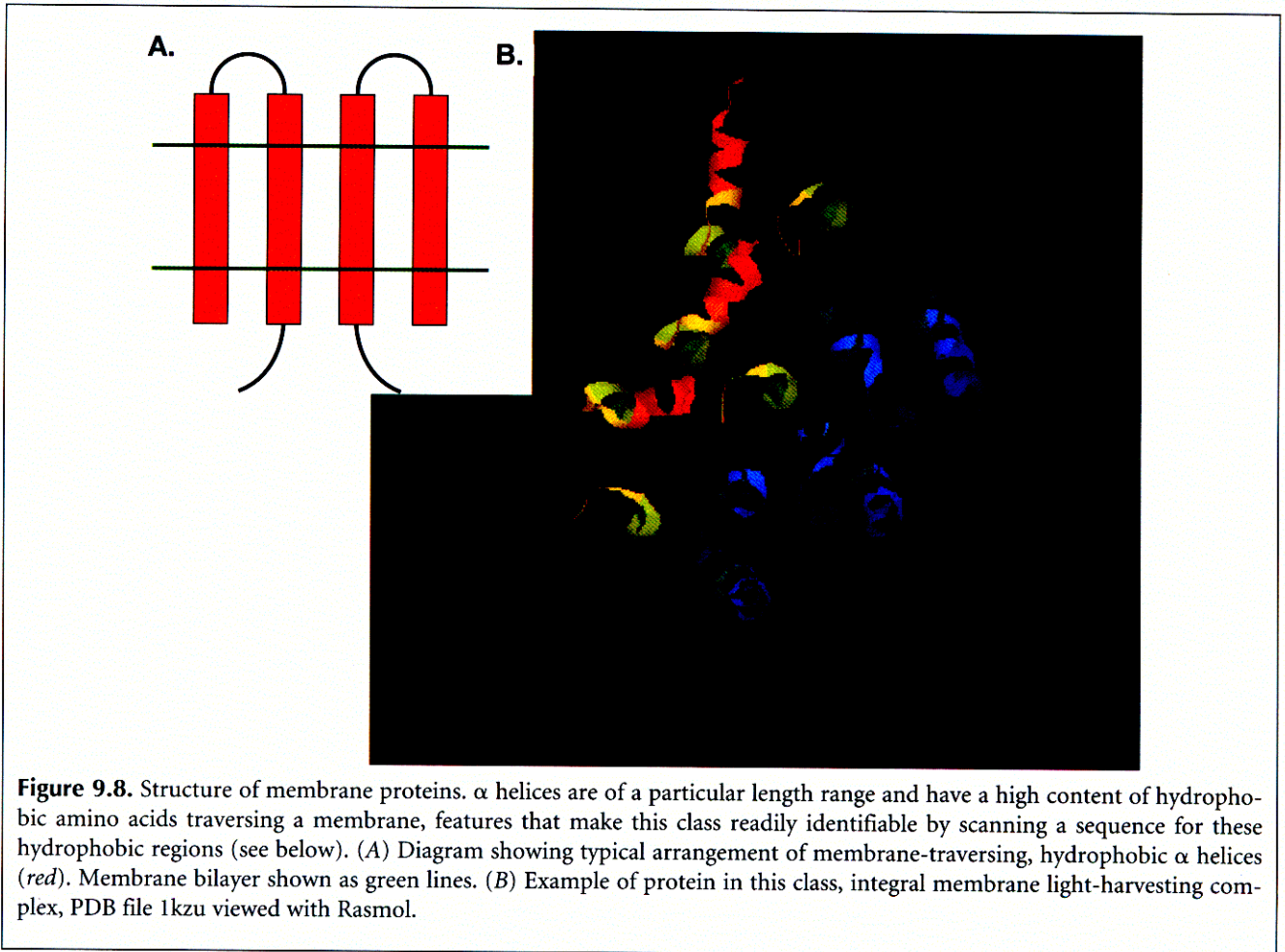
**Figure 9.5.** Structure of  $\beta$  class proteins. (A) Diagram showing typical arrangement of the antiparallel  $\beta$  strands (blue arrows) joined by loops (black lines) in  $\beta$  sheet. (B) Example of protein in this class, T-cell receptor CD8, PDB file 1cd8, image from <http://expasy.hcuge.ch/pub/Graphics/IMAGES/>.

The three-dimensional structure of a protein is usually obtained by making crystals of the protein and using X-ray diffraction to determine the positions of molecules that are fixed within the crystal. The technique of nuclear magnetic resonance (NMR) is also used to obtain protein structures. Once the three-dimensional coordinates of each atom in the protein molecule have been found, a table of these coordinates is deposited with the Brookhaven Data Bank as a PDB entry. PDB entries such as shown in Table 9.3 give the atomic coordinates of the amino acids in proteins, protein fragments, or proteins bound to substrates or inhibitors. PDB files may be easily retrieved from the PDB Web site (<http://www.rcsb.org/pdb/>) and displayed with a molecular viewer such as Rasmol. Structural information may also be stored in forms other than PDB, but PDB is the most accessible for the molecular biologist. There are three different kinds of databases that provide an analysis of proteins, one kind for sequences, a second for structures, and a third for comparing sequences and structures.

As more and more protein structures have been solved by X-ray crystallographic and NMR methods, these structures have been classified by various means into structural databases. This classification is based on comparison and alignment of the protein structures. The types, order, connections, and relative positions of secondary structures are compared using the known atomic coordinates of atoms in each structure and methods described below. This type of information can then be combined with sequence information to identify other proteins that might have similar structural features.







Another type of protein sequence analysis is a sequence alignment of protein sequences discussed in Chapter 3 or a search for similar sequences in the sequence databases, as described in Chapter 7. The alignment will reveal any significant similarity and the degree of amino acid identity between two sequences. Similarity may be present throughout the sequences or localized to certain regions. Localization of sequence similarity can best be performed by global and local sequence alignment methods, as discussed in Chapter 3. The stronger the similarity and identity, the more similar are the three-dimensional folds and other structural features of the proteins. Another level of sequence analysis is examining a group of sequences for common amino acid patterns. Methods for finding different types of patterns, including motifs (short gapped or ungapped patterns), blocks (ungapped patterns), and patterns with gaps (represented by profile scoring matrices and profile HMMs) are discussed in Chapter 4. These patterns may be obtained from sequences of proteins that are already known to have the same function, or they may be obtained by statistical or pattern-finding methods of any set of sequences of biological interest. Depending on the extent and significance of these patterns and additional information about the function of the proteins, their presence may or may not represent structural similarity or an evolutionary relationship among the proteins. A combined form of sequence and structural alignments provides an additional level of analysis.

When proteins of unknown structure are similar to a protein of known structure at the sequence level, multiple sequence alignment and pattern analysis can be used to predict the

**Table 9.3.** Brookhaven Protein Data Bank (PDB) entry 3hhb for deoxy hemoglobin

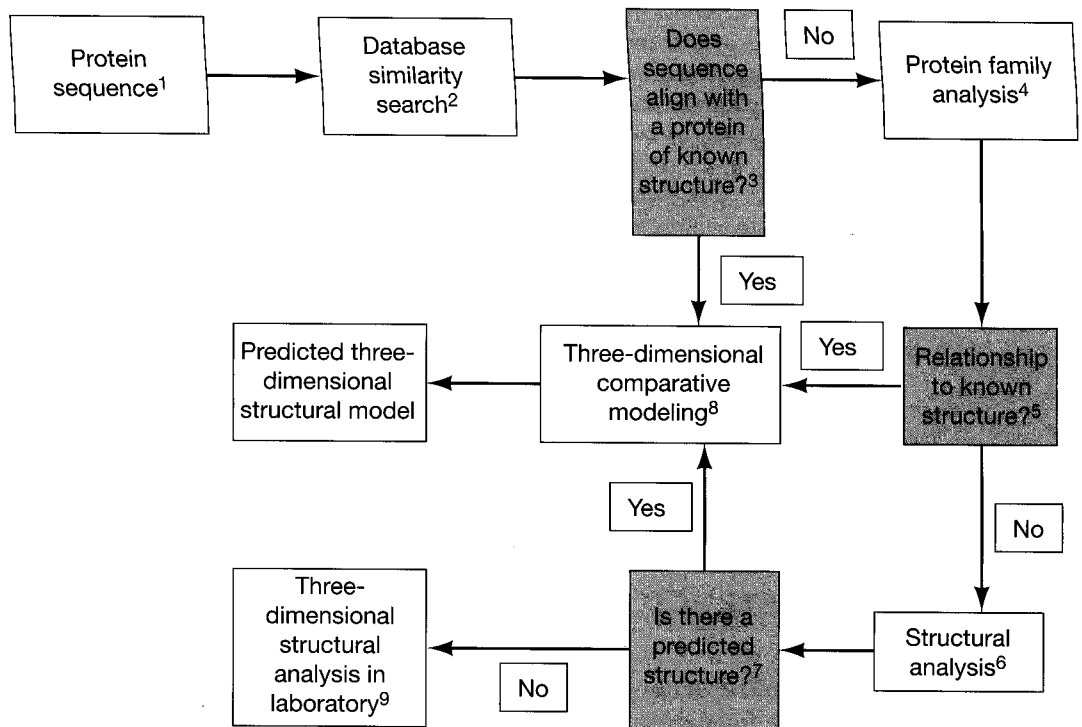
Header compnd	Oxygen transport hemoglobin (deoxy)						13-Jul-93	2hhb	
ATOM	1	N	VAL A	1	5.428	17.064	5.060	1.00	41.29
ATOM	2	CA	VAL A	1	6.168	18.292	4.856	1.00	41.33
ATOM	3	C	VAL A	1	7.676	18.056	5.068	1.00	31.64
ATOM	4	O	VAL A	1	8.120	17.488	6.076	1.00	38.31
ATOM	5	CB	VAL A	1	5.644	19.268	5.884	1.00	52.26
ATOM	6	CG1	VAL A	1	6.044	20.696	5.512	1.00	52.75
ATOM	7	CG2	VAL A	1	4.124	19.120	6.000	1.00	58.75
ATOM	8	N	LEU A	2	8.444	18.512	4.116	1.00	27.63
ATOM	9	CA	LEU A	2	9.896	18.420	4.308	1.00	33.62
ATOM	10	C	LEU A	2	10.360	19.592	5.216	1.00	32.51
ATOM	11	O	LEU A	2	10.128	20.760	4.900	1.00	31.03
ATOM	12	CB	LEU A	2	10.568	18.584	2.932	1.00	34.38
ATOM	13	CG	LEU A	2	10.284	17.488	1.924	1.00	32.23
ATOM	14	CD1	LEU A	2	11.032	17.676	0.580	1.00	36.30
ATOM	15	CD2	LEU A	2	10.576	16.136	2.560	1.00	38.42

Shown is the initial part of the entry showing ATOM records that provide cartesian coordinates of all atoms in the first two amino acids Val and Leu. The last columns give the occupancy and temperature factor for each atom. The occupancy gives the frequency with which the atom is present in the crystal and is usually 1. The temperature gives a measurement of the uncertainty of the position of the atom due to the motion of the atom in the crystal. The units of temperature are Angstroms squared. A typical value of a crystal at room temperature at 2 Å resolution is 20 Å; the higher this value for an atom, the more uncertain the position of that atom. Structural entries sometimes provide the author's assignment of a secondary structure to each amino acid.

structures of these proteins. Databases of such related proteins are available. In another type of analysis, called threading, the sequence of amino acids in a protein of unknown structure is tested for ability to fit into a known three-dimensional structure. The size and chemistry of each amino acid R group and proximity to other amino acids are taken into account. This analysis provides a method for aligning a sequence with a structure.

## METHODS

1. Amino acid sequences of proteins are derived from translation of cDNA sequences or predicted gene structures in genomic DNA sequences. Partial sequences are also derived by translation of expressed sequence tag (EST) sequences or genomic DNA sequences in all six reading frames. These predictions can be improved when genomic and EST sequences can be aligned and when overlapping EST sequences are identified by gene indexing, as described in Chapters 7 and 8.
2. The sequence is used as a query in a database similarity search against the proteins in the Protein Data Bank (PDB), all of which have a known three-dimensional structure. A significant alignment of the query sequence with a PDB sequence is evidence that the query sequence has a similar three-dimensional structure. If a relationship with a PDB protein is not found, then a second database similarity search against a protein sequence database such as SwissProt can be performed. Matching sequences including both closely related and more distantly related ones can then be used in a search against PDB sequences. The PSI-BLAST tool described in Chapter 7 automates and enhances the process of finding related sequences in the protein database. The goal is to discover one or more database sequences that are related both to the query and to a PDB sequence, as illustrated in Figure 7.1.
3. If the database similarity search reveals a significant alignment between the query sequence and a PDB sequence, the alignment between the sequences can be used to position the amino acids of the query sequence in the same approximate three-dimensional structure. Testing the significance of alignment scores is discussed in Chapter 3.



4. Proteins have been classified into families on the basis of sequence similarity. The relationships are depicted in a multiple sequence alignment of the proteins, as described in Chapter 4. Proteins of known three-dimensional structure have also been classified into fold families on the basis of a common arrangement of secondary structures. Sequences of proteins in the same fold family are often not similar, so they cannot be aligned. However, the individual proteins in a particular fold family are often members of families based on sequence similarity. Hence, these similar sequences are also predicted to have the same structural fold as the fold family. The goal of this step in the flowchart is to exploit these structure–sequence relationships. Two questions are addressed: (1) Is the new protein a member of a protein family based on sequence similarity? (2) Does the matched family have a predicted structural fold? The first question is usually addressed by analyzing the test sequence for patterns that represent each family using PSSMs, profile HMMs, and other tools, as described in Chapter 7. Web sites such as Interpro (Table 9.5) include a large, composite collection of patterns and will search a new sequence for matches. 3D-PSSM (Table 9.5) includes a powerful set of scoring matrices based on structural alignments for use in three-dimensional structure prediction. These Web sites usually provide links to related fold families, thus identifying a predicted structural fold for the new protein. Other Web sites employ a cluster analysis of proteins based on pair-wise alignment scores of all of the proteins in the SwissProt database. These sites offer an alternative method for finding relationships between a new sequence and all of the other sequences in SwissProt, and thus for discovering a link to a known protein structure.
5. If the family analysis reveals that the new protein is a member of a family that is predicted to have a structural fold, multiple sequence alignments of these proteins can be used for structural modeling.
6. This step in the flowchart includes several different types of analyses that are described below in the chapter. First, the presence of small amino acid motifs in a protein can be an indicator of a biochemical function. The Prosite catalog can be used to search a new protein sequence for motifs. Second, spacing and arrangement of specific amino acids, e.g., hydrophobic amino acids, provides important structural clues that can be used for modeling. Third, the tendency of certain amino acid combinations to occur in a given type of secondary structure provides methods for predicting where these structures are likely to occur in a new sequence. Fourth, the structural fold families described in note 4 above have been represented by PSSMs and by HMMs that capture the tendency to find each amino acid at a particular position in a structural fold and variations in the fold itself. Other models of three-

dimensional structure represent the size and chemistry of amino acids or the energetic stability associated with amino acid interactions. A new protein sequence can be aligned with these models to determine whether the sequence matches one of them, a procedure known as threading a sequence into a structure.

7. The structural analysis in step 6 provides clues as to the presence of active sites, regions of secondary and three-dimensional structure, and the order of predicted secondary structures. If these predictions are convincing enough, it may be possible to identify a new protein as a member of a known structural class.
8. Sequence or structural alignments of the new protein with a protein of known structure provide a starting three-dimensional model of the protein. By using computer graphics and protein modeling software, the amino acids can then be positioned to accommodate available space and interactions with neighboring amino acids.
9. Proteins that fail to show any relationship to proteins of known structure are candidates for structural analysis. There are approximately 500–600 known fold families, and new structures are frequently found to have an already known structural fold. Accordingly, protein families with no relatives of known structure may represent a novel structural fold.

## VIEWING PROTEIN STRUCTURES

The first major step in displaying a structure is to identify the correct PDB identification code for the structural file. Most sites provide a browser program for searching the structural database for the name of the protein, organism, or other identifying features (see below). There may be a number of choices from which to choose, including domains, folds, or protein fragments, or structures of the protein bound to a substrate or inhibitor. Some databases also include the predicted structure of mutant proteins. The available choices need to be screened carefully for the correct one.

A number of molecular viewers are freely available and run on most computer platforms and operating systems, including Microsoft Windows, Macintosh, and UNIX X-Windows. These programs convert the atomic coordinates into a view of the molecule. They may also recompute information to remove inconsistencies in the database or to supply missing information (Hogue and Bryant 1998a,b). Viewers also provide ways to manipulate the molecule, including rotation, zooming, and creating two images that provide a stereo view. Rotating a molecule by dragging the mouse across the image can illustrate the three-dimensional structure. Viewers can also be used to show a structural alignment of two or more structures or a predicted structure. Unless a very high-resolution view is needed, the simplest way to use a viewer is through a network browser. The browser may be readily configured to run a viewer program automatically when the particular file format used by the viewer is being downloaded from the remote computer. Most sites that provide protein structural files provide several formats allowing a choice of viewers, and they also provide Web links to other sites from which the viewer program may be downloaded. The viewer option usually appears once a particular structural file has been chosen. Shown in Table 9.4 are some representative viewers that are commonly used and their features.

The correct processing of files with molecular structural information through the Web or through E-mail attachments is made possible by the chemical MIME (multipurpose internet mail extension) project (<http://www.ch.ic.ac.uk/chemime/iupac.html>). This project acts as a repository for standard types of MIME files. As an example, if the start of the file includes the label `chemical/x-pdb` (MIME type `chemical` and subtype `x-pdb`), the file is a text file in the Brookhaven Protein Data Bank file format, and a viewer for a `pdb` file such as `Rasmol` or `Chime` is needed. Files intended for viewing by `Rasmol` may also be indicated by MIME type `application/x-rasmol` and the `pdb` file may also be identified by the file-

**Table 9.4.** *Programs for viewing protein molecules*

Viewer	Web location	Features
Chime	<a href="http://www.umass.edu/microbio/chime/">http://www.umass.edu/microbio/chime/</a>	A Web browser plug-in that can be used to display and manipulate structures inside a Web page. There are many mouse-driven controls. Excellent for lecture presentations.
Cn3d <sup>a</sup>	<a href="http://www.ncbi.nlm.nih.gov/Structure/">http://www.ncbi.nlm.nih.gov/Structure/</a> (Hogue 1997)	Provides viewing of three-dimensional structures from Entrez and MMDB. <sup>a</sup> Cn3D runs on Windows, MacOS, and Unix; simultaneously displays structural and sequence alignments; can show multiple superimposed images from NMR studies.
Mage	<a href="http://kinemage.biochem.duke.edu/website/kinhome.html">http://kinemage.biochem.duke.edu/website/kinhome.html</a> (see Richardson and Richardson 1994)	Standard molecular viewing features with animation and kaleidoscope effects.
Rasmol <sup>b</sup>	<a href="http://www.umass.edu/microbio/rasmol/">http://www.umass.edu/microbio/rasmol/</a> (Sayle and Milner-White 1995)	Most commonly used viewer for Windows, MacOS, UNIX, and VMS operating systems. Performs many functions.
Swiss 3D viewer, Spdbv	<a href="http://www.expasy.ch/spdbv/mainpage.html">http://www.expasy.ch/spdbv/mainpage.html</a> (Guex and Peitsch 1997)	Protein models can be built by structural alignments; calculates atomic angles and distances, threading, energy minimation, and interacts with the Swiss Model server.

Additional viewers are accessible from the referenced Web sites. Viewer functions usually include wireframe of C<sub>α</sub> backbone, ribbon of secondary structures, space-filling displays, color schemes to illustrate features such as residues, structures, temperature, mouse-drag rotation, several views including stereo, zooming, and exporting to graphic file formats. Assistance with these viewers is provided at the following Web sites for obtaining molecular coordinates: Molecules R Us at NIH, <http://molbio.info.nih.gov/cgi-bin/pdb>, and NCBI, <http://www.ncbi.nlm.nih.gov/Structure/>. A large list of available graphics viewers may be found at [http://www.csb.yale.edu/user-guides/graphics/csb\\_hm\\_graph.html](http://www.csb.yale.edu/user-guides/graphics/csb_hm_graph.html).

<sup>a</sup> The NCBI structure group has established a new format for databases called ASN.1 (see Chapter 2). The PDB files have been converted into this format to create another database MMDB (Molecular Modelling DataBase) that is highly suitable for structural alignments by vector methods described below. Ambiguities in PDB entries have been made explicit in the MMDB database (Hogue and Bryant 1998a,b; <http://www.ncbi.nlm.nih.gov/Structure/>).

<sup>b</sup> Rasmol and other viewers as well have many features in the molecular viewing window in addition to those described above. These additional features are accessible through a command line window that appears when the program is running.

name extension pdb. There are also additional chemical MIME formats. For Cn3D, chemical/ncbi-asn1-binary and val are the MIME type and filename extension, respectively. Cn3D files are sent as a binary file rather than a text file, meaning that some bytes include characters other than the standard ASCII characters. For MAGE, chemical/x-kinemage and kin are used. Molecules may also be viewed by means of programs called applets written in the JAVA programming language. These programs are sent at the same time as the molecular coordinates and are run by the browser.

In addition to retrieving the three-dimensional coordinates of a molecule, already prepared graphic views of molecules may be obtained from many of the Web sites that provide pdb files. The following FTP site contains a database of stored image files: <http://www.expasy.ch/databases/swiss-3dimage/IMAGES/>. These views include two file formats commonly used on the Web, the JPEG (Joint Photographic Experts Group) format and GIF (graphics interchange format). These formats produce images of a reasonably high quality but have varying levels of detail and resolution. A higher resolution and more detailed rendition of the molecule will have a larger file size and take longer to retrieve over the Internet. These files may be compressed to a smaller size by graphic format conversion programs. Programs such as Raster3D (<http://www.bmsc.washington.edu/raster3d/>) and Molscrip (<http://www.avatar.se/molscrip/>) produce very high-quality images in a number of different formats. These programs require graphics work stations and a more sophisticated level of programming experience.

## PROTEIN STRUCTURE CLASSIFICATION DATABASES

The following databases are accessible on the Web and provide up-to-date structural comparisons for the proteins currently in the Brookhaven PDB and access to the sequences of these proteins. The methods used to classify the protein structures in these databases vary from manual examination of structures to fully automatic computer algorithms. Hence, although one can expect to find roughly the same groupings in each database, there will be some structural relationships that are only identified by one of these methods. Each database has useful information that may be lacking in the others. The MMDB and SARF databases (4 and 5 below) are based on a rapid structural alignment method that is designed to find the most significant alignments in the structural databank. The SCOP, CATH, and FSSP databases (1, 2, and 3) are based on different comparison methods and are likely to provide additional complementary information on relationships among protein structures. These classification schemes have been reviewed previously (Swindells et al. 1998).

1. *The SCOP database.* The SCOP (structural classification of proteins) database (Murzin et al. 1995; Brenner et al. 1996), based on expert definition of structural similarities, is located at <http://scop.mrc-lmb.cam.ac.uk/scop/>. Following classification by class, SCOP additionally classifies protein structures by a number of hierarchical levels to reflect both evolutionary and structural relationships; namely family, superfamily, and fold. Shown in Figure 9.9 is an example of the lineage for the all  $\alpha$  class, globin-like fold, globin-like superfamily, globin, and phycocyanin families, and finally protein domains such as hemoglobin 1 which can be viewed by individual entry in PDB using a molecular viewer.
2. *The CATH database.* The CATH (classification by class, architecture, topology, and homology) protein structure database resides at University College, London (Orengo et al. 1997; <http://www.biochem.ucl.ac.uk/bsm/cath/>). Proteins are classified first into hierarchical levels by class, similar to the SCOP classification except that  $\alpha/\beta$  and  $\alpha+\beta$  proteins are considered to be in one class. Instead of a fourth class for  $\alpha+\beta$  proteins, the fourth class of CATH comprises proteins with few secondary structures. Following class, proteins are classified by architecture, fold, superfamily, and family. Similar structures are found by the program SSAP, described on page 419. An example of a CATH entry is shown in Figure 9.10.
3. *The FSSP database.* The FSSP (fold classification based on structure-structure alignment of proteins) is based on a structural alignment of all pair-wise combinations of the proteins in the Brookhaven structural database by the structural alignment program DALI (Holm and Sander 1996; <http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html>). PDB has a number of redundant structures of proteins whose sequences and structures are 25% or more identical. A subset of representative structures in PDB without these redundant entries was first produced by aligning all of the PDB structures with DALI. Each protein in the subset was then subdivided into individual domains. These domains were then aligned structurally with DALI to identify the common folds. Redundant folds were again eliminated, and a set of representative folds was chosen. From 8320 PDB entries, 947 representative structures, 1484 domains, and 540 structurally distinct fold types were identified in 1997 (Holm and Sander 1998). These fold types represent a unique configuration of secondary structural elements in the domains. For example, one fold might be composed of helix-strand-helix-6 strands joined by loops in a particular configuration.

Corresponding to each representative fold type, there is a cluster of folds that are of the same approximate structure. The domains that have a given cluster of folds are structurally related, and the cluster is represented by structural alignments of these

domains. The higher the statistical score for a given domain alignment and corresponding fold (higher  $Z$  value), the greater the degree to which the atoms occupy similar structural positions.  $Z$  values  $>16$  indicate a very good structural alignment, 8–16 a less good alignment, until a level of 2, which indicates the lowest level of alignment detection, is reached. Thus, fold clusters may be organized in a hierarchical fashion with folds represented by the most low-scoring alignments at the top of the hierarchy, as illustrated in Figure 9.11, FSSP, part D.

In addition, the sequences of the 1000 representative structures were used as probes for a sequence similarity search of the SwissProt protein sequence database. The database search program MAXHOM, which begins with a sequence similarity search and then with an expanded profile search, was used, as discussed in Chapter 7. The resulting homology-derived structures of proteins (HSSP) database (Sander and Schneider 1991; Dodge et al. 1998; <http://www.sander.ebi.ac.uk/hssp/>) contains lists of similar proteins, one list for each representative structure. Given the PDB database number of a known structure, the program will show the closest representative structures, and one or more may be chosen. The program will then show any significant structural alignments between the chosen representative and other representative structures in FSSP. A structural alignment between the chosen representative and each of the matching proteins in the HSSP database entry for that representative may be selected. An example of searching for a structural and sequence similarity using the FSSP and HSSP databases is shown in Figure 9.11.

4. *MMDB (molecular modelling database)*. Proteins of known structure in the Brookhaven PDB have been categorized into structurally related groups in MMDB by the VAST (Vector Alignment Search Tool) structural alignment program (Madej et al. 1995). VAST aligns three-dimensional structures based on a search for similar arrangements of secondary structural elements (see Fig. 9.12). This method provides a method for rapidly identifying PDB structures that are statistically out of the ordinary. MMDB has been further incorporated into the ENTREZ sequence and reference database at <http://www.ncbi.nlm.nih.gov/Entrez> (Hogue et al. 1996). Accordingly, it is possible to perform a simultaneous search for similar sequences and structures, designated neighbors, at the ENTREZ Web site. Structural neighbors within MMDB are based on detailed residue-by-residue alignments.
5. *The SARF database*. The SARF (spatial arrangement of backbone fragments) database at <http://www-lmmb.ncifcrf.gov/~nicka/sarf2.html/> (Alexandrov and Fischer 1996) also provides a protein database categorized on the basis of structural similarity. Like VAST, SARF can find structural similarity rapidly based on a search for secondary structural elements. These structural hierarchies found by this method are in good agreement with those found in the SCOP, CATH, and FSSP databases with several interesting differences. The method also found several new groupings of structural similarity. The SARF Web site provides a similarity-based tree of structures at <http://www-lmmb.ncifcrf.gov/~nicka/tree.html/> and some excellent representations of overlaid structures.

## ALIGNMENT OF PROTEIN STRUCTURES

As more and more protein structures, as well as access to recently developed and rapid methods for comparing protein structures, have become available on the Web, alignment of protein structures has become a task achievable by laboratories not trained in the techniques of structural biology. To perform a sequence alignment, the amino acid sequence of one pro-

## Structural Classification of Proteins





















**Fold: Globin-like**

core: 6 helices; folded leaf, partly opened;

**Lineage:**

1. Root: scop
2. Class: All alpha
3. Fold: Globin-like  
core: 6 helices; folded leaf, partly opened;

**Superfamilies:**

1. Globin-like (2)
  1. Globins (37)  
Heme-binding protein
    1. Hemoglobin I
      1. ark clam (*Scapharca inaequivalvis*) (6) 
      2. clam (*Lucina pectinata*) (2) 
    2. Glycera globin
      1. marine bloodworm (*Glycera dibranchiata*) (2) 
    3. Myoglobin
      1. sperm whale (*Physeter catodon*) (76) 
      2. sea hare (*Aplysia limacina*) (6) 
      3. common seal (*Phoca vitulina*) (1) 
      4. pig (*Sus scrofa*) (8) 
      5. horse (*Equus caballus*) (6) 
      6. human (*Homo sapiens*) (5) 
      7. asian elephant (*Elephas maximus*) (1) 
      8. Loggerhead sea turtle (*Caretta caretta*) (2) 
      9. yellowfin tuna (*Thunnus albacares*) (1) 
    4. Erythrocrurin
      1. Midge (*Chironomus thummi thummi*), Fraction III (4) 
    5. Leghemoglobin
      1. yellow lupin (*Lupinus luteus* L) (17) 
      2. Soybean (*Glycine max*), isoform A (2) 
    6. Hemoglobin, alpha-chain
      1. human (*Homo sapiens*) (39)  
      2. horse (*Equus caballus*) (3) 
      3. deer (*Odocoileus virginianus*) (1) 
      4. bovine (*Bos taurus*) (1) 



- 5. pig (*Sus scrofa*) (1)
- 6. trout (*Oncorhynchus mykiss*) (2)
- 7. antarctic fish (*Pagothenia bernacchii*) (2)
- 8. teleost fish *leiosomus xanthurus* (1)
- 7. Hemoglobin, beta-chain
  - 1. human (*Homo sapiens*) (38)
  - 2. human fetus (*Homo sapiens*), gamma-chain (1)
  - 3. horse (*Equus caballus*) (3)
  - 4. deer (*Odocoileus virginianus*) (1)
  - 5. bovine (*Bos taurus*) (1)
  - 6. pig (*Sus scrofa*) (1)
  - 7. trout (*Oncorhynchus mykiss*) (2)
  - 8. antarctic fish (*Pagothenia bernacchii*) (2)
  - 9. teleost fish *leiosomus xanthurus* (1)
- 8. Lamprey globin
  - 1. sea lamprey (*Petromyzon marinus*) (1)
- 9. Ascaris hemoglobin, domain 1
  - 1. pig roundworm (*Ascaris suum*) (1)
- 10. Hemoglobin
  - 1. innkeeper worm (*Urechis caupo*) (1)
- 11. Hemoglobin
  - 1. sea cucumber (*Caudina (Molpadia) arenicola*) (2)
- 12. Flavohemoglobin, N-terminal domain
  - 1. *Alcaligenes eutrophus* (1)
- 2. **Phycocyanins** (6)
 

*oligomers of two different types of homologous subunits  
each subunit contains 2 additional helices at the N-terminus  
binds a chromophore*

  - 1. Phycocyanin
    - 1. red alga (*Cyanidium caldarium*) (1)
  - 2. C-phycocyanin
    - 1. cyanobacterium (*Fremyella diplosiphon*) (1)
  - 3. Allophycocyanin
    - 1. (*Spirulina platensis*) (1)
  - 4. R-phycoerythrin
    - 1. red algae (*Polysiphonia urceolata*) (1)
  - 5. B-phycoerythrin
    - 1. Red alga (*Porphyridium sordium*) (1)
  - 6. Phycoerythrocyanin
    - 1. Thermophilic cyanobacterium (*Mastigocladus laminosus*) (1)

Enter search key:

Generated from scop database 1.37 Development with scopm 1.087 on Mon Feb 16 20:03:02 1998  
 Copyright © 1994-1998 The scop authors / scop@mrc-lmb.cam.ac.uk

Figure 9.9. A portion of the SCOP structural classification showing the hierarchy of all  $\alpha$ -class, globin, and globin-like proteins.

## A. CATH Search

### Search results for 2reb

**Name:** Rec a protein (e.c.3.4.99.37)

**Source:** (*escherichia coli*)

Summary:

*for multi-chain proteins, click on any chain to see a more detailed description...*

PDB Code	Chain	Status
2reb	-	In CATH

#### 2 assigned domains

Domain 1: residues 27 to 269

[Goto CATH entry](#)

<b>Class</b>	3	Alpha Beta
<b>Architecture</b>	40	3-Layer(aba) Sandwich
<b>Topology</b>	50	Rossmann fold (Nitrogenase Molybdenum-Iron Protein, subunit A, domain 3)
<b>Homologous superfamily</b>	1200	2reb domain 1

Domain 2: residues 270 to 328

[Goto CATH entry](#)

<b>Class</b>	3	Alpha Beta
<b>Architecture</b>	30	2-Layer Sandwich
<b>Topology</b>	250	Rec A Protein, domain 2
<b>Homologous superfamily</b>	10	2reb domain 2

[cath@biochem.ucl.ac.uk](mailto:cath@biochem.ucl.ac.uk)

**Figure 9.10.** CATH entry for *E. coli* RecA protein (PDB 2reb). (A) CATH classification of the protein. (B) Ancillary information provided by CATH database including structure, sequence-secondary structure alignment, a structural image, and links to other databases. *Figure continues on next pages.*

tein is written above the amino acid sequence of a second protein. Similar or identical amino acids are placed in the same columns and gaps are placed at positions where there is no matching character. In performing structural alignments, the three-dimensional structure of one protein domain is superimposed upon the three-dimensional structure of a second protein domain, fitting together the atoms as closely as possible so that the average deviation between them is minimum. Sequence alignments are performed to discover sequence similarity, and structural alignments are done to discover structural similarity (evidence that the structures share a common fold). New structural relationships are being constantly discovered. Just as a laboratory may discover a remote sequence similarity between two protein domains reflecting a family or superfamily relationship, so may the same laboratory discover a previously unknown structural relationship between two proteins.

There is one important difference between sequence and structural similarity, however. Statistically significant sequence similarity is an indicator of an evolutionary relationship between sequences. In contrast, significant structural similarity is common, even among

B.  View 1 **PDB code: 2reb**

READ ME  RasMol script  RasMol  VRML v.1.0

**Self-cleavage stimulation**

**Structure:** *Rec a protein*  
**Source:** (*Escherichia coli*)

**Resolution:** 2.30Å. **R-factor:** 0.210.

**Authors:** R.M.Story, T.A.Steitz - **Date:** 06-Mar-92

**Further information:** [PDB header](#) (including references), [3DB Browser](#) and coords, complete [MacroMolecule](#), [MMDB](#) entry, [CATH](#) and [SCOP](#) classifications, [FSSP](#) structural alignments, [PROCHECK](#) summary, [PDBREPORT](#), [PROMOTIF](#) analyses.

**Enzyme Classification number:** (from PDB file: [E.C.3.4.99.37](#))

**SWISS-PROT entry:** [RECA\\_ECOLI](#)

**Molecule(s) in PDB file 2reb:**

Protein: 303 residues

- **CATH classification:**  
 Domain 1: [3.40.560.10](#) -> *Class: Alpha Beta. Architecture: 3-Layer(aba) Sandwich.*  
 Domain 2: [3.30.250.10](#) -> *Class: Alpha Beta. Architecture: 2-Layer Sandwich.*

RasMol domains Protein coloured by domain.

Secondary structure plot

- **PROMOTIF summary:**  
 3 sheets, 14 strands, 14 helices, 16 beta turns, 5 gamma turns, 8 beta bulges, 6 beta hairpins, 2 beta alpha beta units.
- **TOPS** protein topology cartoon

- SAS - annotated FASTA alignment of related sequences in the PDB

- **PROSITE** pattern present in this chain:-

PROSITE pattern PS00321 **PS00321** - RECA.  
 Ala214->Arg222: ALKFYA VR

[Help!](#)

- **MolScript picture** (PostScript file)

136 water molecules.

Enter new **PDB code**

Figure 9.10. Continued.



### A. FSSP: select structural neighbours of 1tupA

Please cite: L. Holm and C. Sander (1996) Science 273(5275):595-60.

#### Select (check) structural neighbours to display

**3D superimposition** **Multiple alignment** **Multiple families** **Reset selection**

STRID2 Z RMSD LALI LSEQ2 %IDE PROTEIN

<input type="checkbox"/>	<u>1tupA</u>	38.8	0.0	196	196	100	tumor suppressor p53 DNA (5'-d(t
<input type="checkbox"/>	<u>1tsrA</u>	38.5	0.0	196	196	100	p53 tumor suppressor DNA
<input type="checkbox"/>	<u>1ycsA</u>	34.6	0.4	191	191	100	p53 fragment 53bp2 fragment (p53
<input type="checkbox"/>	<u>1tupB</u>	33.9	0.8	194	194	100	tumor suppressor p53 DNA (5'-d(t
<input type="checkbox"/>	<u>1tsrB</u>	33.9	0.8	194	194	100	p53 tumor suppressor DNA
<input type="checkbox"/>	<u>1tsrC</u>	33.6	0.8	194	195	100	p53 tumor suppressor DNA
<input type="checkbox"/>	<u>1tupC</u>	33.6	0.8	194	195	100	tumor suppressor p53 DNA (5'-d(t
<input type="checkbox"/>	<u>1a02N</u>	7.4	3.8	139	280	8	nfat fragment (nf-at) biological
<input type="checkbox"/>	<u>1ba1A</u>	6.3	4.7	93	997	9	beta-galactosidase
<input type="checkbox"/>	<u>1a3aA</u>	6.1	3.7	119	285	5	nuclear factor-kappa-b p52 fragm
<input type="checkbox"/>	<u>1rhoA</u>	6.1	3.0	102	145	4	rho gdp-dissociation inhibitor 1
<input type="checkbox"/>	<u>3dpa</u>	5.4	3.3	103	218	6	PapD
<input type="checkbox"/>	<u>1ctn</u>	5.0	2.9	86	538	11	Chitinase a (ph 5.5, 4 degrees c
<input type="checkbox"/>	<u>1mspA</u>	5.0	3.0	93	124	5	major sperm protein (msp)
<input type="checkbox"/>	<u>1mfa</u>	4.8	2.7	88	229	7	Fv fragment (murine se155-4) com
<input type="checkbox"/>	<u>1f13A</u>	4.8	3.9	116	721	8	cellular coagulation factor xiii
<input type="checkbox"/>	<u>1ddt</u>	4.7	3.7	114	523	7	Diphtheria toxin (dimeric)
<input type="checkbox"/>	<u>1eut</u>	4.7	3.6	97	601	15	sialidase (neuraminidase)
<input type="checkbox"/>	<u>1hcz</u>	4.7	3.5	105	250	6	cytochrome f
<input type="checkbox"/>	<u>1xbrA</u>	4.6	3.5	106	184	6	t protein fragment DNA
<input type="checkbox"/>	<u>1cdy</u>	4.6	2.6	78	178	10	t-cell surface glycoprotein cd4
<input type="checkbox"/>	<u>1clc</u>	4.4	3.4	91	541	3	endoglucanase celd (1,4-beta-d-g
<input type="checkbox"/>	<u>1ten</u>	4.4	2.7	83	89	1	Tenascin (third fibronectin type
<input type="checkbox"/>	<u>1fnf</u>	4.4	2.8	83	368	10	fibronectin
<input type="checkbox"/>	<u>1tcrA</u>	4.4	3.5	92	202	7	alpha, beta t-cell receptor (vb8
<input type="checkbox"/>	<u>1neu</u>	4.3	3.0	88	115	5	myelin p0 protein fragment
<input type="checkbox"/>	<u>1vcaA</u>	4.3	3.1	81	199	6	human vascular cell adhesion mol
<input type="checkbox"/>	<u>1lla</u>	4.3	2.8	98	600	7	Hemocyanin (subunit type ii)
<input type="checkbox"/>	<u>1nkr</u>	4.3	2.8	86	195	9	p58-cl42 kir fragment (killer ce
<input type="checkbox"/>	<u>1tf4A</u>	4.2	3.6	106	605	8	t. fusca endoEXO-CELLULASE E4 CA
<input type="checkbox"/>	<u>1tvdA</u>	4.0	3.1	88	116	7	t cell receptor fragment (es204
<input type="checkbox"/>	<u>1ah1</u>	4.0	3.2	92	129	12	ctla-4 fragment (cd152) biologic
<input type="checkbox"/>	<u>1bec</u>	3.9	3.4	99	238	5	14.3.D t cell antigen receptor M
<input type="checkbox"/>	<u>1aohA</u>	3.9	3.3	100	143	11	cellulosome-integrating protein

Continues on next page

<input type="checkbox"/>	<u>2mcm</u>	3.9	3.2	83	112	9	Macromomycin
<input type="checkbox"/>	<u>1oakL</u>	3.8	3.1	90	212	3	nmc-4 igg1 fragment von willebra
<input type="checkbox"/>	<u>1rsy</u>	3.8	3.0	86	135	7	Synaptotagmin i (first c2 domain
<input type="checkbox"/>	<u>1bf5A</u>	3.8	3.4	119	545	7	stat-1 biological_unit DNA
<input type="checkbox"/>	<u>1iatB</u>	3.8	3.8	96	444	5	igg2a intact antibody - mab231
<input type="checkbox"/>	<u>ZahlA</u>	3.8	3.3	110	293	5	alpha-hemolysin (alphatoxin) bio
<input type="checkbox"/>	<u>1cd8</u>	3.8	3.4	89	114	9	Cd8 (t cell c0-receptor, n-termi
<input type="checkbox"/>	<u>1nbcA</u>	3.8	3.5	90	155	8	cellulosomal scaffolding protein
<input type="checkbox"/>	<u>1cto</u>	3.7	3.4	88	109	6	granulocyte colony-stimulating f
<input type="checkbox"/>	<u>1exa</u>	3.7	3.1	88	110	11	Exo-1,4-beta-d-glycanase (cellul
<input type="checkbox"/>	<u>1tit</u>	3.6	2.7	77	89	5	titin, i27 (connectin i27, titin
<input type="checkbox"/>	<u>1aof</u>	3.6	3.0	77	639	8	Galactose oxidase (ph 4.5)
<input type="checkbox"/>	<u>1hnf</u>	3.6	3.3	75	179	7	Cd2 (human)
<input type="checkbox"/>	<u>2ncm</u>	3.6	3.2	84	99	10	neural cell adhesion molecule fr
<input type="checkbox"/>	<u>1ebaA</u>	3.6	3.2	77	212	8	epo receptor fragment (ebp) epo
<input type="checkbox"/>	<u>1edhA</u>	3.6	2.9	82	211	6	e-cadherin (epithelial cadherin
<input type="checkbox"/>	<u>1kb5B</u>	3.6	3.2	86	117	4	kb5-c20 t-cell antigen receptor
<input type="checkbox"/>	<u>1cfp</u>	3.6	3.0	82	205	5	Drosophila neuroglial (chymotryp
<input type="checkbox"/>	<u>1bihA</u>	3.5	3.4	83	391	4	hemolin
<input type="checkbox"/>	<u>1cid</u>	3.5	2.5	67	177	9	Cd4 (domains 3 and 4)
<input type="checkbox"/>	<u>1axiB</u>	3.4	3.4	81	191	4	growth hormone (hgh) Mutant grow
<input type="checkbox"/>	<u>1amx</u>	3.4	3.8	91	150	7	collagen adhesin fragment (cbd19
<input type="checkbox"/>	<u>1xsoA</u>	3.4	3.1	86	150	8	Cu, zn superoxide dismutase
<input type="checkbox"/>	<u>1zxa</u>	3.4	3.2	82	192	7	intercellular adhesion molecule-
<input type="checkbox"/>	<u>1bauA</u>	3.3	4.3	75	208	4	gp130 fragment
<input type="checkbox"/>	<u>1lrhI</u>	3.2	3.0	74	95	7	antibody a6 fragment interferon-
<input type="checkbox"/>	<u>1tlk</u>	3.2	3.1	77	103	11	Telokin
<input type="checkbox"/>	<u>1aba</u>	3.2	3.6	98	858	9	chitobiase (beta-n-acetylhexosam
<input type="checkbox"/>	<u>1pamA</u>	3.1	3.0	69	686	13	cyclodextrin glucanotransferase
<input type="checkbox"/>	<u>1rlw</u>	3.1	3.3	79	126	15	phospholipase a2 fragment (calb
<input type="checkbox"/>	<u>1bdvA</u>	3.1	3.8	81	123	4	protein kinase c fragment (pkc)
<input type="checkbox"/>	<u>1who</u>	3.0	3.1	69	94	6	allergen phl p 2 (phl p ii)
<input type="checkbox"/>	<u>1itbB</u>	3.0	3.6	85	310	5	interleukin-1 beta biological_un
<input type="checkbox"/>	<u>1kcw</u>	3.0	3.7	98	1017	6	ceruloplasmin biological_unit
<input type="checkbox"/>	<u>1bd9A</u>	2.9	3.4	101	180	6	phosphatidylethanolamine binding
<input type="checkbox"/>	<u>1aadB</u>	2.8	3.4	73	99	4	b*0801 fragment (b8) beta-2 micr
<input type="checkbox"/>	<u>1iakA</u>	2.8	3.3	74	182	7	mhc class ii i-ak hen eggwhite l
<input type="checkbox"/>	<u>1plc</u>	2.8	3.7	82	99	5	Plastocyanin (cu2+, ph 6.0)
<input type="checkbox"/>	<u>1acc</u>	2.8	3.1	83	665	5	anthrax protective antigen (pa)
<input type="checkbox"/>	<u>1ksr</u>	2.7	3.3	80	100	8	gelation factor fragment (abp-12
<input type="checkbox"/>	<u>1bhaA</u>	2.7	3.6	92	611	7	beta-glucuronidase (gus gene pro
<input type="checkbox"/>	<u>1nwpA</u>	2.7	3.4	88	128	6	azurin
<input type="checkbox"/>	<u>1dixB</u>	2.7	3.3	82	561	10	phosphoinositide-specific phosph
<input type="checkbox"/>	<u>1aac</u>	2.7	2.7	72	104	7	amicyanin
<input type="checkbox"/>	<u>1kum</u>	2.6	3.6	76	108	9	glucoamylase fragment (1,4-alpha

Figure 9.11. Continued.

<input type="checkbox"/>	<a href="#">1aozA</a>	2.6	3.6	65	552	11	Ascorbate oxidase
<input type="checkbox"/>	<a href="#">1aol</a>	2.5	4.0	93	227	10	gp70 fragment (su)
<input type="checkbox"/>	<a href="#">1ar1B</a>	2.5	5.2	89	252	10	cytochrome c oxidase (cytochrome
<input type="checkbox"/>	<a href="#">1ahsA</a>	2.4	4.4	84	126	6	african horse sickness virus (se
<input type="checkbox"/>	<a href="#">1iakB</a>	2.4	3.6	70	185	10	mhc class ii i-ak hen eggwhite l
<input type="checkbox"/>	<a href="#">4kbpA</a>	2.3	3.2	93	424	5	purple acid phosphatase
<input type="checkbox"/>	<a href="#">1preA</a>	2.3	4.4	71	449	11	proaerolysin
<input type="checkbox"/>	<a href="#">1cwpA</a>	2.3	2.9	72	149	14	cowpea chlorotic mottle virus (c
<input type="checkbox"/>	<a href="#">7paz</a>	2.3	3.8	74	123	6	pseudoazurin Mutant biological_u
<input type="checkbox"/>	<a href="#">1dupA</a>	2.2	3.7	86	136	2	deoxyuridine 5'-triphosphate nuc
<input type="checkbox"/>	<a href="#">1cvx</a>	2.2	3.7	85	158	4	cyoa fragment Mutant biological_
<input type="checkbox"/>	<a href="#">1etb1</a>	2.2	3.5	71	118	9	Transthyretin (prealbumin) mutan
<input type="checkbox"/>	<a href="#">8atcB</a>	2.2	3.0	61	146	5	Aspartate carbamoyltransferase (
<input type="checkbox"/>	<a href="#">1iktA</a>	2.1	3.8	73	104	2	tailspike protein fragment (late
<input type="checkbox"/>	<a href="#">1svb</a>	2.1	2.5	62	395	15	tick-borne encephalitis virus gl
<input type="checkbox"/>	<a href="#">1bvp1</a>	2.1	4.2	86	349	7	Bluetongue virus 10 (usa) vp7 (b
<input type="checkbox"/>	<a href="#">1nls</a>	2.1	4.3	91	237	10	concanavalin a biological_unit
<input type="checkbox"/>	<a href="#">1qly</a>	2.1	3.8	86	146	7	cd40 ligand fragment
<input type="checkbox"/>	<a href="#">1rcy</a>	2.1	3.7	81	151	8	rusticyanin biological_unit
<input type="checkbox"/>	<a href="#">1ciy</a>	2.0	4.4	103	577	6	cryia(a)
<input type="checkbox"/>	<a href="#">1bv8</a>	2.0	3.8	93	137	9	alpha-2-macroglobulin fragment
<input type="checkbox"/>	<a href="#">2cbp</a>	2.0	3.4	65	96	6	cucumber basic protein
<input type="checkbox"/>	<a href="#">2tbvA</a>	2.0	3.5	76	283	4	Tomato bushy stunt virus

return to [FSSP home page](#) / [Dali Domain Dictionary](#)

(C) L. Holm, EMBL-EBI, Hinxton, May 1996

Figure 9.11. Continued.

Continues on next page

proteins that do not share any sequence similarity or evolutionary relationship. Thus, structural similarity may or may not be an indicator of an evolutionary relationship. Further light may be shed on this question by a close examination of the similarity. The similarity may be quite simple, such as a common arrangement and spacing of several secondary structural elements. Alternatively, there may be a highly significant alignment of many of the proteins through the same sequence of secondary structures and loops, and many of the atoms in the two proteins may be quite superimposable. Such structural closeness may be an indication of a possible evolutionary relationship. The results of a search for remote sequence similarity by sensitive statistical methods (Gibbs sampling, expectation maximization methods, and Bayesian alignment methods discussed in Chapter 4) may be found to provide further support for such a possibility. The ability to make such comparisons has depended on the development and availability of fast and efficient methods for performing structural comparisons.

Structural comparison methods share some of the features of methods for comparing sequences, but with additional considerations. For comparing two sequences, one searches for a row of amino acids in one sequence that matches a row in the second, allowing for substitutions and the insertion of gaps in one sequence to make up for extra characters in the other. For comparing structures, positions of atoms in two three-dimensional structures are compared. These methods initially examine the positions of secondary structural elements,  $\alpha$  helices and  $\beta$  strands, within a protein domain to determine whether or not

## B. FSSP: structural neighbours of 1tupA

Please cite: L. Holm and C. Sander (1996) *Science* 273(5275):595-60.

Structural alignment by Dali

Notation: Uppercase: structurally equivalent with 1tupA; lowercase: structurally non-equivalent with 1tupA

Identities computed with respect to sequence: (1) 1tupA

Colored by: identity+property

```

1 [ . . . . . : . . . . . 60
1 1tupA 100.0% SSSVPS-----
2 2tbvA  2.9% gvtvtshreyltqvnssgfvvnggigvnsiqlnpsngtlfswlpalasnfdqysfnsvv

61 . . . . . 1 . . . . . 120
1 1tupA 100.0% -----
2 2tbvA  2.9% ldyvplcgttevgvrvalyfdkdsqdepadrvelanfgvketapwaeamlriptdkvkr

121 . . . . . : . . . . . 180
1 1tupA 100.0% -----
2 2tbvA  2.9% ycn dsatvdqklidlgqlgiatyggagadavgelflarsvtlyfpqptntlkrldltgsl

181 . . . . . 2 . . . . . 240
1 1tupA 100.0% QKTYQGSYGFRLLGFLHSGTAKSVTCTYSPALNKMFCQLAKTCPVQLWVDSTPPPGRVRA
2 2tbvA  2.9% ADATGP---GYLV-----lTRTPT---VLTHTFRA-----tgTFNLS

241 . . . . . 3 300
1 1tupA 100.0% MAIYKQSQHMTVEVRRCPHHERCSDSDGLAPPQHLIRVEGNL----RVEYLDDRNTFR-H
2 2tbvA  2.9% GGL-----rcltSLTLGATgavviNDILAIidnvgtasD

301 . . . . . : . . . . . 360
1 1tupA 100.0% SVVVPYEPPEVGSDCCTTIHNYMNCSSCMGMNRRPILTIITLEDSSGNLLGRNSFEVRV
2 2tbvA  2.9% YFLNCTVSS----LPATVTFVSG-----vAAGILLVGRARANvvnll-----

361 . . . . . ] 375
1 1tupA 100.0% CACPGRDRRTEEENL
2 2tbvA  2.9% -----

```

mview 1.16 Copyright (c) Nigel P. Brown, EMBL-EBI 1997.

return to [FSSP home page](#) / [Dali Domain Dictionary](#)

(C) L. Holm, EMBL-EBI, Hinxton, May 1996

### Figure 9.11. Continued.

(B) Two representative structures that can be aligned with 1tupA with a high level of significance. Amino acid colors reflect side-chain chemistry and use the multiple alignment display program of N.P. Brown (Brown et al. 1998), which can be obtained from the author (see FSSP Web site). The aligned amino acids represent a structural alignment obtained with program DALI, not a sequence alignment. The capitalized amino acids match 1tupA structurally; lowercase amino acids do not match. Note that the percent sequence identity between the p53 sequence of 1tupA and the other two proteins is quite low at 11% for chitinase A (structure 1ctn) and 15% for sialidase-neuraminidase (structure 1eut).



## c. FSSP Fold Tree

### The FSSP database

The FSSP database includes all protein chains from the Protein Data Bank which are longer than 30 residues. The chains are divided into a **representative set** and **sequence homologs** of structures in the representative set. Sequence homologs have more than 25 % sequence identity, and the representative set contains no pair of such sequence homologs. An all-against-all structure comparison is performed on the representative set. The resulting alignments are reported in the FSSP entries for individual chains. In addition, FSSP entries include the structure alignments of the search structure with its sequence homologs.

### Reference

L. Holm and C. Sander (1998) Touring protein fold space with Dali/FSSP. *Nucl. Acids Res.* 26, 316-319.

### Availability

Free academic use. No commercial use. No incorporation into other databases.

### This table

is a fold classification of the representative set. A hierarchical clustering method is used to construct a tree based on the structural similarities from the all-against-all comparison. Family indices are constructed by cutting the tree at levels of 2, 3, 4, 5, 10 and 15 standard deviations above database average.

### Related tables

PROTEIN INDEX is sorted according to PDB codes. See the accompanying README file for additional information.

### Hyperlinks

Click on Family index for a summary of aligned pairs. Click on PDB-code for the complete FSSP entry. Click on alignment to view the structural alignments.

Family index	PDB-code	Alignments	compound
1.1.1.1.1.1	<a href="#">1af5</a>	<a href="#">alignment</a>	"i-crei (DNA endonuclease i-crei) Mutant"
2.1.1.1.1.1	<a href="#">1iba</a>	<a href="#">alignment</a>	"glucose permease fragment"
3.1.1.1.1.1	<a href="#">1aie</a>	<a href="#">alignment</a>	"p53 fragment"
4.1.1.1.1.1	<a href="#">1bba</a>	<a href="#">alignment</a>	"Bovine pancreatic polypeptide (bpp) (NMR, mean struct
4.1.2.1.1.1	<a href="#">1ppt</a>	<a href="#">alignment</a>	"Avian pancreatic polypeptide"
5.1.1.1.1.1	<a href="#">1emn</a>	<a href="#">alignment</a>	"fibrillin fragment"
6.1.1.1.1.1	<a href="#">1hcaB</a>	<a href="#">alignment</a>	"Blood coagulation factor xa"
7.1.1.1.1.1	<a href="#">1pft</a>	<a href="#">alignment</a>	"tfiib fragment (pftfiibn)"
8.1.1.1.1.1	<a href="#">1avp</a>	<a href="#">alignment</a>	"RNA polymerase ii fragment"
8.1.1.2.1.1	<a href="#">1tfti</a>	<a href="#">alignment</a>	"Transcriptional elongation factor sii (tfiis, nucleic
9.1.1.1.1.1	<a href="#">1baf</a>	<a href="#">alignment</a>	"stat-4 fragment"

Figure 9.11. *Continued.*

*Continues on next page*

115.1.1.1.1.1	<u>2plc</u>	<u>alignment</u>	"phosphatidylinositol-specific phospholipase c (pi-plc
115.1.1.2.1.1	<u>1uroA</u>	<u>alignment</u>	"uroporphyrinogen decarboxylase (uro-d, urod) biologic
115.1.1.2.1.2	<u>1a0cA</u>	<u>alignment</u>	"xylose isomerase (glucose isomerase) biological_unit'
115.1.1.2.1.2	<u>4xis</u>	<u>alignment</u>	"Xylose isomerase complex with xylose and mnCl2"
115.1.1.2.2.1	<u>1a0A</u>	<u>alignment</u>	"1,3-1,4-beta-glucanase (1,3-1,4-beta-d-glucan 4-gluc
115.1.1.2.2.1	<u>1bqA</u>	<u>alignment</u>	"beta-galactosidase"
115.1.1.2.2.1	<u>1bhqA</u>	<u>alignment</u>	"beta-glucuronidase (gus gene product) biological_unit
115.1.1.2.2.1	<u>1ceo</u>	<u>alignment</u>	"cellulase celc (1,4-beta-d-glucan-glucanohydrolase, e
115.1.1.2.2.1	<u>1eceA</u>	<u>alignment</u>	"endocellulase e1 fragment (endo-1,4-beta-d-glucanase;
115.1.1.2.2.1	<u>1eda</u>	<u>alignment</u>	"endoglucanase a fragment (endo-(1,4)-beta-glucanase,
115.1.1.2.2.1	<u>1gwaA</u>	<u>alignment</u>	"beta-glycosidase biological_unit"
115.1.1.2.2.1	<u>2myr</u>	<u>alignment</u>	"myrosinase (thioglucoiside glucohydrolase) biological_
115.1.1.2.2.2	<u>1byb</u>	<u>alignment</u>	"Beta-amylase reacted with 200 mm maltose and comple
115.1.1.2.2.2	<u>1xvzA</u>	<u>alignment</u>	"1,4-beta-d-xylan-xylanohydrolase (endo-1,4-beta-xylc
115.1.1.2.2.3	<u>1aba</u>	<u>alignment</u>	"chitobiase (beta-n-acetylhexosaminidase, n-acetyl-bet
115.1.1.2.2.4	<u>1cnv</u>	<u>alignment</u>	"concanavalin b"
115.1.1.2.2.4	<u>1ctn</u>	<u>alignment</u>	"Chitinase a (ph 5.5, 4 degrees c)"
115.1.1.2.2.4	<u>1nar</u>	<u>alignment</u>	"Narbonin"
115.1.1.2.2.4	<u>2ebn</u>	<u>alignment</u>	"Endo-beta-n-acetylglucosaminidase f1 (endoglycosidase
115.1.1.2.3.1	<u>1onrA</u>	<u>alignment</u>	"transaldolase b"
115.1.1.2.4.1	<u>1nsj</u>	<u>alignment</u>	"phosphoribosyl anthranilate isomerase (prai)"
115.1.1.2.4.2	<u>1igs</u>	<u>alignment</u>	"indole-3-glycerolphosphate synthase (igps)"
115.1.1.2.4.2	<u>1pii</u>	<u>alignment</u>	"N-(5'phosphoribosyl)anthranilate isomerase complex wi
115.1.1.2.4.2	<u>2tysA</u>	<u>alignment</u>	"tryptophan synthase Mutant biological_unit"
115.1.1.2.4.3	<u>1aj2</u>	<u>alignment</u>	"dihydropteroate synthase (dhps) biological_unit"
115.1.1.2.4.4	<u>1aw5</u>	<u>alignment</u>	"5-aminolevulinatase dehydratase (porphobilinogen synthc
-			
407.1.1.1.1.1	<u>1hev</u>	<u>alignment</u>	"Hevein (NMR, 6 structures)"
407.1.1.1.2.1	<u>9waaA</u>	<u>alignment</u>	"Wheat germ agglutinin (isolectin 2)"

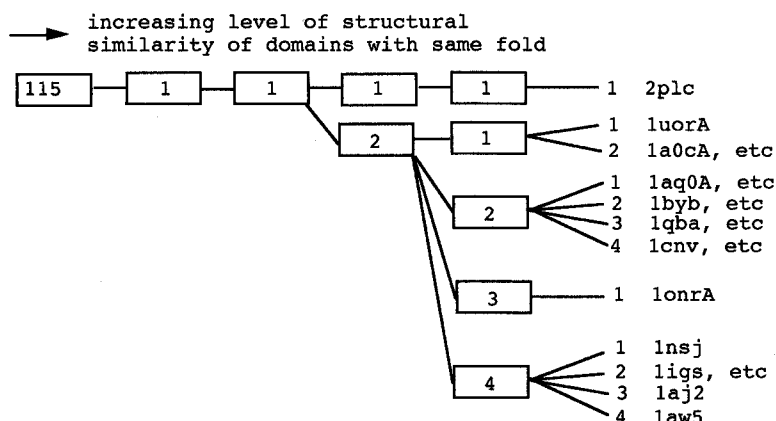


Figure 9.11. Continued.

(C) Hierarchical clustering of folds and domains. All of the current folds represented by domain alignments in FSSP have been organized into a dendrogram that indicates the relationships among them. The dendrogram for fold 115 is first illustrated, and a tabular representation is then shown. Domains are identified by the PDB file from which they were derived. If only one of several domains is represented by the fold, the domain is identified by the PDB file name plus a letter code; e.g., 1uorA is a domain of the structure 1uor. Domains that are grouped on the right are the most structurally alike and give a high statistical score for an alignment of the representative fold (a certain combination of secondary structures in space and their connections) when they are aligned with the DALI program. Although these domains have very little sequence similarity, their very close structural similarity suggests that they could possibly be homologous and represent a superfamily. Domains that are joined in deeper branches of the dendrogram, e.g., 1uorA and 1aq0A, are less structurally alike, and the score for their alignment is lower. Although domain 1pic has the same fold as the rest of the domains, its atoms align the least well with the other domains. The structures and alignments represented can be viewed by the links on the Web page. The page is accessible from the main page of the FSSP database.

### D. FSSP: family alignment around 1tupA

Please cite: L. Holm and C. Sander (1996) *Science* 273(5275):595-60.

Structures aligned by Dali with sequence neighbours from HSSP

Notation: parent structure[:domain-identifier]Swissprot-identifier; uppercase: structurally equivalent to 1tupA; lowercase: bounds sequence insertion; - deletion from HSSP; ~ structurally nonequivalent to 1tupA

Identities computed with respect to sequence: (1) 1tupA  
Colored by: identity+property

1 1tupA	100.0%	1 [	SSSVPSQKTYQGSYGFRLLGFLHSGTAKSVTCTYSPALNKMFQQLAKTQPVQLWVDSTPPP	60
2 1tupAlp53_human	100.0%		SSSVPSQKTYQGSYGFRLLGFLHSGTAKSVTCTYSPALNKMFQQLAKTQPVQLWVDSTPPP	
3 1tupAlp53_macmu	97.4%		SSSVPSQKTYHGSYGFRLLGFLHSGTAKSVTCTYSPDLNKMFQQLAKTQPVQLWVDSTPPP	
4 1tupAlp53_cerae	97.4%		SSSVPSQKTYHGSYGFRLLGFLHSGTAKSVTCTYSPDLNKMFQQLAKTQPVQLWVDSTPPP	
44 1tupAlp53_oryla	59.4%		~TTVPVTTDYPGSEYELERFQKSGTAKSVTSTYSETLNKLYCQLAKTSPIEVRVSKEPPK	
45 1tupAlp53_plafe	58.0%		SSTVPVVTDYPGEYGFQLRFQKSGTAKSVTSTFSELLKLYCQLAKTSPVEVLLSKEPPQ	
1 1tupA	100.0%	61	GTRVRAMAIYKQSQHMTEVVRRP PHERCSDSDGLAPPQHLIRVEGNL RVEYLLDRNTFR	120
2 1tupAlp53_human	100.0%		GTRVRAMAIYKQSQHMTEVVRRP PHERCSDSDGLAPPQHLIRVEGNL RVEYLLDRNTFR	
3 1tupAlp53_macmu	97.4%		GSRVRAMAIYKQSQHMTEVVRRP PHERCSDSDGLAPPQHLIRVEGNL RVEYSDDRNTFR	
4 1tupAlp53_cerae	97.4%		GSRVRAMAIYKQSQHMTEVVRRP PHERCSDSDGLAPPQHLIRVEGNL RVEYSDDRNTFR	
44 1tupAlp53_oryla	59.4%		GAILRATAVYKKT EHVADVRRP PHHQN---EDSVEHRSHLIRVEGSQLAQYFEDPYTKR	
45 1tupAlp53_plafe	58.0%		GAVLRATAVYKKT EHVADVRRP PHHQT---EDTAEHRSHLIRLEGSQRALYFEDPHTKR	
1 1tupA	100.0%	121	HSVVVPYEPPEVGSDDTTIHYNMNCSSMGGMNR RPILTIITLEDSSGNLLGRNSFEVR	180
2 1tupAlp53_human	100.0%		HSVVVPYEPPEVGSDDTTIHYNMNCSSMGGMNR RPILTIITLEDSSGNLLGRNSFEVR	
3 1tupAlp53_macmu	97.4%		HSVVVPYEPPEVGSDDTTIHYNMNCSSMGGMNR RPILTIITLEDSSGNLLGRNSFEVR	
4 1tupAlp53_cerae	97.4%		HSVVVPYEPPEVGSDDTTIHYNMNCSSMGGMNR RPILTIITLEDSSGNLLGRNSFEVR	
44 1tupAlp53_oryla	59.4%		QSVTVPYEPPQPGSEM TILLSYMNCSSMGGMNR RPILTILTLET-EGLVLGRRCFEVR	
45 1tupAlp53_plafe	58.0%		QSVTVPYEPPQLGSETTA ILLSFMNCSSMGGMNR RQILTILTLET PDGLVLGRRCFEVR	
1 1tupA	100.0%	181	VCAQ PGRDRRTEEEENL	196
2 1tupAlp53_human	100.0%		VCAQ PGRDRRTEEEENL	
3 1tupAlp53_macmu	97.4%		VCAQ PGRDRRTEEEENF	
4 1tupAlp53_cerae	97.4%		VCAQ PGRDRRTEEEENF	
44 1tupAlp53_oryla	59.4%		ICAQ PGRDRKTEEES~	
45 1tupAlp53_plafe	58.0%		VCAQ PGRDRKTEEES~	

mview 1.16 Copyright (c) Nigel P. Brown, EMBL-EBI 1997.

return to [FSSP home page](#) / [Dali Domain Dictionary](#)

(C) L. Holm, EMBL-EBI, Hinxton, May 1996

Figure 9.11. Continued.

(D) Structural alignment of 1tupA with other protein sequences in SwissProt that are similar in sequence to p53. The information on matching sequences is stored in the HSSP database described in the text. Shown on each row of the alignment are the PDB structure identification, matched SwissProt sequence, percent sequence identity between the sequence of the structural entry and the SwissProt sequence, and the multiple sequence alignment of the sequence with the other matching SwissProt sequences based on a structural alignment. This alignment reveals which amino acid residues in these proteins are predicted to occupy the same structural position. Sequence notations are indicated on the page. Only a portion of the alignment is shown.

VAST Structure Neighbors

## Structures similar to MMDB 2890, 2REB domain 1

Rec A Protein (E.C.3.4.99.37)

 View / Save Alignments

 New

[Get Cn3D 2.0 Now!](#)

## Options:

- Launch Viewer  
 See File  
 Save File

## Viewer:

- Cn3D v2.0 (asn.1)  
 Mage (Kinemage)  
 (PDB)

## Complexity:

- Aligned Chains only     Alpha Carbons only  
 All Chains     All Atoms

	PDB	C	D	RMSD	NRES	%Id	Description
<input type="checkbox"/>	1REA	1		0.3	275	100.0	Rec A Protein (E.C.3.4.99.37) Complex With Adenosine Diphosphate (Rec A-Adp)
<input type="checkbox"/>	1SKY	E	5	3.1	160	14.4	Crystal Structure Of The Nucleotide Free Alpha3beta3 Sub-Complex Of F1-AtPase From The Thermophilic Bacillus Ps3
<input type="checkbox"/>	1THM			4.1	99	9.1	Thermitase (E.C.3.4.21.66)
<input type="checkbox"/>	1UAA	B	5	2.5	93	17.2	Structure Of The Rep Helicase-Single Stranded Dna Complex At 3.0 Angstroms Resolution
<input type="checkbox"/>	1UAG	2		4.5	81	8.6	Udp-N-Acetylmuramoyl-L-Alanine:d-Glutamate Ligase
<input type="checkbox"/>	1POX	A	3	3.1	106	4.7	Pyruvate Oxidase (E.C.1.2.3.3) Mutant With Pro 178 Replaced By Ser, Ser 188 Replaced By Asn, And Ala 458 Replaced By Val (P178s,S188n,A458v)
<input type="checkbox"/>	1CL1	A	1	4.2	106	8.5	Cystathionine Beta-Lyase (Cbl) From Escherichia Coli
<input type="checkbox"/>	1AST	1		2.7	114	14.9	Crystal Structure Of The Delta Prime Subunit Of The Clamp-Loader Complex Of Escherichia Coli Dna Polymerase Iii
<input type="checkbox"/>	1ETS	2		4.3	106	16.0	Signal Recognition Particle Receptor From E. Coli
<input type="checkbox"/>	1GPL	1		3.4	79	8.9	Rp2 Lipase
<input type="checkbox"/>	1AAT	1		3.1	88	9.1	Cytosolic Aspartate Aminotransferase (E.C.2.6.1.1) Complex With 2-Oxo-Glutamic Acid
<input type="checkbox"/>	1RLA	B		3.8	97	6.2	Three-Dimensional Structure Of Rat Liver Arginase, The Binuclear Manganese Metalloenzyme Of The Urea Cycle
<input type="checkbox"/>	2TPL	A	3	3.8	86	8.1	Tyrosine Phenol-Lyase From Citrobacter Intermedius Complex With 3-(4'-Hydroxyphenyl)propionic Acid, Pyridoxal-5'-Phosphate And Cs+ Ion
<input type="checkbox"/>	2DRI	1		3.2	72	12.5	D-Ribose-Binding Protein Complexed With Beta-D-Ribose
<input type="checkbox"/>	8ABP	1		3.1	72	11.1	L-Arabinose-Binding Protein (Mutant With Met 108 Replaced By Leu) (M108L) Complex With D-Galactose
<input type="checkbox"/>	1CYD	D		4.4	97	8.3	Carbonyl Reductase Complexed With Nadph And 2-Propanol
<input type="checkbox"/>	1N2C	E		2.8	80	16.2	Nitrogenase Complex From Azotobacter Vinelandii Stabilized By

**Figure 9.12.** Example of searching for structural neighbors identified by the VAST algorithm. Shown is the result of a search for neighbors to chain 1 of the *E. coli* RecA protein structure (PDB identifier 2reb). If the rightmost column box next to any of the listed structural neighbors and the view/save structures box are sequentially checked, then an overlay view of the structures is provided by ENTREZ for viewing by Cn3d or Mage. In the output table of structural neighbors, PDB is a four-character PDB-identifier of the structural neighbor, C is the PDB chain name, D is the MMDB domain identifier, RMSD is the root mean square deviation in Angstroms between the superimposed atoms, NRES is the number of equivalent pairs of  $C_{\alpha}$  atoms super-

					Adp-Tetrafluoroaluminate
<input type="checkbox"/>	<a href="#">1NGS A 2</a>	2.7	87	12.6	Complex Of Transketolase With Thiamin Diphosphate, Ca2+ And Acceptor Substrate Erythrose-4-Phosphate
<input type="checkbox"/>	<a href="#">1GDH A 1</a>	3.0	65	7.7	D-Glycerate Dehydrogenase (Apo Form) (E.C.1.1.1.29)
<input type="checkbox"/>	<a href="#">1A4S A 2</a>	3.0	70	11.4	Betaine Aldehyde Dehydrogenase From Cod Liver
<input type="checkbox"/>	<a href="#">1AK1 2</a>	3.2	70	5.7	Ferrochelatae From Bacillus Subtilis
<input type="checkbox"/>	<a href="#">1ZIN 1</a>	2.9	80	13.8	Adenylate Kinase With Bound Ap5a
<input type="checkbox"/>	<a href="#">1MIO D 13</a>	3.0	59	5.1	Nitrogenase Molybdenum-Iron Protein
<input type="checkbox"/>	<a href="#">1RRE</a>	3.5	52	9.6	Non-Myristoylated Rat Adp-Ribosylation Factor-1 Complexed With Gdp, Monomeric Crystal Form
<input type="checkbox"/>	<a href="#">1RVV 1</a>	2.1	63	14.3	SynthaseRIBOFLAVIN SYNTHASE COMPLEX OF BACILLUS SUBTILIS
<input type="checkbox"/>	<a href="#">1CEY</a>	2.5	58	10.3	Chey Complexed With Magnesium (Nmr, 46 Structures)
<input type="checkbox"/>	<a href="#">1DEK A 1</a>	2.7	77	13.0	Deoxynucleoside Monophosphate Kinase Complexed With Deoxy-Gmp
<input type="checkbox"/>	<a href="#">1MIO D 11</a>	3.0	70	4.3	Nitrogenase Molybdenum-Iron Protein
<input type="checkbox"/>	<a href="#">1RAA C 5</a>	2.1	45	17.8	Aspartate Transcarbamoylase (E.C.2.1.3.2) (Aspartate Carbamoyltransferase) (T State) Complexed With Ctp (Fast Cooling Sa Refinement After 250 Steps Of Equilibration Of Preliminary Refined Model)
<input type="checkbox"/>	<a href="#">1BNC A 1</a>	2.8	54	11.1	Mol_id: 1; Molecule: Biotin Carboxylase; Chain: A, B; Ec: 6.3.4.14
<input type="checkbox"/>	<a href="#">1UAG 3</a>	3.0	62	14.5	Udp-N-Acetylmuramoyl-L-Alanine:d-Glutamate Ligase
<input type="checkbox"/>	<a href="#">1MIO C 9</a>	2.3	52	0.0	Nitrogenase Molybdenum-Iron Protein
<input type="checkbox"/>	<a href="#">1ITB A</a>	2.2	41	14.6	Crystal Structure Of Iibcellobiose From Escherichia Coli
<input type="checkbox"/>	<a href="#">1IOW 1</a>	2.0	36	5.6	Complex Of Y216f D-Ala:d-Ala Ligase With Adp And A Phosphoryl Phosphinate
<input type="checkbox"/>	<a href="#">1BPL A 1</a>	2.5	60	6.7	Glycosyltransferase
<input type="checkbox"/>	<a href="#">1XAN 3</a>	1.7	30	13.3	Human Glutathione Reductase In Complex With A Xanthene Inhibitor

**Display / Sort Hits**
**Display Subset:**

- Non-redundant; BLAST p-value 10e-7
- Non-redundant; BLAST p-value 10e-40
- Non-redundant; BLAST p-value 10e-80
- Non-identical sequences
- All of MMDB

**Sorted by:**

- VAST Score
- VAST P-value
- Rmsd
- Aligned residues
- Identities

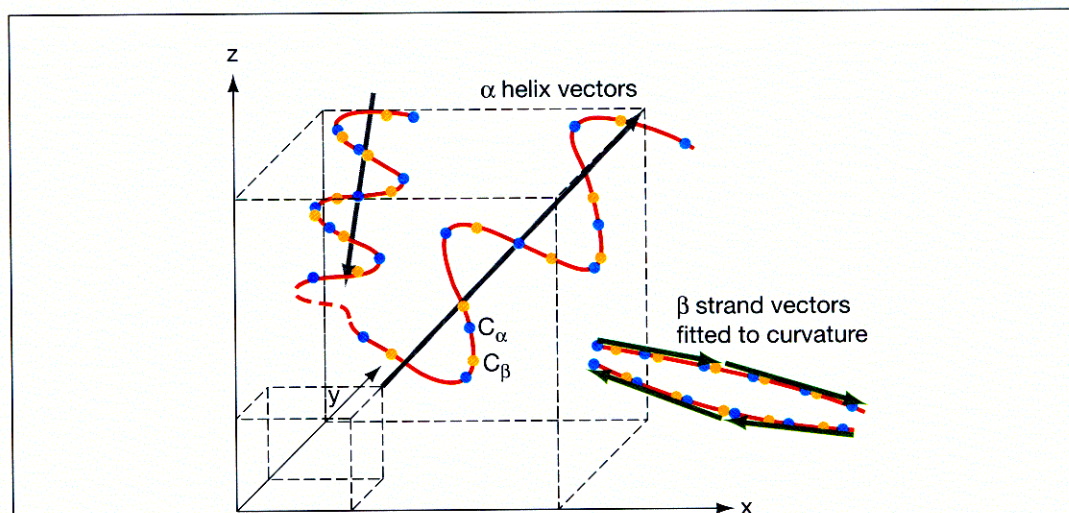
**Column Format:**

- RMSD, NRES, %Id
- All values

imposed between the two structures, %Id is the percent identical residues in the aligned sequence region, and description of the neighbor is taken from the PDB database entry. The options in the lower part of the page influence the number of matches reported in the table, and this number may be varied by mouse-clicking the "display subset" box. The MMDB database is organized into groups based on sequence similarity, and only a representative member of each group is included. Groups are based on extensive BLAST searches for sequence similarity followed by clustering by a neighbor-joining procedure (see Chapter 6). Several different levels of clustering based on different ranges of BLAST scores are shown. The lower the score chosen, the more group members are reported. Note that the format of a structural entry in the MMDB database is different from that in the PDB database and requires visualization by the Cn3d viewer.

the number, type, and relative positions of these elements are similar or if the proteins have a similar architecture. Distances between the  $C_\alpha$  or  $C_\beta$  atoms within these structures are then examined in detail to determine the degree to which the structures may be superimposed. If a few elements can be aligned and are joined by a similar arrangement of loops, the proteins share a common fold. As the arrangement, joining, and alignment of secondary structural elements within the proteins increase, the degree of structural similarity between the proteins becomes more and more convincing and significant.

To specify a three-dimensional structure, positions of molecules are expressed as  $x$ ,  $y$ , and  $z$  cartesian coordinates within a fixed frame of reference, as shown in Figure 9.13. The direction of the bond angles and the interatomic distances between amino acids along the polypeptide chain may also be represented as a vector. Secondary structures can also be represented by a vector that starts at the beginning of the secondary structural element, extends for the length of the element, and has a direction that reveals the orientation of the element in the overall structure. Comparison of these structural representations in two proteins provides a framework for comparing the structures of the proteins. In many structural comparison methods, distances between  $C_\alpha$  or between  $C_\beta$  atoms in two protein structures are used for comparison purposes. A more detailed comparison of the structures can be made by adding information on side chains such as the amount of outside area of the side chain



**Figure 9.13.** Alignment of the three-dimensional structure of proteins by their secondary structures. Representation of arrangement of secondary structures in three-dimensional space is shown on a two-dimensional projection. In the structural alignment programs VAST and SARF, the atoms of each secondary structural element in each protein are replaced by a vector of position, length, and direction determined by the positions of the  $C_\alpha$  or  $C_\beta$  atoms along the element. Shown are projections of two  $\alpha$  helices and two  $\beta$  strands and their vector representations as gray and green arrows, respectively, from a common  $x$ ,  $y$ ,  $z$  cartesian coordinate system. The three-dimensional cartesian coordinates of the start and end of one  $\alpha$ -helical vector are diagrammed as wide dashed lines. Only these two sets of coordinates are needed to specify the location of the vector, whereas many such sets are required to locate the  $C_\alpha$  or  $C_\beta$  atoms in the corresponding  $\alpha$  helix. An element that is curved is approximated by two or more sequential vectors, as depicted for the two  $\beta$  strands, which are bent due to the twist of their composite  $\beta$  sheet. The joining of the helices by a short loop is also recognized by the algorithm. The vector representations of two proteins are then compared. If the type and arrangement of the elements are similar in two proteins within a reasonable margin of error and level of significance, the three-dimensional structures of the proteins are predicted to be similar.

that is buried under other molecules so that the chain is not accessible to water molecules. Distances and bond angles to other atoms in the structure may also be compared. Several of the parameters used for structural comparisons may also be used to classify the environment of a particular amino acid, e.g., a buried, hydrophobic amino acid in a  $\beta$  strand.

There are two reasons that it is more difficult to align structures than sequences. First, a similar structure may form by many different foldings of the amino acid  $C_\alpha$  backbone. As a result, matched regions may not necessarily be in the same order in the two proteins so that two matching segments are often separated by unmatched segments. Second, although the local environments of many molecules in two proteins may be similar, there may also be some local differences. For example, central positions, but not the ends, of secondary structures in two proteins may match closely. For this reason, structural alignment methods often smooth out the comparisons by comparing several molecules at the same time and choosing an average result.

Structural biologists have been working on the problem of finding similar structural features in proteins for a long time, and a variety of methods have been devised for performing comparisons of protein structures (for review, see Blundell and Johnson 1993; Holm and Sander 1994, 1996; Alexandrov and Fischer 1996; Gibrat et al. 1996; Orengo and Taylor 1996). A complete discussion of this subject is beyond the scope of this text. Programs publicly accessible at Web sites, SSAP and DALI, and two programs that utilize a fast search for common arrangements of secondary structures, VAST and SARE, are described below.

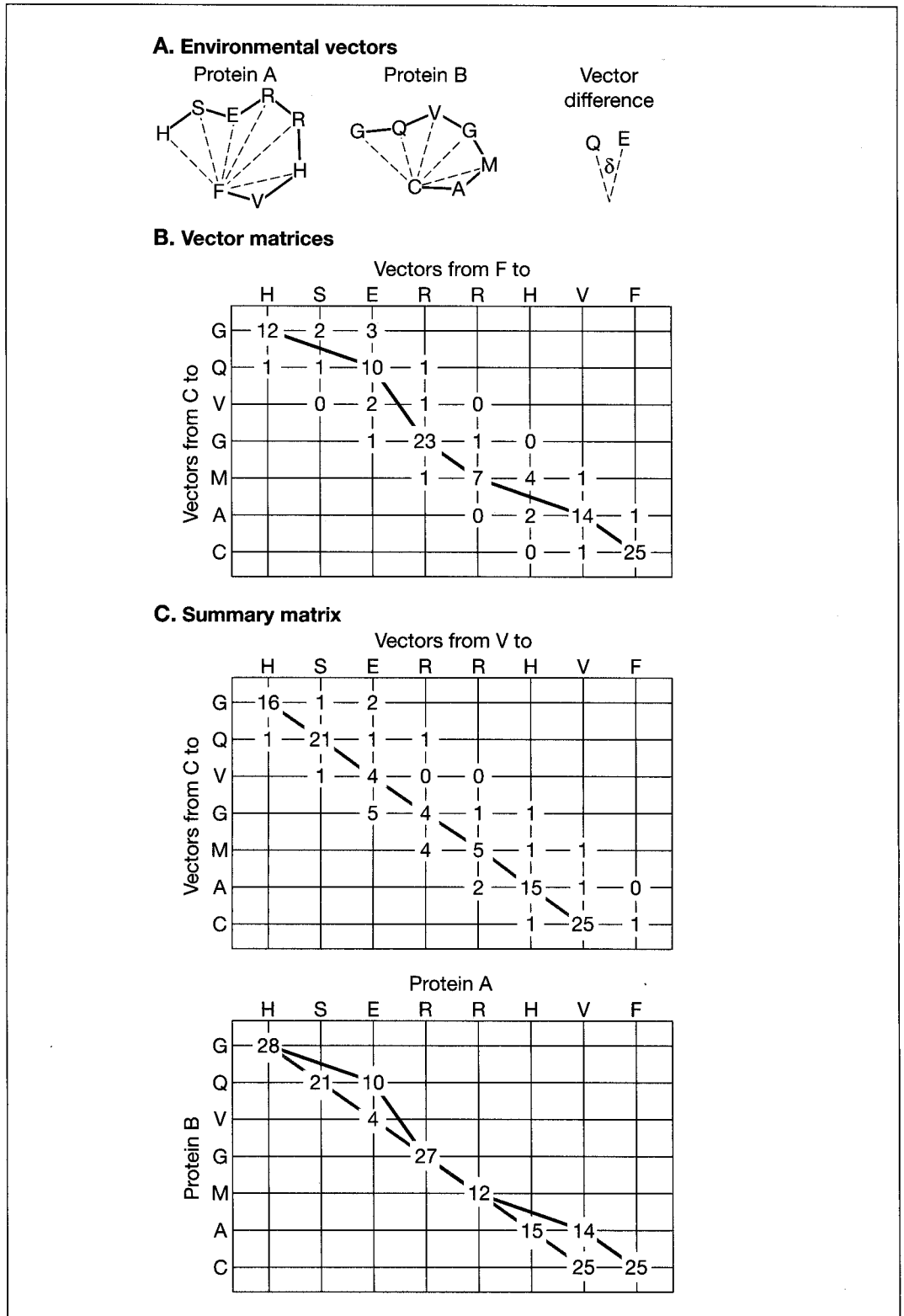
## Dynamic Programming

Algorithms like those used for sequence alignment have also been used for aligning structures. For aligning sequences, the object is to bring as many identical or similar sequence characters into vertical register in the alignment with a minimum cost of insertions and deletions. For aligning structures, the local environment of each amino acid expressed in interatomic distances, bond angles, or R group is given a coded value or vector representation that reflects the environment of that amino acid. Alternatively, a scoring matrix much like the amino acid scoring used for sequence alignments may be made. For protein structures, each sequential column in the scoring matrix gives a score for the fit of any of the 20 amino acids to a single position in the structure (more on matrices below). An optimal alignment between these sets of values by dynamic programming is then found.

The alignment program SSAP (*secondary structure alignment program*) uses a method called double dynamic programming to produce a structural alignment between two proteins (Taylor and Orengo 1989; Orengo et al. 1993; Orengo and Taylor 1996). A local structural environment is independently defined for each residue in each sequence, and the method then matches residues by comparing these structural environments. The environment assigned to each amino acid takes into account the degree of burial in the hydrophobic core and type of secondary structure. As in sequence alignment by dynamic programming, a scoring matrix is derived and the highest-scoring regions in this matrix define the optimal structural alignment of the two proteins. One of the environmental variables that is used is a representation of the geometry of the protein by drawing a series of vectors from the  $C_\beta$  atoms of an amino acid to the  $C_\beta$  atoms of all of the other amino acids in the protein. If the resulting geometric views in two protein structures are similar, the structures must also be similar. The double dynamic programming method of aligning structures using  $C_\beta$  vectors is illustrated in Figure 9.14.

Because each sequential pair of amino acids is compared, an alignment will be possible only if the two protein chains follow the same approximate conformational changes throughout their lengths. If the proteins follow the same changes along some of their lengths, then

diverge, then return again, it is difficult to align them through the divergent region by the above method, as described. The problem is similar to trying to choose a gap penalty for sequence alignments, but in the structural case, many kinds of rearrangements are possible.





Another version of SSAP (SSAP1) has been developed for identifying conserved folds/motifs, and this method circumvents the above alignment problem. This program uses all of the vector matrix values in the summary matrix and then uses a local alignment version of the dynamic programming algorithm to locate the most alike regions in the structures. The algorithm has been greatly speeded up by comparing only pairs of amino acids with similar torsional angles ( $\Phi$  and  $\Psi$ ) and extent of residue burial/lack of water accessibility. SSAP is used to cluster proteins in the CATH database in a fully automated manner (Orengo et al. 1997).

## Distance Matrix

The distance method uses a graphic procedure very similar to a dot matrix to identify the atoms that lie most closely together in the three-dimensional structure. If two proteins have a similar structure, the graphs of these structures will be superimposable. Distances between  $C_\alpha$  atoms along the polypeptide chain and between  $C_\alpha$  atoms within the protein structure can be compared by a two-dimensional matrix representation of the structure, as shown in Figure 9.15. Instead of aligning environmental variables of each successive amino acid in

**Figure 9.14.** The double dynamic programming method for structural alignment. (A) Vectors from the  $C_\beta$  atom of one amino acid to a set of other nearby amino acids in each of two protein segments are shown as two-dimensional projections. These vectors are given the same coordinate axes. Hence, one vector may be subtracted from the other to compare the relative positions of the  $C_\beta$  atoms in the two protein segments, shown in A as a vector difference. The smaller the differences, the more alike the structures. In SSAP, the vectors are subtracted (the resulting difference is  $\delta$ ) and the difference added to an empirically derived number, 10. The resulting value is then divided into a second empirically derived number, 500, to give a score  $S$  for the vector difference. For example, if the vector difference is  $10^\circ$ , then  $S = 500 / (10 + 10) = 25$ . (B) Two vector matrices that represent differences between the geometric view from one amino acid position in one protein and a view for one amino acid in the second protein. The set of vectors of one protein are listed across the top of the matrix and the set for the other are listed down the right side. The matrix is then filled with scores of vector differences. For example, if the vector from F to H in protein 1 less the vector from C to G in protein 2 is  $31^\circ$ , then the score placed in the upper right corner is  $S = 500 / (31 + 10) = 12$ . The remaining difference scores are calculated in a similar manner. Although vectors to neighboring amino acids are shown in this example, vectors to immediate neighbor positions are actually not used to reduce effect of local secondary structure. An optimal alignment, shown as a red path through the matrix, is then found through the vector matrix by a global form of the dynamic programming algorithm, using a constant deletion penalty of 50. For performing a structural alignment by this method, a similar set of vector differences are determined between the next amino acid V in protein A and the amino acid in protein B, as shown in the lower matrix in B, and an optimal path (*blue*) is obtained. This procedure is repeated until vector views between all amino acid positions have been compared. Two vector matrices are shown, comparing one position in protein A to each of two positions in protein B. (C) The resulting alignments (shown as red and blue paths) and the scores on the alignment path are transferred to a summary matrix. If two optimal alignment paths cross the same matrix position, the scores of those positions in the two alignments are summed. One part of the alignment path (*black*) is found in both comparisons, thereby providing corroborative evidence of vector similarity in these regions. In the example shown, the sum of the upper right positions in the two vector matrices is  $12 + 16 = 28$ . When all of the alignments have been placed into the summary matrix, a second dynamic programming alignment is performed through this matrix. The final alignment found represents the optimal alignment between the protein structures. The logarithm of the final score is scaled such that a maximum value of 100 is possible. An adjusted score of 80 indicates a close structural relationship; one of 60–70 indicates a probable common fold. Other types of environmental variables other than the position of the  $C_\beta$  atoms in this example may also be aligned with this double dynamic programming method, as described in the text. (Adapted from an example in Orengo and Taylor 1996.)

two protein structures, the distance matrix method compares geometric relationships between the structures without regard to alignment. The sequence of the protein is listed both across the top and down the side of the matrix. Each matrix position represents the distance between the corresponding  $C_\alpha$  atoms in the three-dimensional structure. The smallest distances represent the more closely packed atoms within secondary structures and regions of tertiary structure. Positions of closest packing are marked with a dot to highlight them, much as in a dot matrix. Distance matrices are produced for each three-dimensional structure of interest. Similar groups of secondary structural elements are superimposed as closely as possible into a common core structure by minimizing the sum of the atomic distances between the aligned  $C_\alpha$  atoms. The method is outlined in Figure 9.15.

The program DALI (*d*istance *a*lignment tool) uses this method to align protein structures (Vriend and Sander 1991). The existing structures have been exhaustively compared to each other by DALI and the results organized into a database, the FSSP database, which may be accessed at <http://www2.embl-ebi.ac.uk/dali/fssp/fssp.html>. A newly found structure may be compared to the existing database of protein structures using DALI at <http://www2.embl-ebi.ac.uk/dali/>. The network version of DALI uses fast comparison methods to determine whether a new structure is similar to one already present in the FSSP database.

### **Alignment in DALI**

The assembly step of the original DALI algorithm uses a Monte Carlo simulation that performs a random search strategy for submatrices that can be aligned using the similarity score defined below as a guide. The algorithm is similar to the genetic and simulated annealing algorithms (Chapter 4) in using a probabilistic method to improve previously found alignments. There is no existing algorithm for direct alignment of two structures; such an algorithm would have to find the closest alignment of two sets of points in three-dimensional space, a very difficult problem computationally. Hence, the need for an approximate solution. Other methods for aligning structures that are described below also use simulations to find alignments. The Internet version of the DALI program utilizes more rapid search methods than those described above to compare new structures to existing structures in the FSSP database, but the overall analysis is very similar.

The similarity score for a structural alignment of two proteins by the distance method is based on the degree to which all of the matched elements can be superimposed. In the example shown in Figure 9.16, the score for a matching set of helices is the sum of the similarity scores of all of the atom pairs using a particular scheme for scoring each pair. Suppose that two helices a and b have been found to interact in protein A, and that a pair of helices a' and b' in protein B are superimposable on a and b. A certain pair of  $C_\alpha$  atoms that are very close in the model, one in helix a ( $i^A$ ) and a second in helix b ( $j^A$ ), is identified. This set will correspond to a matched pair  $i^B$  in helix a' and  $j^B$  in helix b' of protein B. If the distance between  $i^A$  and  $j^A$  is  $d_{ijA}$  and the distance between  $i^B$  and  $j^B$  is  $d_{ijB}$ , then the similarity score for this pair of atoms is derived from the fractional deviation  $|d_{ijA} - d_{ijB}| / d_{ij*}$ , where  $d_{ij*}$  is the average of  $d_{ijA}$  and  $d_{ijB}$ . If two atom pairs can be superimposed, they are given a threshold similarity score of 0.20; otherwise they are given a similarity score of the threshold less the above fractional deviation. A deviation of 0.20 will correspond to adjacent  $\beta$  strands matching to within 1 Å and to  $\alpha$  helices and helix strands matching to within 2–3 Å. As these scores are summed over all of the atoms in the match-

ing helices, the contributions of more distant atoms are down-weighted by an exponential factor to allow for bending and other distortions. The result of using this scoring system is that the similarity score for matching the two helix pairs in proteins A and B will increase in proportion to the number of superimposable atoms in the two helices. As additional matching elements are added to the structural alignment of the two proteins, the similarity scores for matching each individual pair of secondary structures are added to give a higher similarity score that reflects the full alignment of the structures.

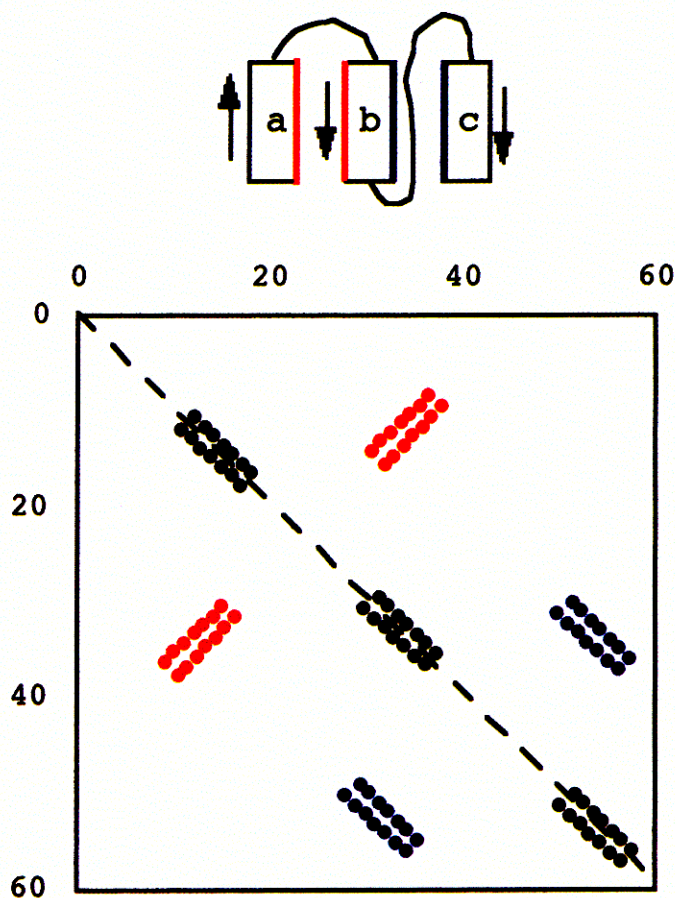
The DALI method provides one convenient method, in addition to the others described herein, to compare a new structure to existing structures in the Brookhaven structural database, and is accessible from a Web site.

### Fast Structural Similarity Search Based on Secondary Structure Analysis

One class of structural alignment methods performs a comparison of the types and arrangements of  $\alpha$  helices and  $\beta$  strands in one protein structure with the  $\alpha$  helices and  $\beta$  strands in a second structure, as well as the ways in which these elements are connected (for review, see Gibrat et al. 1996). If the elements in two structures are similarly arranged, the corresponding three-dimensional structures are also similar. Because there are relatively few secondary structural elements in proteins and the relative positions of these elements may be quite adequately described by vectors giving their position, direction, and length, vector methods provide a fast and reliable way to align structures. It is a much simpler computational problem to compare vector representations of secondary structures than to compare the positions of all of the  $C_\alpha$  or  $C_\beta$  atoms in those structures. If an element of a given type and orientation within a given tolerance level is found in the same relative position in both structures, they possess a basic level of structural similarity. Elements that do not match within the tolerance level are not considered to be structurally similar. VAST and SARF are examples of programs that are available on the Web that use this methodology (Hogue et al. 1996; Alexandrov and Fischer 1996; see <http://www.ncbi.nlm.nih.gov/Entrez> and <http://www-lmmb.mcicrf.gov/~nicka/sarf2.html/>). Vector methods do not use the structure authors' assigned secondary structures in the PDB entry, but rather use automatic methods to assign secondary structure based on the molecular coordinates of atoms on the structure. Different methods are used for defining the number and extent of secondary structural elements and for the thresholds that make up an acceptable match (Bryant and Lawrence 1993; Madej et al. 1995; Gibrat et al. 1996; Alexandrov and Fischer 1996). Until one of these methods is shown to be superior, it is advisable to try all to increase the chance of a finding a biologically important match.

Once individually aligning sets of secondary structural elements have been identified, they are clustered into larger alignment groups. For example, if three matching sets of  $\alpha$  helices have been found in two structures, a similarly oriented group of three  $\alpha$  helices must be present in the structures. The same arrangements of a small number of secondary structural elements are commonly found in protein structures, thus this method often finds new occurrences of a previously found arrangement. An arrangement with a large number of secondary elements is less common and therefore more significant. This clustering step generates a large number of possible groups of secondary structural elements from which the most likely ones must be selected. Some methods use the clusters with the largest number of secondary structures as the most significant. Other methods perform a more detailed analysis of the aligned secondary structures. For example, the atomic coordinates of an  $\alpha$  helix in one protein structure will be aligned with those of the matched  $\alpha$  helix in the second structure, and the root mean square deviation (rmsd) will be calculated. The quality of this new alignment provides an indication of which secondary structure

A.



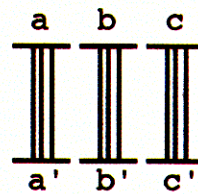
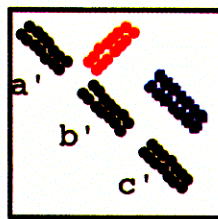
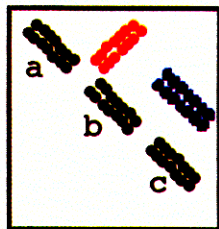
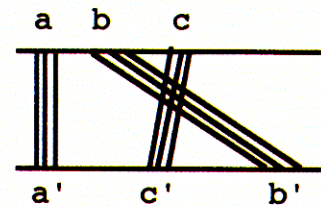
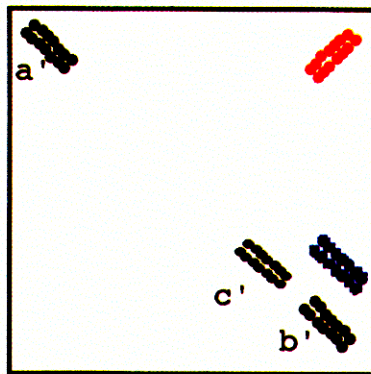
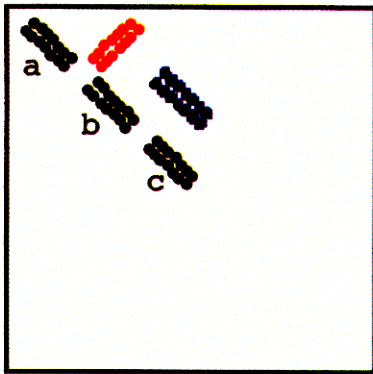
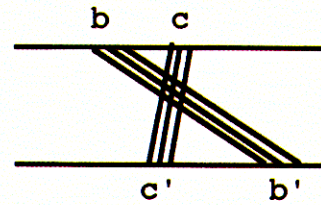
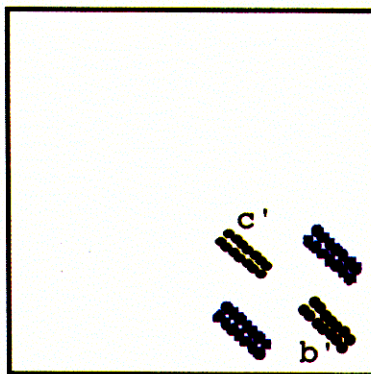
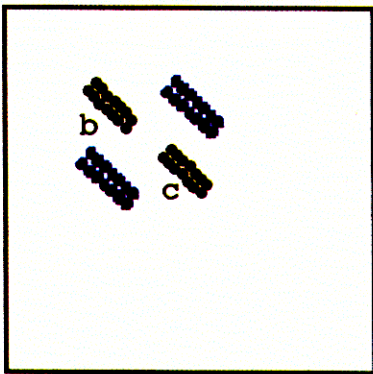
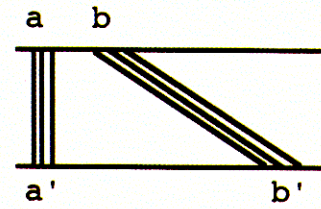
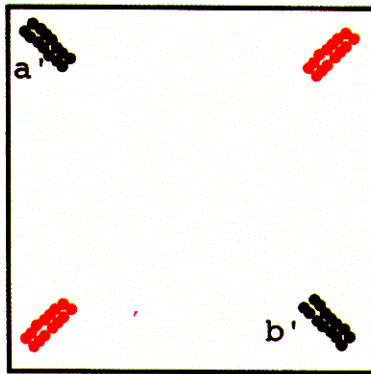
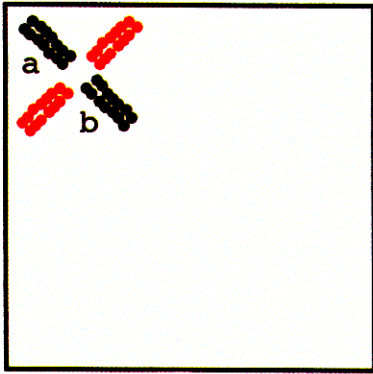
**Figure 9.15.** Distance matrix of hypothetical three-helix structure. (A) Matrix positions that represent closest distances of approximately  $<12 \text{ \AA}$  between the  $C_{\alpha}$  atoms in the known three-dimensional structures of the protein are marked by filling them with dots. Positions marked with black dots drawn just above the main downward-pointing diagonal (*dashed line*) from upper left to lower right represent amino acid sequential positions aa1-aa2, aa2-aa3, etc., that are close to each other because they are in the  $\alpha$  helix. Marked regions of shortest  $C_{\alpha}$ - $C_{\alpha}$  distances along this diagonal thus indicate positions of the  $\alpha$  helices. Other marked diagonal regions (*red and blue dots*) indicate tertiary structural interactions, including those between adjacent secondary structural elements. Helices a and b are close to each other and have opposite chemical polarities so that aa10-aa11-aa12 . . . are close to aa40-aa39-aa38 . . . on the red surface of the helices. An upward-running diagonal (*red dots*) from lower left to upper right reveals this spatial relationship. Helices b and c are also close to each other but have the same polarity so that aa30-aa31-aa32 . . . are close to aa50-aa51-aa52 . . . , producing a downward-directed diagonal (*blue dots*). If another protein has a matrix pattern similar to that of the above example, then the two protein structures have the same three-helical arrangement and the loops joining the helices are of approximately the same length and conformation. The distance alignment method will find such three-helix patterns, even when the loop patterns are not similar. (B) Search for a common structural pattern in proteins A and B by DALI. A hypothetical example of a three-helix architecture is again used. In the top row, DALI first searches the entire distance matrix of protein A for a set of matching helices, a and b, indicated by an

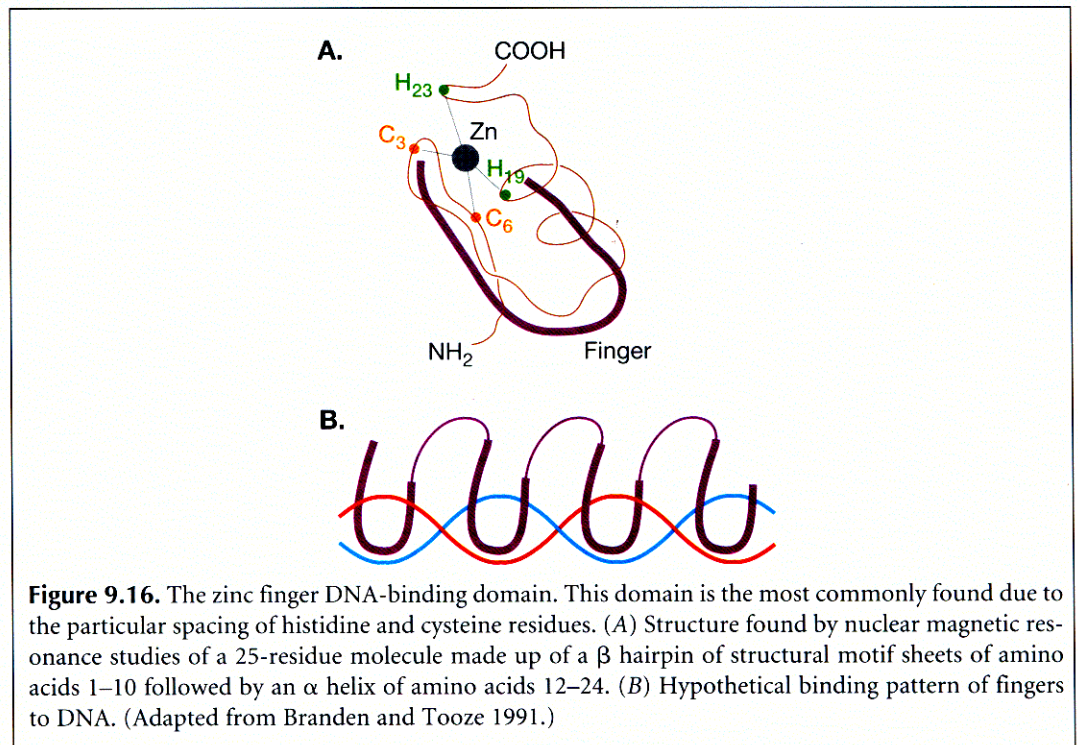
upward-directed diagonal whose position is the intersection of the locations of the helices in the sequence of protein A (*left column*). A similar search is performed for a corresponding pair of helices a' and b' in the distance matrix of protein B. In practice, the algorithm breaks down each full-sized matrix into a set of overlapping submatrices of size  $6 \times 6$  amino acids. Distance patterns within the submatrices from each protein are then compared to locate similar structural configurations. Some matches will be longer than 6 amino acids and will therefore be found in several neighboring submatrices. A computationally sophisticated assembly step in the algorithm (see below) combines these overlaps into a complete structural alignment. Once found, individual matches are assembled. If a pair of helices is found in each structure, a beginning structural alignment of the sequences may be made (*right column*). A search for a third pair of helices c and c' that interact with helices b and b' in proteins A and B, respectively, is then made, as illustrated in the second row. A hypothetical pair common to A and B is shown. In this case, the order of regions b' and c' on the sequence of protein B is reversed from that of b and c. The composite matrices and alignment of all helices a, b, c and a', b', c' are shown in the third row. Only the top one-half of the matrix is shown, leaving out the mirror image. Finally, DALI removes the insertions and deletions in the matrices and rearranges the sequence of the protein B to produce a parallel alignment of the elements in the two sequences (*bottom row*). By following these steps, an alignment of helices a, b, and c and a', b', and c' in structures A and B is found by DALI, but the arrangements of sequences that produce this common architecture are different. Structural features that include  $\beta$  strands in proteins are found in the same manner. (Diagram derived from Holm and Sander 1993, 1996.)

B. Protein A

Protein B

One-dimensional alignment





clusters are the most feasible. After starting with an alignment that includes the highest matching number of elements, the VAST algorithm examines alternative alignments that might increase the alignment score using the Gibbs sampling algorithm described in Chapter 4.

Like other structural alignment methods, VAST and SARF are available on Web pages and may be used for comparing new structures to the existing databases or for viewing structural similarities within the existing databases. An important aspect of searches for structural similarity by the vector method and other methods is the extent of the alignment found, or as Gibrat et al. (1996) state, is the alignment “surprising”?

## Significance of Alignments of Secondary Structure

As in sequence alignment, it is important to estimate the reliability or statistical significance of a structural alignment. The problem is to determine the probability with which a given cluster of secondary structural elements would be expected between unrelated structures. The analogous problem with sequences is to determine whether or not an alignment score between two test sequences would also be found between random or unrelated sequences.

When comparing the arrangement of secondary elements in protein structures, a very large number of possible alignments are commonly found (Gibrat et al. 1996). The probability of a chance alignment of a few elements in two large but structurally unrelated proteins that have many such elements is quite high. Therefore, alignment of only a few elements in an actual comparison of two test sequences is not particularly significant. The probability of an alignment between most of the elements in large, unrelated proteins, however, is extremely low. Hence, such an alignment between structures is highly significant. The problem of significance thus boils down to assessing the number of possible ways of aligning elements in two unrelated proteins.

For calculating the probability of an alignment, the VAST algorithm uses a statistical theory very similar to that of the BLAST algorithm to calculate this probability. Recall that BLAST calculates a probability (or expect value) that a sequence alignment score at least as high as that found between a test sequence and a database sequence would also be found by alignment of random sequences. Sequence alignment scores are derived by using amino acid substitution matrices and suitable alignment gap penalties, and the probabilities that alignments of random sequences could score as high as actual scores are calculated using the extreme value distribution. The equivalent VAST score is the number of superimposed secondary structural elements found in comparing two structures. The greater the number of elements that can be aligned, the more believable and significant the alignment. The statistical significance of a score is the likelihood that such a score would be seen by chance alignment of unrelated structures. This likelihood is calculated from the product of two numbers—the probability that such a score would be found by picking elements randomly from each protein domain and the number of alternative element pair combinations. Thus, if the chance of picking the number of matching elements found is  $10^{-8}$  and the number of combinations is  $10^4$ , the likelihood of an alignment of the same number of elements between unrelated structures is  $10^{-8} \times 10^4 = 10^{-4}$ .

## Displaying Protein Structural Alignments

The programs and Web sites that perform a structural alignment or that provide access to databases of similar structures will transmit coordinates of the matched regions. The aligned regions may then be viewed with a number of molecular viewing programs, including Rasmol, Cn3d, and Spdbv. Cn3d also shows a second window with the matching sequence alignment, and aligned structures may be highlighted starting from this window. The program JOY provides a method for annotating sequence alignments with three-dimensional structural information (<http://www-cryst.bioc.cam.ac.uk/~joy>; Mizuguchi et al. 1998b).

## STRUCTURAL PREDICTION

### Use of Sequence Patterns for Protein Structure Prediction

Although the sequences of 86,000 proteins are available, the structures of only 12,500 of these proteins are known. The increasing rate of genome sequencing can also be expected to outpace the rate of solving protein structures. Protein structural comparisons described above have shown that newly found protein structures often have a similar structural fold or architecture to an already-known structure. Thus, many of the ways that proteins fold into a three-dimensional structure may already be known. Structural comparisons have also revealed that many different amino acid sequences in proteins can adopt the same structural fold, and these sequences have been organized into databases described above. Further examination of sequences in structures has also revealed that the same short amino acid patterns may be found in different structural contexts. Amino acid sequences present in secondary structures have been entered into databases that are useful for structure prediction. Many proteins in the sequence databases also have conserved sequence patterns upon which they may be further categorized.

If two proteins share significant sequence similarity, they should also have similar three-dimensional structures. The similarity may be present throughout the sequence

lengths or in one or more localized regions having relatively short patterns that may or may not be interrupted with gaps. When a global sequence alignment is performed, if more than 45% of the amino acid positions are identical, the amino acids should be quite superimposable in the three-dimensional structure of the proteins. Thus, if the structure of one of the aligned proteins is known, the structure of the second protein and the positions of the identical amino acids in this structure may be reliably predicted. If less than 45% but more than 25% of the amino acids are identical, the structures are likely to be similar, but with more variation at the lower identity levels at the corresponding three-dimensional positions.

### ***Protein Classification Schemes***

Proteins have been classified on the basis of sequence similarity or the presence of common amino acid patterns. First, they have been organized into families and superfamilies on the basis of the level of sequence similarity in sequence alignments. The current method of organizing proteins by this method at the Protein Information Resource (PIR) (<http://www-nbrf.georgetown.edu>) is that each entry in the PIR protein sequence database is searched against the remaining entries using the FASTA algorithm. Similar sequences are then aligned with the Genetics Computer Group multiple sequence alignment program PILEUP. This level of comparison based on sequence alignment was originally made by the PIR founded by M. Dayhoff. Using present-day classification schemes (Barker et al. 1996), families are composed of proteins that align along their entire lengths with a level of sequence identity of usually 50% or better.

More recent analyses of amino acid patterns in protein sequences have revealed that many proteins are made up of modules, short regions of similar amino acid sequence that correspond to a particular function or structure. Furthermore, sets of proteins from widely divergent biological sources may share several such modules and the modules may not be in the same order. Hence, it has become necessary to redefine the concepts of family and superfamily. Proteins that comprise the same set of similar homology domains (extended regions of sequence similarity) in the same order are referred to as homeomorphic protein families. Protein families, members of which have the same domains in the same order, but also have dissimilar regions, are designated as a homeomorphic superfamily (Barker et al. 1996). The superfamily classification of a newly identified protein sequence may be analyzed at several Web sites (Table 9.5).

The second method of classifying proteins is based on the presence of amino acid patterns. Proteins with the same biochemical function have been examined for the presence of strongly conserved amino acid patterns that represent an active site or other important feature. The resulting database is known as the Prosite catalog (A. Bairoch and colleagues; Hofmann et al. 1999) (Table 9.5). Proteins have also been categorized on the basis of the occurrence of common amino acid patterns—motifs and conserved gapped and ungapped regions in multiple sequence alignments. These patterns are found by extracting them from multiple sequence alignments, by pattern-finding algorithms that search unaligned sequences for common patterns, and by several statistical methods that search through unaligned sequences. The patterns vary in length, presence of gaps, and degree of substitution. The algorithms that are used include pattern-finding methods, hidden Markov models, the expectation maximization method, and the Gibbs sampling method. These methods and the computer programs and Web sites that provide them are described in Chapter 4. Listed in Table 9.5 are several databases that categorize proteins based on the occurrence of common patterns. Also shown are databases of amino acid patterns that



determine cellular localization of proteins or sites of protein modification (signal or transit peptides). FSSP, a structural family database, is listed in this table because it includes links to information on sequence families and superfamilies.

A given protein sequence may be classified by using one of the resources in Table 9.5 for sequence patterns that are characteristic of a group or family of proteins. Because most of these databases are derived by quite different methods of pattern analysis, statistics, and database similarity searching, they can be expected to provide complementary information. Thus, a given database may include a sequence pattern that is not identified in others, and this pattern may provide an important link to structure or function for one group of proteins. Another database may provide patterns more suitable for classifying a different group of proteins. Therefore, a wise choice would be to use as many of these resources as possible for classifying a new sequence. However, note the availability of Web sites that have combined the resources of separate protein classification databases into a single database (e.g., INTERPRO; Table 9.5). In one new field of endeavor, protein taxonomy, genomic databases that list the entire set of proteins produced by a particular organism are searched for matches. Such searches can provide a wealth of information on protein evolution (Pellegrini et al. 1999).

### ***Clusters***

Another, more recently introduced, method for classifying proteins is to use clustering methods. In these methods, every protein in a sequence database such as SwissProt is compared to every other sequence using a database search method including the BLAST, FASTA, and Smith-Waterman dynamic programming methods described in Chapter 7. Thus, each protein in the database receives a sequence similarity score with every other sequence. A similar method is used to identify families of paralogous proteins encoded by a single genome (p. 501). Matching sequences are further aligned by a pair-wise alignment program like LALIGN to recalculate the significance of the alignment score (see Chapter 3 flowchart, p. 58). In a cluster analysis, sequences are represented as vertices on a graph, and those vertices representing each pair of related sequences are joined by an edge that is weighted by the degree of similarity between the pair (see Fig. 10.4). In a first step, the clustering algorithm detects the sets of proteins that are joined in the graph by strongly weighted edges. In subsequent steps, relationships between the initial clusters found in the first step are identified on the basis of weaker, but still significant, connections between them. These related clusters are then merged in a manner that maximizes the strongest global relationships (see Web sites for ProtoMap and SYSTERS; Table 9.5). Clustering has been used to identify groups of proteins that lack a relative with a known structure and hence are suitable for structural analysis (Portugaly and Linial 2000). Additional information on clustering methods is provided in Chapter 10.

### ***Proteins Comprise Motifs, Modules, and Other Sequence Elements of Structural Significance***

The above analysis describes the types and distribution of motifs in proteins from the same or different organisms. A motif can represent an individual folded structure or active-site residues. Several different motifs widely separated in the same protein sequence are often found. These motifs represent conserved regions that lie in the core of the protein structure. Hence, their presence in two sequences predicts a common structural core (for review, see Henikoff et al. 1997).

**Table 9.5.** *Databases of patterns and sequences of protein families*

Name	Web address	Description	Reference
3D-Ali	<a href="http://www.embl-heidelberg.de/argos/ali/ali_info.html">http://www.embl-heidelberg.de/argos/ali/ali_info.html</a>	aligned protein structures and related sequences using only secondary structures assigned by author of the structures	Pascarella and Argos (1992)
3D-PSSM	<a href="http://www.bmm.icnet.uk/3dpssm">http://www.bmm.icnet.uk/3dpssm</a>	uses a library of scoring matrices based on structural similarity given in the SCOP classification scheme (p. 402) for alignment with matrices based on sequence similarity	Kelley et al. (2000)
BLOCKS	<a href="http://blocks.fhcrc.org/">http://blocks.fhcrc.org/</a>	ungapped blocks in families defined by the Prosite catalog	Henikoff and Henikoff (1996); Henikoff et al. (1998)
COGS (Clusters of Orthologous Groups database and search site)	<a href="http://www.ncbi.nlm.nih.gov/COG">http://www.ncbi.nlm.nih.gov/COG</a>	clusters of similar proteins in at least three species collected from available genomic sequences	Tatusov et al. (1997)
DIP (Database of Interacting Proteins)	<a href="http://dip.doe-mbi.ucla.edu">http://dip.doe-mbi.ucla.edu</a>	database of interacting proteins	Xenarios et al. (2000)
eMOTIF	<a href="http://dna.Stanford.EDU/emotif/">http://dna.Stanford.EDU/emotif/</a>	common and rare amino acid motifs in the BLOCKS and HSSP databases	Nevill-Manning et al. (1998)
HOMSTRAD	<a href="http://www-cryst.bioc.cam.ac.uk/~homstrad/">http://www-cryst.bioc.cam.ac.uk/~homstrad/</a>	structure-based alignments organized at the level of homologous families <sup>a</sup>	Mizuguchi et al. (1998a)
HSSP	<a href="http://swift.embl-heidelberg.de/hssp/">http://swift.embl-heidelberg.de/hssp/</a> <a href="http://www.sander.ebi.ac.uk/hssp/">http://www.sander.ebi.ac.uk/hssp/</a>	sequences similar to proteins of known structure	Dodge et al. (1998)
INTERPRO integrated resource of protein domains and functional sites <sup>b</sup>	<a href="http://www.ebi.ac.uk/interpro">http://www.ebi.ac.uk/interpro</a>	combination of Pfam, PRINTS, Prosite, and current SwissProt/TrEMBL sequence	see Web site
LPFC	<a href="http://www-camis.stanford.edu/projects/helix/LPFC/">http://www-camis.stanford.edu/projects/helix/LPFC/</a>	a library of protein family cores based on multiple sequence alignment of protein cores using amino acid substitution matrices based on structure (see Chapter 3)	see Web page
NetOGlyc 2.0 prediction server	<a href="http://www.cbs.dtu.dk/services/NetOGlyc/">http://www.cbs.dtu.dk/services/NetOGlyc/</a>	predicts glycosylation sites in mammalian proteins by neural network analysis	Hansen et al. (1997)
NNPSL	<a href="http://predict.sanger.ac.uk/nnpsl/">http://predict.sanger.ac.uk/nnpsl/</a>	predicts subcellular location of proteins by neural network	see Web site
Pfam	<a href="http://www.sanger.ac.uk/Pfam">http://www.sanger.ac.uk/Pfam</a>	profiles derived from alignment of protein families, each one composed of similar sequence and analyzed by hidden Markov models	Sonnhammer et al. (1998)
PIR	<a href="http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml">http://www-nbrf.georgetown.edu/pirwww/pirhome.shtml</a>	family and superfamily classification based on sequence alignment	Barker et al. (1996)
PRINTS	<a href="http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html">http://www.biochem.ucl.ac.uk/bsm/dbbrowser/PRINTS/PRINTS.html</a>	protein fingerprints or sets of unweighted sequence motifs from aligned sequence families	Attwood et al. (1999)

Table 9.5. *Continued.*

Name	Web address	Description	Reference
PROCLASS	<a href="http://www-nbrf.georgetown.edu/gfserver/proclass.html">http://www-nbrf.georgetown.edu/gfserver/proclass.html</a>	database organized by Prosite patterns and PIR superfamilies; neural network system for protein classification into superfamily	Wu (1996); Wu et al. (1996)
PRODOM	<a href="http://protein.toulouse.inra.fr/prodom.html">http://protein.toulouse.inra.fr/prodom.html</a>	groups of sequence segments or domains from similar sequences found in SwissProt database by BLASTP algorithm; aligned by multiple sequence alignment	Corpet et al. (1998)
Prosite	<a href="http://www.expasy.ch/prosite">http://www.expasy.ch/prosite</a>	groups of proteins of similar biochemical function on basis of amino acid patterns	Bairoch (1991); Hofmann Bairoch et al. (1999)
ProtoMap	<a href="http://protomap.cornell.edu">http://protomap.cornell.edu</a>	classification of SwissProt and TrEMBL proteins into clusters	Yona et al. (1999)
PSORT	<a href="http://psort.nibb.ac.jp">http://psort.nibb.ac.jp</a>	predicts presence of protein localization signals in proteins	see Web site
SignalP Web server	<a href="http://www.cbs.dtu.dk/services/SignalP/">http://www.cbs.dtu.dk/services/SignalP/</a>	predicts presence and location of signal peptide cleavage sites in proteins of different organisms by neural network analysis	Nielsen et al. (1997)
SMART	<a href="http://smart.embl-heidelberg.de">http://smart.embl-heidelberg.de</a>	database of signaling domain sequences with accurate alignments	Schultz et al. (1998)
SYSTEMS	<a href="http://www.dkfz-heidelberg.de/tbi/services/cluster/systemsform">http://www.dkfz-heidelberg.de/tbi/services/cluster/systemsform</a>	classification of all sequences in the SwissProt database into clusters based on sequence similarity	Krause et al. (2000)
TargetDB	<a href="http://molbio.nmsu.edu:81/">http://molbio.nmsu.edu:81/</a>	database of peptides that target proteins to cellular locations	see Web site

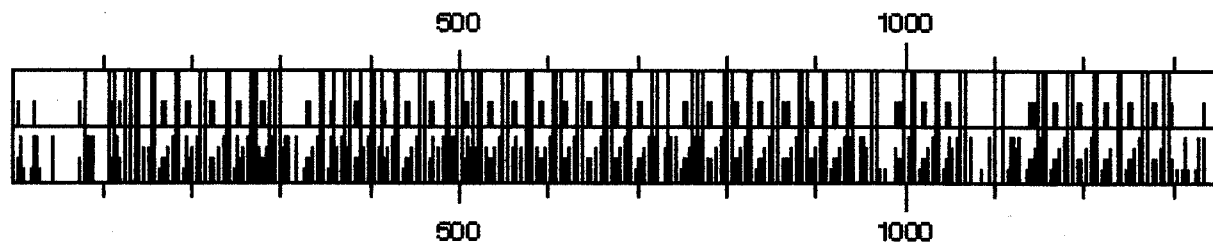
A list of Web sites with protein sequence/structure databases is maintained at <http://www.imb-jena.de/ImgLibDoc/help/db/>. Many protein family databases are accessible through the European Bioinformatics Institute (<http://srs.ebi.ac.uk/>). Information on the available protein family databases is also found on the MetaFam site at <http://metafam.ahc.umn.edu/>.

<sup>a</sup> Sequence alignments of each family shown with residues labeled by solvent accessibility, secondary structure, H bonds to main-chain amide or carbonyl group, disulfide bond, and positive  $\Phi$  angle.

<sup>b</sup> A combination of Pfam 5.0, PRINTS 25.0, Prosite 16, and current SwissProt and TrEMBL data. Additional merges with other protein pattern databases are planned.

A more detailed analysis of motifs has revealed that they are components of a more fundamental unit of structure and function, the protein module. Proteins may have several modules corresponding to different units of function, and these modules may be present in a different order (Henikoff et al. 1997). These diverse arrangements suggest that a biologically important module has been repeatedly employed in protein evolution by gene duplication and rearrangement mechanisms that are discussed in Chapter 6 and Chapter 10. The presence of modules also provides a further system of protein classification into module-based families.

An example of an important motif is the C<sub>2</sub>H<sub>2</sub> (2 cysteines and 2 histidines) zinc finger DNA-binding motif X<sub>fin</sub> of *Xenopus laevis* illustrated in Figure 9.16. The zinc finger is one of the most commonly identified motifs, in part due to the characteristic spacing of C and H residues in the motif sequence. As indicated in Figure 9.17, the zinc atom forms bonds with these residues to create the finger-like projection. When present in tandem copies, the finger is thought to lie in an alternating pattern in the major groove of DNA. A simple plot



**Figure 9.17.** Graph of the *Xenopus laevis* XFIN protein sequence which is in the Cys-Cys-His-His class of zinc finger DNA-binding proteins (Branden and Tooze 1991). The graph was produced using the AA Window, Cys + His map option of DNA STRIDER vers. 1.2 on a Macintosh computer. The bottom panel shows amino acids Y, C, F, L, and H, respectively, as bars of increasing length. The top panel shows H and C as half- and full bars, respectively. The fingers appear in the top panel as double half-bars (two Cys residues separated by 2 amino acids) followed by double full bars (two His residues separated by 2 amino acids). This type of graphic representation is extremely useful for visualizing amino acid patterns in proteins.

of the positions of C and H residues on the protein sequence as shown in Figure 9.17 provides a very simple way to locate zinc fingers in a protein sequence.

Pfam is a Web site that provides a listing of proteins that carry the zinc finger sequence motif. As shown in Figure 9.18, the zinc finger is one of the most commonly recognized motifs, and proteins that carry the motif have been classified into a family. Two other families of zinc finger proteins with 4 cysteine or 3 cysteine and 1 histidine residues interacting with the Zn atom, and additional variations in the basic structure of zinc fingers, have also been identified. Descriptions and alignments of these proteins are provided at the Pfam Web site, as illustrated in Figure 9.19. Other families in the Pfam classification are given a description that best reflects the extent and complexity of the conserved sequence patterns, be it a domain, module, repeat, or motif. In general, all of these patterns represent a conserved unit of structure or function.

### ***Structural Features of Some Proteins Are Readily Identified by Sequence Analysis***

The above section indicates that a newly identified protein may be classified on the basis of the presence of sequence motifs, modules, or other sequence elements that represent structure or function. The zinc finger motif is one structural motif that may be readily identified on the basis of the order and spacing of a conserved pattern of cysteine and histidine residues in the sequence. Other classes of proteins have characteristic amino acid composition and patterns such that the structure can often be reliably predicted from the amino acid sequence. Some other examples of structure recognition on the basis of sequence are given below.

***Leucine zippers and coiled coils.*** The leucine zipper motif is typically made up of two antiparallel  $\alpha$  helices held together by interactions between hydrophobic leucine residues located at every seventh position in each helix, as illustrated in Figure 9.20A. The zipper holds protein subunits together. The leucines are located at approximately every two turns of the  $\alpha$  helix. It is this repeated occurrence of leucines that makes the motif readily identifiable. In the transcription factors Gcn4, Fos, Myc, and Jun, the binding of the subunits forms a scissor-like structure with ends that lie on the major groove of DNA, as shown in Figure 9.20B. If the amino acids in each helical region are plotted as a spiral of 3.6 amino acid residues per turn, representing a view looking down the helix from the end starting at residue 1 on the inside of the spiral, then the result shown in Figure 9.20C is found. The leucine residues are found on approximately the same side of the helix, slightly out of phase

# Pfam

## Browse Families

Browse available alignments and models



[Pfam \(St. Louis\)](#) | [Pfam \(Cambridge\)](#) | [Pfam \(Stockholm\)](#) | [HMMER software](#) | [WashU Dept. of Genetics](#) |  
[Home](#) | [Analyze a sequence](#) | [Browse alignments](#) | [Text search](#) | [Swisspfam](#) | [Help & more information](#) |

### Pfam 3.4: available alignments and models

The families are grouped under the first letter of their name, regardless of case. All families starting with a number are found in 'Number'

Available sections: [Numbers](#) [A](#) [B](#) [C](#) [D](#) [E](#) [F](#) [G](#) [H](#) [I](#) [J](#) [K](#) [L](#) [M](#) [N](#) [O](#) [P](#) [Q](#) [R](#) [S](#) [T](#) [U](#) [V](#) [W](#) [X](#) [Y](#) [Z](#) [Top twenty families](#)

Top twenty families							
Name	acc number	#seed	#full	av. len	av. %id	structure	Description
<a href="#">GP120</a>	<a href="#">PF00516</a>	24	13408	131 aa	53%	<a href="#">1gc1</a>	Envelope glycoprotein GP120
<a href="#">zf-C2H2</a>	<a href="#">PF00096</a>	200	4991	23 aa	35%	<a href="#">1zaa</a>	Zinc finger, C2H2 type
<a href="#">ig</a>	<a href="#">PF00047</a>	65	3495	65 aa	20%		Immunoglobulin domain
<a href="#">RuBisCO large</a>	<a href="#">PF00016</a>	17	3007	401 aa	77%	<a href="#">3rub</a>	Ribulose biphosphate carboxylase, large chain
<a href="#">pkinase</a>	<a href="#">PF00069</a>	67	2942	212 aa	24%		Eukaryotic protein kinase domain
<a href="#">cytochrome b N</a>	<a href="#">PF00033</a>	9	2866	152 aa	69%		Cytochrome b(N-terminal)/b6/petB
<a href="#">EGF</a>	<a href="#">PF00008</a>	73	2388	34 aa	35%	<a href="#">1apo</a>	EGF-like domain
<a href="#">Collagen</a>	<a href="#">PF01391</a>	15	2125	59 aa	42%		Collagen triple helix repeat (20 copies)
<a href="#">fn3</a>	<a href="#">PF00041</a>	109	2103	85 aa	20%		Fibronectin type III domain
<a href="#">efhand</a>	<a href="#">PF00036</a>	86	1773	28 aa	27%	<a href="#">1osa</a>	EF hand
<a href="#">LRR</a>	<a href="#">PF00560</a>	300	1753	47 aa	23%	<a href="#">1bnh</a>	Leucine Rich Repeat (2 copies)
<a href="#">MHC II beta</a>	<a href="#">PF00969</a>	165	1688	44 aa	66%	<a href="#">1seb</a>	Class II histocompatibility antigen, beta domain
<a href="#">zf-CCHC</a>	<a href="#">PF00098</a>	122	1678	17 aa	57%	<a href="#">1ncp</a>	Zinc finger, CCHC class
<a href="#">ank</a>	<a href="#">PF00023</a>	95	1663	33 aa	27%		Ank repeat
<a href="#">ryp</a>	<a href="#">PF00077</a>	34	1508	95 aa	79%	<a href="#">1ida</a>	Retroviral aspartyl proteases

**Figure 9.18.** The Pfam Web mirror site at Washington University (<http://pfam.wustl.edu/browse.shtml>). Shown are the 20 most common protein families classified according to the motifs that are present. Note the presence of the Pfam entry for zf-C2H2, the name assigned to the C<sub>2</sub>H<sub>2</sub> (2 cysteines and 2 histidines) Zn finger DNA-binding motif, accession no. PF00096. Any family may be examined by clicking the mouse on the first letter of the family name. Fig. 9.19 is an example of the entry for the PF00096.

*Continues on next page*

<u>WD40</u>	<u>PF00400</u>	37	1482	39 aa	24%		WD domain, G-beta repeat
<u>homeobox</u>	<u>PF00046</u>	45	1431	49 aa	41%	<u>lahd</u>	Homeobox domain
<u>7tm_1</u>	<u>PF00001</u>	64	1423	230 aa	19%		7 transmembrane receptor (rhodopsin family)
<u>gag_p17</u>	<u>PF00540</u>	4	1319	104 aa	77%	<u>2hmx</u>	gag gene protein p17 (matrix protein).
<u>oxidored_q1</u>	<u>PF00361</u>	33	1315	220 aa	32%	<u>lmin</u>	NADH-Ubiquinone/plastoquinone (complex I), various chains

1407 families.



Figure 9.18. *Continued.*

with the rotational symmetry of the helix. The predicted structure is that of a coiled coil, as shown in Figure 9.20D (Branden and Tooze 1991).

Coiled-coil structures typically comprise two to three  $\alpha$  helices coiled around each other in a left-handed supercoil in a manner that slightly distorts the helical repeat so that it is 3.5 residues per turn instead of the usual 3.6, or an integral number of 7 residues every second turn (Lupas 1996). They occur in fibrous proteins such as keratin and fibrinogen, and are also thought to occur in leucine zippers, as there is a repeat of leucine at every seventh residue (Branden and Tooze 1991). If the spiral wheel in Figure 9.20C is plotted so that there are 7 residues every second turn instead of 7.2, then the residues align more uniformly on one face of the helix. Consequently, the leucine zipper has been hypothesized to adopt a coiled-coil structure.

Coiled-coil regions may be predicted by searching for the 7-residue (heptad) periodicity observed in the sequence of these proteins. Naming these respective positions a, b, c, d, e, f, and g, then a and d are usually hydrophobic amino acids and the remaining amino acids are hydrophilic because coiled coils are generally fibrous, solvent-exposed structures. As more and more of these sequential patterns are observed along a sequence, one can be more convinced that the prediction is reliable. If there are at least 5–10 of these heptads and the hydrophobicity pattern is strongly conserved, the prediction is a good one. Poorer quality patterns come into doubt.

A program COILS2 has been developed for predicting coiled-coil regions with greater reliability than simple pattern searching for heptad repeats (Lupas et al. 1991; Lupas 1996; program description at <http://www.embl-heidelberg.de/predictprotein/>). There are two Web sites for predicting the occurrence of coiled-coil regions in protein sequences using the COILS program—<http://www.isrec.isb-sib.ch/software/software.html> and <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>. The program may also be obtained from these sites for running on a local server. Central to the method is the generation of a profile scoring matrix, with each column showing the distribution of amino acids in each of the seven positions, a–g, found in all of the known coiled-coil proteins.

Pfam 3.4 (St. Louis) : [Home](#) | [Analyze a sequence](#) | [Browse alignments](#) | [Text search](#) | [Swisspfam](#) | [Help](#) |

## Pfam entry: zf-C2H2

Accession number: PF00096  
 Definition: Zinc finger, C2H2 type  
 Author: Bateman A, Boehm S, Sonnhammer ELL  
 Source of seed members: Boehm S  
 Alignment method of seed: Manual  
 HMM build command line: hmmbuild HMM SEED  
 HMM build command line: hmmscalibrate --seed 0 HMM  
 Gathering method: hmmssearch -T 15 --domT 5  
 Trusted cutoffs: 15.00 5.00  
 Noise cutoffs: 14.80 17.50  
 Reference Number: [1]  
 Reference Medline: [97315340](#)  
 Reference Title: Variations of the C2H2 zinc finger motif in the yeast genome and classification of yeast zinc finger proteins.  
 Reference Author: Boehm S, Frishman D, Mewes HW;  
 Reference Location: Nucleic Acids Res 1997;25:2464-2469.  
 Database Reference: PROSITE; [PDOC00028](#);  
 Database Reference: PRINTS; [PR00048](#);  
 Database Reference: SCOP; 1zaa; fa; [[SCOP-USA](#)][[CATH-PDBSUM](#)]  
 Comment: The C2H2 zinc finger is the classical zinc finger domain.  
 Comment: The two conserved cysteines and histidines co-ordinate a zinc ion. The following pattern describes the zinc finger.  
 Comment: #-X-C-X(1-5)-C-X3-#-X5-#-X2-H-X(3-6)-[H/C]  
 Comment: Where X can be any amino acid, and numbers in brackets indicate the number of residues. The positions marked # are those that are important for the stable fold of the zinc finger. The final position can be either his or cys.  
 Comment: The C2H2 zinc finger is composed of two short beta strands followed by an alpha helix. The amino terminal part of the helix binds the major groove in DNA binding zinc fingers.  
 Number of members: 4991

### Retrieve a Pfam alignment for zf-C2H2

Which alignment:  ↕

What format:  ↕

Output straight text. (Default is HTML-ized text.)

Pfam 3.4 (St. Louis) : [Home](#) | [Analyze a sequence](#) | [Browse alignments](#) | [Text search](#) | [Swisspfam](#) | [Help](#) |

Comments, questions, flames? Email [pfam@genetics.wustl.edu](mailto:pfam@genetics.wustl.edu).

**Figure 9.19.** The Pfam entry for family zf-C2H2 (accession no. PF00096). The mouse was clicked on the entry for zf-C2H2 shown in the above figure. The Pfam database is based on a statistical analysis of sequences with the same motif using hidden Markov models. The result is a profile of the sequences with matches, mismatches, and gaps. The entry describes how this profile was produced by the HMMER program, and also provides references and a link to a multiple sequence alignment of the sequences. As discussed in Chapter 3, this hidden Markov model of the sequences can be used to produce the multiple sequence alignment by choosing the most probable path through the model.

globular residues in GenBank. These scores will vary with each window size and option chosen, and the scores may be normalized to give a better impression of their range. A false positive can occur with sequences that have a biased amino acid distribution; these false positives can be identified by the program option of weighting the two hydrophobic positions a and d the same as the five hydrophilic positions b, c, e, f, and g. Normally, these positions are weighted 2.5 times more heavily during the scoring procedure. False positives will continue to have a high score whereas true positives will not.

For candidate protein sequences, Lupas recommends using both types of weighting and both MTK and MTIDK matrices. The program reliably predicts known coiled-coil regions (Lupas 1996). An example of the program output from the ISREC Web site is shown in Figure 9.21, using as input the sequence of Gcn4 (identified as GCN4\_YEAST in the SwissProt database), which has a leucine zipper region. The protein is scanned for the number of occurrences of coiled coils in a sliding window of 7, 14, 21, or 28 residues.

Another method for predicting coiled coils is based on an analysis of correlations between pairs of amino acids (Berger et al. 1995), and the program is accessible at <http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html>.

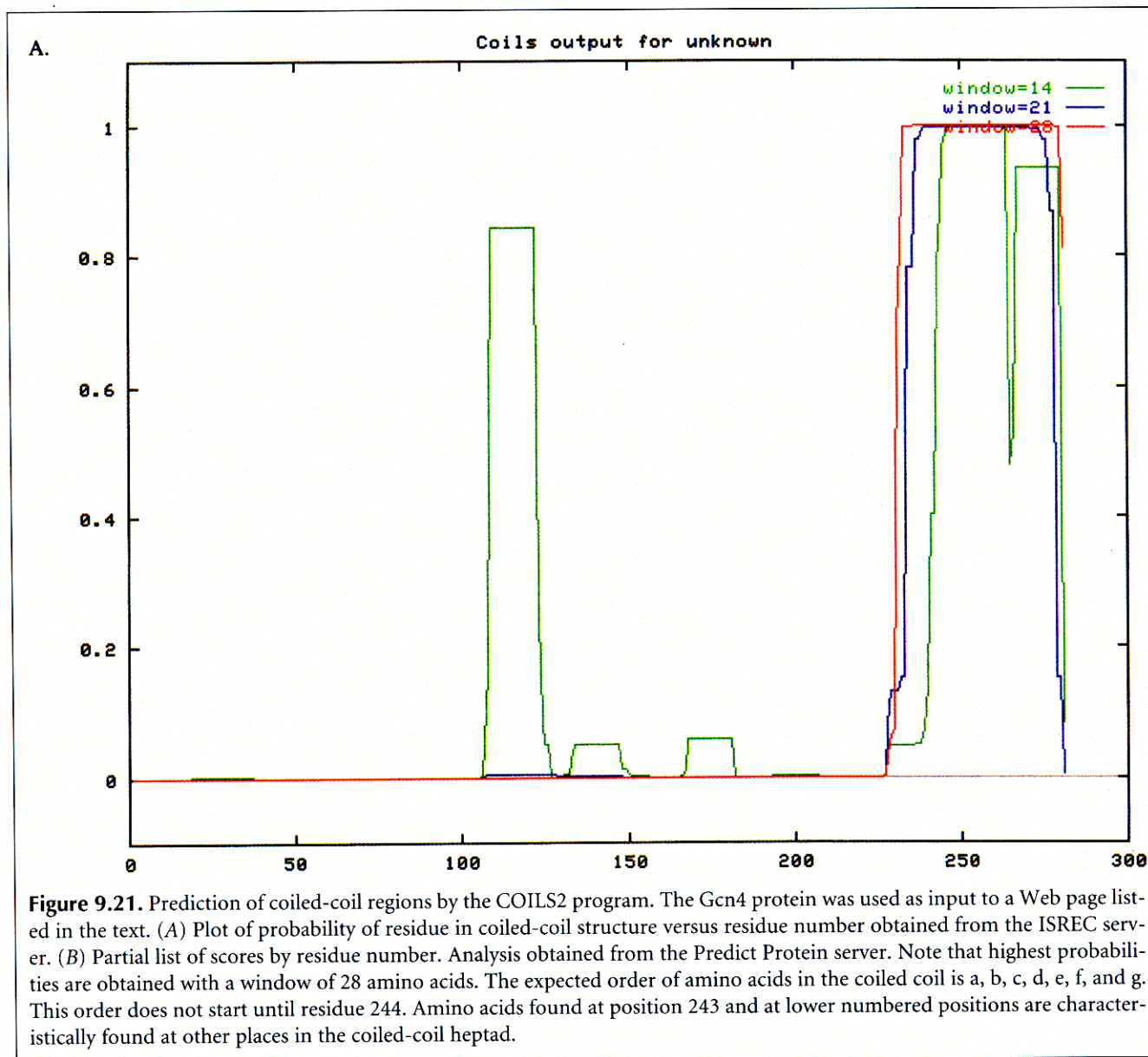
### ***Transmembrane-spanning Proteins***

The all- $\alpha$  superfamily of membrane proteins (see classification of membrane proteins at the SCOP structural database at <http://scop.mrc-lmb.cam.ac.uk/scop/>) is composed of proteins that traverse membranes back and forth through a series of  $\alpha$  helices comprising amino acids with hydrophobic side chains. The typical length, 20–30 residues, and strong hydrophobicity of these helices provide a simple method for scanning a candidate sequence for such features. An example of such a structure is illustrated in Figure 9.22.

Membrane-spanning hydrophobic  $\alpha$  helices can be quite accurately located by scanning for hydrophobic regions about 19 residues in length in the amino acid sequence (Kyte and Doolittle 1982). The occurrence of such regions in a candidate protein of unknown structure is a good indicator that the region spans a membrane. In Figure 9.23, such an analysis is shown for subunit M of the above molecule. Membrane-spanning helices are different from  $\alpha$  helices that are located on the surface of a protein structure. The surface helices tend to have hydrophobic residues located on the core-facing side (inside) and the hydrophilic residues on the solvent-facing side (outside) of the helix. These surface-exposed helices can be recognized by this separation of hydrophobic residues through a helical moment analysis described below. Membrane  $\alpha$  helices are more like  $\alpha$  helices that are buried in the structural core of a protein, which also have a high proportion of amino acids with hydrophobic side groups located throughout their lengths. In an effort to distinguish different classes of  $\alpha$  helices, several methods for improving the prediction of transmembrane regions have been devised and are available on Web sites.

One such method is one of the program choices of the PHD (profile-fed neural network system from Heidelberg) server for protein structure prediction at <http://www.embl-heidelberg.de/predictprotein/predictprotein.html>. The membrane-spanning helix predict program is named PHDhtm (PHD for *helical transmembrane* proteins). Briefly, a machine learning method called a neural network (see below) is trained to recognize the sequence patterns and sequence variations of a set of  $\alpha$ -helical transmembrane proteins of known three-dimensional structure. A candidate sequence is then scanned for the presence of similar sequence variations and a prediction is made as to the occurrence and location of  $\alpha$ -helical domains in the candidate protein. The specific steps were as follows. First, each of the small number of structurally identified  $\alpha$ -helical transmembrane proteins was used to





search the SwissProt protein sequence database for additional sequences in this superfamily using the BLAST or FASTA algorithms. Second, the sequences found were assembled first into a multiple sequence alignment and then into a motif by the program MAXHOM. Sequences less than 30% identical, and therefore least likely to be in the superfamily, were not included. The most-alike sequences in the alignment were also removed to provide a representative and statistically reasonable range of amino acid substitutions in each column of the motif. The neural network was then trained to differentiate between columns in the motif representing the  $\alpha$ -helical domains and the flanking nonhelical domains. The training method is described in greater detail below. The orientation of the predicted  $\alpha$ -helical domains with respect to the inside (cytoplasmic) or outside of the membrane is also predicted based on the observed preponderance of positively charged amino acids on the cytoplasmic side of solved structures (Rost et al. 1995). An illustrative example of a PHDhtm analysis on protein Iprc\_M is shown in Figure 9.24. As shown, the program correctly predicts five transmembrane helices, but positions of the ends of these helices are not

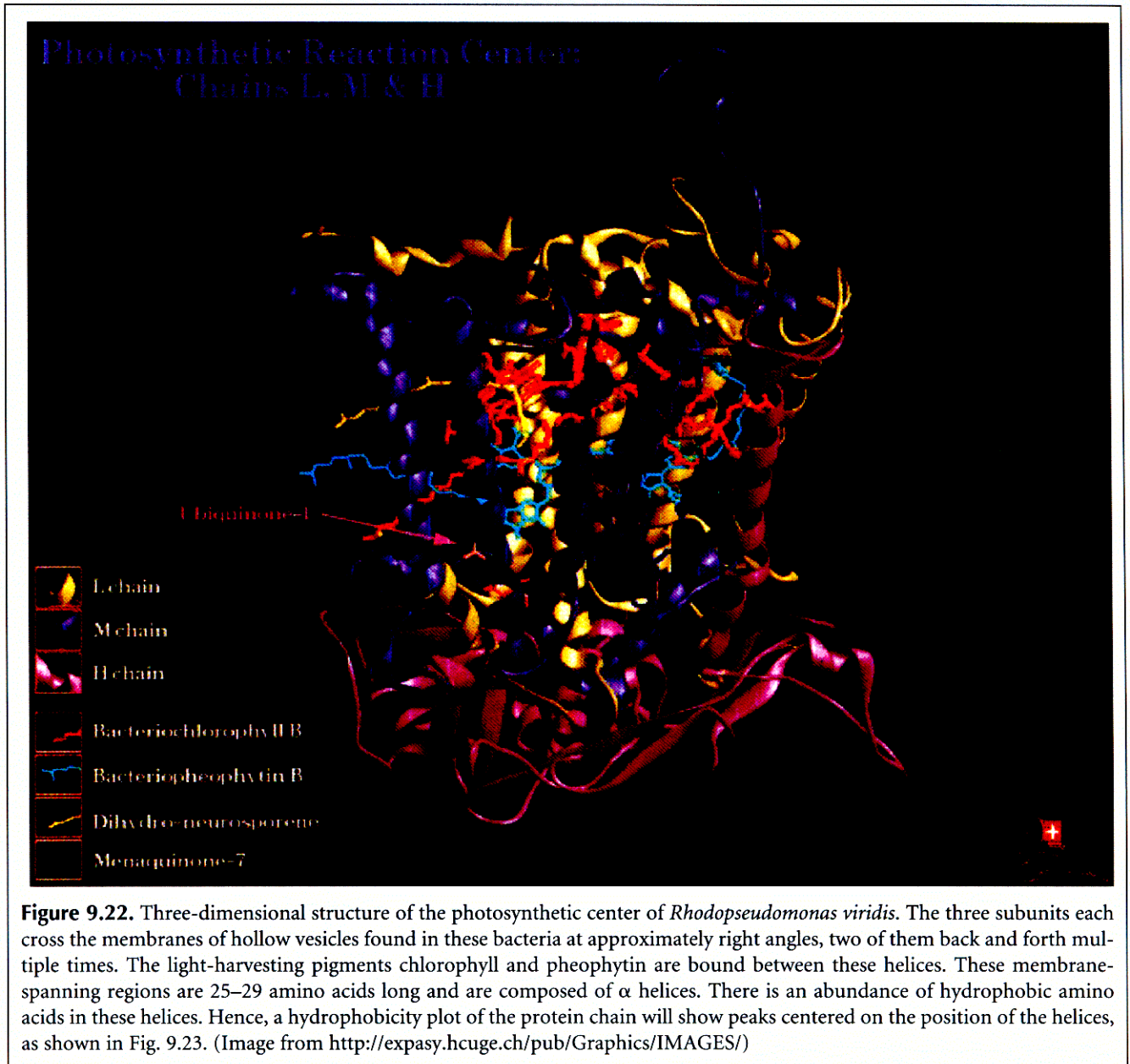
B.  
 COILS version 2.1  
 using MTK matrix.  
 weights: a,d=2.5 and b,c,e,f,g=1.0  
 Input file is /home/phd/server/work/predict\_h24138-21300.fasta  
 >prot (#) ppOld, gcn4 /home/phd/server/work/predict\_h24138

Residue	Window=14		Window=21		Window=28	
	Score	Probability	Score	Probability	Score	Probability
239 A	g 1.240	0.011	d 1.529	0.419	d 1.546	0.852
240 R	a 1.240	0.011	e 1.581	0.585	e 1.546	0.852
241 R	e 1.446	0.051	f 1.581	0.585	f 1.546	0.852
242 S	f 1.446	0.051	g 1.581	0.585	g 1.546	0.852
243 R	g 1.450	0.052	a 1.581	0.585	a 1.551	0.862
244 A	b 1.529	0.093	b 1.607	0.664	b 1.643	0.968
245 R	c 1.592	0.145	c 1.607	0.664	c 1.643	0.968
246 K	d 1.669	0.238	d 1.607	0.664	d 1.643	0.968
247 L	e 2.433	0.994	e 1.843	0.978	e 1.984	1.000
248 Q	f 2.433	0.994	f 1.988	0.997	f 2.041	1.000
249 R	g 2.433	0.994	g 2.018	0.998	g 2.052	1.000
250 M	a 2.433	0.994	a 2.054	0.999	a 2.052	1.000
251 K	b 2.433	0.994	b 2.054	0.999	b 2.052	1.000
252 Q	c 2.433	0.994	c 2.054	0.999	c 2.052	1.000
253 L	d 2.433	0.994	d 2.054	0.999	d 2.052	1.000
254 E	e 2.433	0.994	e 2.054	0.999	e 2.052	1.000
255 D	f 2.433	0.994	f 2.054	0.999	f 2.052	1.000
256 K	g 2.433	0.994	g 2.054	0.999	g 2.052	1.000
257 V	a 2.433	0.994	a 2.054	0.999	a 2.052	1.000
258 E	b 2.433	0.994	b 2.054	0.999	b 2.052	1.000
259 E	c 2.433	0.994	c 2.054	0.999	c 2.052	1.000
260 L	d 2.433	0.994	d 2.054	0.999	d 2.052	1.000
261 L	e 2.433	0.994	e 2.054	0.999	e 2.052	1.000
262 S	f 2.421	0.993	f 2.054	0.999	f 2.052	1.000
263 K	g 2.421	0.993	g 2.054	0.999	g 2.052	1.000
271 V	a 2.026	0.848	a 2.004	0.998	a 2.052	1.000
272 A	b 2.026	0.848	b 1.968	0.996	b 2.052	1.000
273 R	c 2.026	0.848	c 1.943	0.994	c 2.052	1.000
274 L	d 2.026	0.848	d 1.943	0.994	d 2.052	1.000
275 K	e 2.026	0.848	e 1.883	0.987	e 2.052	1.000
276 K	f 2.026	0.848	f 1.883	0.987	f 2.052	1.000
277 L	g 2.026	0.848	g 1.776	0.948	g 1.986	1.000
278 V	a 2.026	0.848	a 1.776	0.948	a 1.949	1.000
279 G	b 2.026	0.848	b 1.631	0.732	b 1.868	0.999
280 E	c 2.026	0.848	c 1.631	0.732	c 1.868	0.999
281 R	a 1.378	0.030	d 1.090	0.003	d 1.381	0.263

Figure 9.21. Continued.

always correctly predicted, as revealed by a lack of correlation between the predicted regions (H) and the known regions (\*).

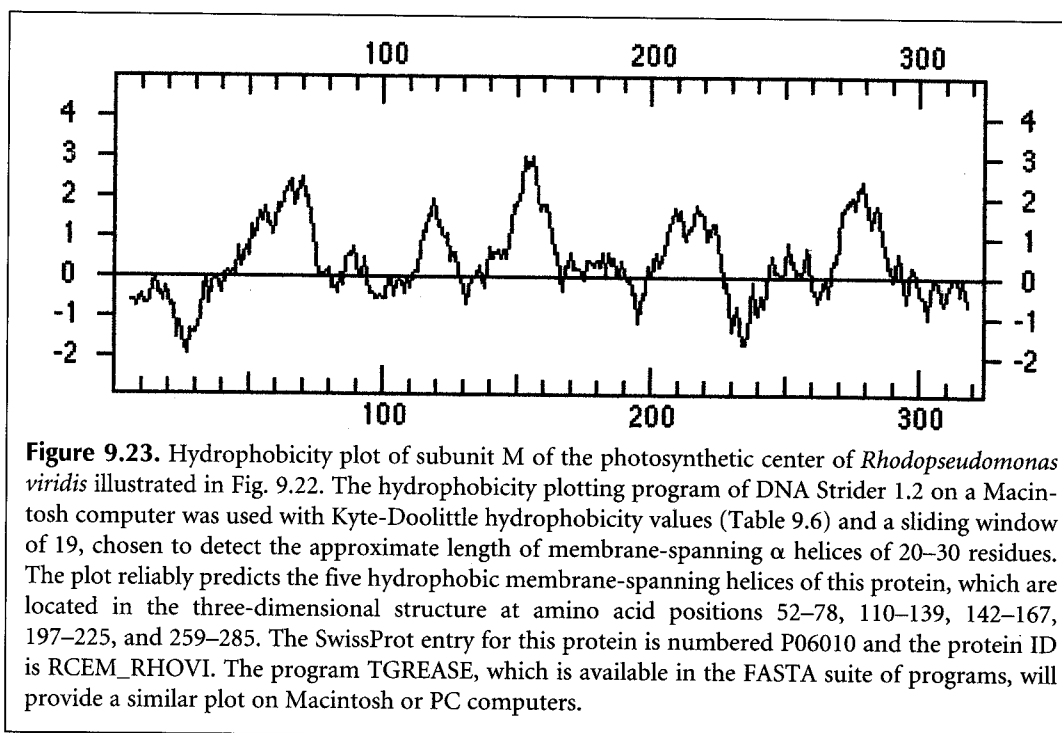
A second method for prediction of transmembrane  $\alpha$  helices is by the TMpred server. This method scans a candidate sequence for matches to a sequence scoring matrix obtained by aligning the sequences of all of the transmembrane  $\alpha$ -helical regions that are known from structures. These sequences have been collected into a database (TMbase) of such sequences. An example of a transmembrane analysis of 1prc\_M by this method is shown in Figure 9.25. As shown, the program correctly predicted five  $\alpha$ -helical transmembrane segments. Two alternative models were predicted, the first more highly favored, but neither one matched the known ends of these regions. These examples serve to illustrate that these methods can be expected to identify membrane-spanning  $\alpha$ -helical proteins quite reliably



but not the ends of such regions. A simple hydrophobicity plot may also be used as shown in Table 9.3. The number and extent of these regions can also be predicted from the peaks in this plot. This method is unsuitable for scanning genomic sequences for possible membrane-spanning proteins; the automatic methods are much more suitable for this purpose.

### Prediction of Protein Secondary Structure from the Amino Acid Sequence

Accurate prediction as to where  $\alpha$  helices,  $\beta$  strands, and other secondary structures will form along the amino acid chain of proteins is one of the greatest challenges in sequence analysis. At present, it is not possible to predict these events with very high reliability. As methods have improved, prediction has reached an average accuracy of 64–75% with a higher accuracy for  $\alpha$  helices, depending on the method used. These predictive methods

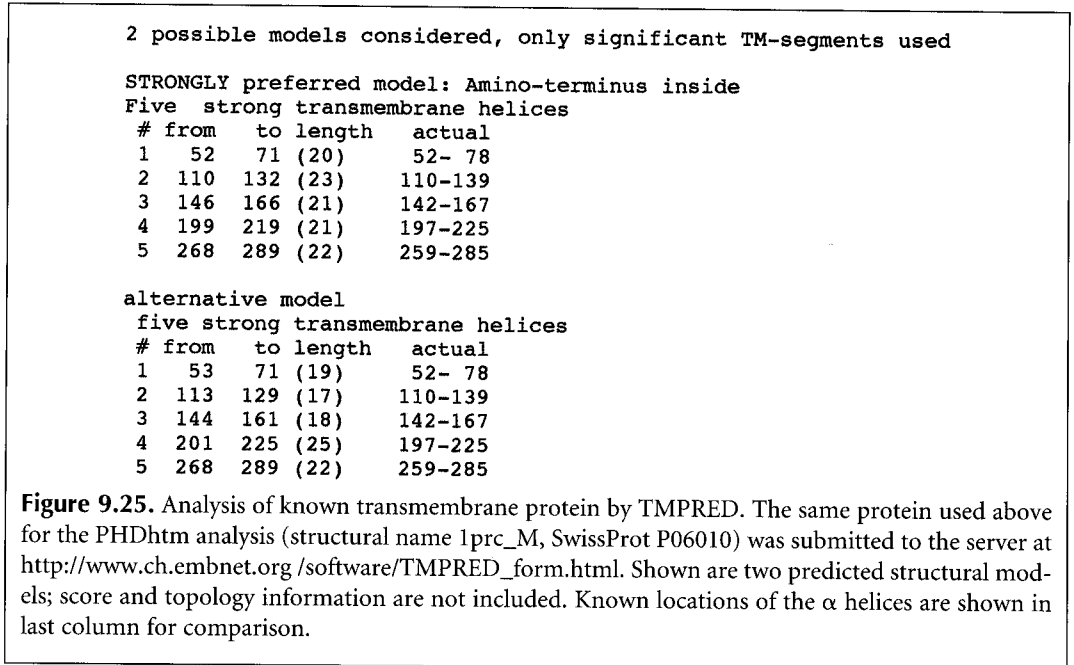


can be made especially useful when combined with other types of analyses discussed in this chapter. For example, a search of a sequence database or a protein motif database for matches to a candidate sequence may discover a family or superfamily relationship with a protein of known structure. If significant matches are found in regions of known secondary or three-dimensional structure, the candidate protein may share the three-dimensional structural features of the matched protein. Several Web sites provide such an enhanced analysis of secondary structure. These sites and others that provide secondary structure analysis of a query protein are given in Table 9.7. The main methods of analyses used at these sites are described below.

Methods of structure prediction from amino acid sequence begin with an analysis of a database of known structures. These databases are examined for possible relationships between sequence and structure. When secondary structure predictions were first being made in the 1970s and 1980s, only a few dozen structures were available. This situation has now changed with present databases including approximately 500 independent structural folds. The combination of more structural and sequence information presents a new challenge to investigators who wish to develop more powerful predictive methods.

The ability to predict secondary structure also depends on identifying types of secondary structural elements in known structures and determining the location and extent of these elements. The main types of secondary structures that are examined for sequence variation are  $\alpha$  helices and  $\beta$  strands. Early efforts focused on more types of structures, including other types of helices, turns, and coils. To simplify secondary structure prediction, these additional structures that are not an  $\alpha$  helix or  $\beta$  strand were subsequently classified as coils. Assignment of secondary structure to particular amino acids is sometimes included in the PDB file by the investigator who has solved the three-dimensional structure. In other cases, secondary structure must be assigned to amino acids by examination of the structural coordinates of the atoms in the PDB file. Methods for comparing three-dimensional structures, described above, frequently assign these features automatically, but not always





**Table 9.6.** *Hydrophobicity scales for the amino acids*

Residue		Value
Ala	A	1.8
Arg	R	-4.5
Asn	N	-3.5
Asp	D	-3.5
Cys	C	2.5
Gln	Q	-3.5
Glu	E	-3.5
Gly	G	-0.4
His	H	-3.2
Ile	I	4.5
Leu	L	3.8
Lys	K	-3.9
Met	M	1.9
Phe	F	2.8
Pro	P	-1.6
Ser	S	-0.8
Thr	T	-0.7
Trp	W	-0.9
Tyr	Y	-1.3
Val	V	4.2

These values are based on adjusted values derived from several sets of experimental measurements (Kyte and Doolittle 1982). The most hydrophobic amino acids are printed in green, the least hydrophobic amino acids in red. A number of additional scales are also available (von Heijne 1987).

**Table 9.7.** Selected programs for performing protein secondary structure prediction

Program	Web address	Method	Reference
Baylor College of Medicine (BCM)	<a href="http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html">http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html</a>	collection of methods and linked to other servers	see Web site and text
DSC	<a href="http://www.bmm.icnet.uk/dsc/">http://www.bmm.icnet.uk/dsc/</a>	linear discrimination	King et al. (1997)
J-Pred structure prediction server	<a href="http://jura.ebi.ac.uk:8888/">http://jura.ebi.ac.uk:8888/</a>	NNSSP, DSC, Predator, Mulpred, <sup>b</sup> Zpred, <sup>c</sup> Jnet, <sup>e</sup> and PHD	Cuff et al. (1998); and see text
NNPRED	<a href="http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html">http://www.cmpharm.ucsf.edu/~nomi/nnpredict.html</a>	neural networks enhanced to detect sequence periodicity	Kneller et al. (1990)
NPS@ server, MLR combination for secondary structure prediction <sup>a</sup>	<a href="http://pbil.ibcp.fr/NPSA/">http://pbil.ibcp.fr/NPSA/</a>	combination of prediction methods using multivariate linear regression to optimize the predictions	Guermeur et al. (1999)
Protein Sequence Analysis (PSA) System <sup>d</sup>	<a href="http://bmerc-www.bu.edu/psa/index.html">http://bmerc-www.bu.edu/psa/index.html</a>	discrete space models (hidden Markov models) for patterns of $\alpha$ helices, $\beta$ strands, tight turns, and loops in specific structural classes	Stultz et al. (1993, 1997); White et al. (1994)
PREDATOR	<a href="http://www.embl-heidelberg.de/argos/predator/predator_info.html">http://www.embl-heidelberg.de/argos/predator/predator_info.html</a>	based on analysis of long- and short-range amino acid interactions and alignments of sequence pairs	Frishman and Argos (1995, 1996, 1997)
Predict Protein server	<a href="http://www.embl-heidelberg.de/predictprotein/predictprotein.html">http://www.embl-heidelberg.de/predictprotein/predictprotein.html</a> ; see also mirror sites	neural networks of multiple sequence alignment	Rost and Sander (1994); Rost (1996)
PSSP	<a href="http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html">http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html</a>	nearest neighbor enhanced by non-intersecting local and multiple sequence alignments	Salamov and Solovyev (1995, 1997)
Simpa96	<a href="http://pbil.ibcp.fr/NPSA/">http://pbil.ibcp.fr/NPSA/</a>	nearest-neighbor method	Levin (1997)
SOPM, SOPMA	<a href="http://pbil.ibcp.fr/NPSA/">http://pbil.ibcp.fr/NPSA/</a>	nearest-neighbor method based on sequence alignments	Geourjon and Deleage (1994, 1995)
SSP	<a href="http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html">http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html</a>	linear discriminant analysis based on amino acid composition of local and adjacent regions	see H option for this program on Web page
UCLA-DOE structure prediction server	<a href="http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html">http://www.doe-mbi.ucla.edu/people/frsvr/frsvr.html</a>	collection of methods and linked to other servers	Fischer and Eisenberg (1996)

<sup>a</sup>Consensus option provides a user-defined combination of methods.

<sup>b</sup>See Cuff et al. (1998).

<sup>c</sup>Zpred server is also available at <http://kestrel.ludwig.ucl.ac.uk/zpred.html>. The program predicts secondary structure based on physicochemical information and GOR prediction scores.

<sup>d</sup>This server will also predict 3D structural class.

<sup>e</sup>Jnet uses multiple sequence alignments and a trained neural network to make secondary structure predictions (Cuff and Barton 2000).

and Sander 1984) and 8 (Sudarsanam 1998) residues in length can be found in different secondary structures. An 11-residue-long amino acid “chameleon” sequence has been found to form an  $\alpha$  helix when inserted into one part of a primary protein sequence and a  $\beta$  sheet when inserted into another part of the sequence (Minor and Kim 1996). More distant interactions may account for the observation that  $\beta$  strands are predicted more poorly by analysis of local regions (Garnier et al. 1996). However, the methods that have been used to predict the secondary structure of an amino acid residue all perform less well when amino acids more distant than in the small window of sequence are used.

The number of possible amino acid combinations in a sequence window of 17 amino acids is very large ( $17^{20} = 14 \times 10^{24}$ ). If many combinations influence one type of secondary structure, examination of a large number of protein structures is required to discover the significant patterns and correlations within this window. Earlier methods for predicting secondary structure assumed that each amino acid within the sequence window of 13–17 residues influences the local secondary structure independently of other nearby amino acids; i.e., there is no interaction between amino acids in influencing local secondary structure. Later methods assumed that interactions between amino acids within the window could play a role.

Neural network models described below have the ability to detect interactions between amino acids in a sequence window, including conditional interactions. A hypothetical example of the interactions that might be discovered illustrates the possibilities. If the central amino acid in the sequence window is Leu and if the second upstream amino acid toward the amino terminus is Asn, the Leu is in an  $\alpha$  helix; however, if the neighboring amino acid is not Asn, the Leu is in a  $\beta$  strand. In another method of secondary structure prediction, the nearest-neighbor method, sequence windows in known structures that are most like the query sequence are identified. This method bypasses the need to discover complex amino acid patterns associated with secondary structure. Protein secondary structure has also been modeled by hidden Markov models, also described as discrete state-space models, which are described below (Stultz et al. 1993; White et al. 1994).

### ***Accuracy of Secondary Structure Prediction***

One method of assessing accuracy of secondary structure prediction is to give the percentage of correctly predicted residues in sequences of known structure, called  $Q_3$ . This measure, however, is not very effective by itself, because even a random assignment of structure can achieve a high score by this test (Holley and Karplus 1991). Another measure is to report the fraction of each type of predicted structure that is correct. A third method is to calculate a correlation coefficient for each type of predicted secondary structure (Mathews 1975). The coefficient indicating success of predicting residues in the  $\alpha$ -helical configuration,  $C_\alpha$ , is given by

$$C_\alpha = (p_\alpha n_\alpha - u_\alpha o_\alpha) / \sqrt{([n_\alpha + u_\alpha][n_\alpha + o_\alpha][p_\alpha + u_\alpha][p_\alpha + o_\alpha])} \quad (2)$$

where  $p_\alpha$  is the number of correct positive predictions,  $n_\alpha$  is the number of correct negative predictions,  $o_\alpha$  is the number of overpredicted positive predictions (false positives), and  $u_\alpha$  is the number of underpredicted residues (misses). The closer this coefficient is to a value of 1, the more successful the method for predicting a helical residue. An overall level of prediction accuracy does not provide information on the accuracy of the number of predicted secondary structures, and their lengths and location in the sequence. One simple index of success is to compare the average of the predicted lengths with the known average (Rost and Sander 1993).

Another factor to consider in prediction accuracy is that some protein structures are more readily predictable than others, such that the spectrum of test proteins chosen will influence the frequency of success. A representative set of proteins that have limited similarity will provide the most objective test. Rost and Sander (1993) have chosen a set of 126 globular and 4 membrane proteins that have less than 25% pair-wise similarity and have used this set for training and testing neural network models. A newer set of 540 structurally



distinct fold types in the FSSP database provides an even larger set of training and test structures of unique structure and sequence (Holm and Sander 1998). In the often-used jackknife test, one protein in a set of known structure is left out of a calibration or training step of the program being tested. The rest of the proteins are used to predict the structure of the left-out one, and the procedure is cycled through all of the sequences. The overall frequency of success of predicting the secondary structural features of the left-out sequence is used as an indicator of success. An even more comprehensive approach to the problem of accuracy is to examine the predictions for different structural classes of proteins. Because some classes are much more difficult to predict, the overall success rate with respect to protein class is an important index of success. Prediction accuracy is discussed further below.

A valuable addition to secondary structure prediction is giving the degree of reliability of the prediction at each position. Some prediction methods produce a score for each of the three types of structures (helix, strand, coil or loop) at each residue position. If one of these scores is much higher than the other two, the score is considered to be more reliable, and a high reliability index may be assigned that reflects high confidence in the prediction. If the scores are more similar, the index is lower. By examining predictions for known structures, as in a jackknife experiment, the accuracy of these reliability indices may be determined. What has been found is that a prediction with a high index score is much more accurate (Yi and Lander 1993; and see PHD server below), thus increasing confidence in the prediction of these residues.

### ***Methods for Secondary Structure Prediction***

Three widely used methods of protein secondary structure prediction, (1) the Chou-Fasman and GOR methods, (2) neural network models, and (3) nearest-neighbor methods, are discussed below. An additional method that models structural families by hidden Markov models is then described. These methods can be further enhanced by examining the distribution of hydrophobic, charged, and polar amino acids in protein sequences.

#### ***Chou-Fasman/GOR Method***

The Chou-Fasman method (Chou and Fasman 1978) was based on analyzing the frequency of each of the 20 amino acids in  $\alpha$  helices,  $\beta$  sheets, and turns of the then-known relatively small number of protein structures. It was found, for example, that amino acids Ala (A), Glu (E), Leu (L), and Met (M) are strong predictors of  $\alpha$  helices, but that Pro (P) and Gly (G) are predictors of a break in a helix. A table of predictive values for each type of secondary structure was made for each of the  $\alpha$  helices,  $\beta$  strands, and turns. To produce these values, the frequency of amino acid  $i$  in structure  $s$  is divided by the frequency of all residues in structure  $s$ . The resulting three structural parameters ( $P_{\alpha}$ ,  $P_{\beta}$ , and  $P_t$ ) vary roughly from 0.5 to 1.5 for the 20 amino acids.

To predict a secondary structure, the following set of rules is used. The sequence is first scanned to find a short sequence of amino acids that has a high probability for starting a nucleation event that could form one type of structure. For  $\alpha$  helices, a prediction is made when four of six amino acids have a high probability  $>1.03$  of being in an  $\alpha$  helix. For  $\beta$  strands, the presence in a sequence of three of five amino acids with a probability of  $>1.00$  of being in a  $\beta$  strand predicts a nucleation event for a  $\beta$  strand. These nucleated regions are extended along the sequence in each direction until the prediction values for four amino acids drops below 1. If both  $\alpha$ -helical and  $\beta$ -strand regions are predicted, the higher probability prediction is used.

Turns are predicted somewhat differently. Turns are modeled as a tetrapeptide, and two probabilities are calculated. First, the average of the probabilities for each of the four amino acids being in a turn is calculated as for  $\alpha$  helix and  $\beta$  strand predictions. Second, the probabilities of amino acid combinations being present at each position in the turn tetrapeptide (i.e., the probability that a particular amino acid such as Pro is at position 1, 2, 3, or 4 in the tetrapeptide) are determined. These probabilities for the four amino acids in the candidate sequence are multiplied to calculate the probability that the particular tetrapeptide is a turn. A turn is predicted when the first probability value is greater than the probabilities for an  $\alpha$  helix and a  $\beta$  strand in the region and when the second probability value is greater than  $7.5 \times 10^{-5}$ . In practice, the Chou-Fasman method is only about 50–60% accurate in predicting secondary structural domains.

Garnier et al. (1978) developed a somewhat more involved method for protein secondary structure prediction that is based on a more sophisticated analysis. The method is called the GOR (Garnier, Osguthorpe, and Robson) method. Whereas the Chou-Fasman method is based on the assumption that each amino acid individually influences secondary structure within a window of sequence, the GOR method is based on the assumption that amino acids flanking the central amino acid residue influence the secondary structure that the central residue is likely to adopt. In addition, the GOR method uses principles of information theory to derive predictions (Garnier et al. 1996).

As in the Chou-Fasman method, known secondary structures are scanned for the occurrence of amino acids in each type of structure. However, the frequency of each type of amino acid at the next 8 amino-terminal and carboxy-terminal positions is also determined, making the total number of positions examined equal to 17, including the central one. In the original GOR method, three scoring matrices, containing in each column the probability of finding each amino acid at one of the 17 positions, are prepared. One matrix corresponds to the central (eighth) amino acid being found in an  $\alpha$  helix, the second for the amino acid being in a  $\beta$  strand, the third a coil, and the fourth, a turn. Later versions omitted the turn calculation because these were the most variable features and were consequently the most difficult to predict. A candidate sequence is analyzed by each of the three to four matrices by a sliding window of 17 residues. Each matrix is positioned along a candidate sequence and the matrix giving the highest score predicts the structural state of the central amino acid. At least 4 residues in a row have to be predicted as an  $\alpha$  helix and 2 in a row for a  $\beta$  strand for a prediction to be validated.

Matrix values are calculated in somewhat the same manner as amino acid substitution matrices (described in Chapter 3), in that matrix values are calculated as log odds units representing units of information. The information available as to the joint occurrence of secondary structural conformation  $S$  and amino acid  $a$  is given by (Garnier et al. 1996)

$$I(S; a) = \log [ P(S | a) / P(S) ] \quad (3)$$

where  $P(S | a)$  is the conditional probability of conformation  $S$  given residue  $a$ , and  $P(S)$  is the probability of conformation  $S$ . By Bayes' rule (see Chapter 3, p. 120), the probability of conformation  $S$  given amino acid  $a$ ,  $P(S | a)$  is given by

$$P(S | a) = P(S, a) / P(a) \quad (4)$$

where  $P(S, a)$  is the joint probability of  $S$  and  $a$  and  $P(a)$  is the probability of  $a$ . These probabilities can be estimated from the frequency of each amino acid found in each structure and the frequency of each amino acid in the structural database. Given these frequencies,

$$I(S; a) = \log(f_{S,a} / f_S) \quad (5)$$

where  $f_{S,a}$  is the frequency of amino acid  $a$  in conformation  $S$  and  $f_S$  is the frequency of all amino acid residues found to be in conformation  $S$ .

The GOR method maximizes the information available in the values of  $f_{S,a}$  and avoids data size and sampling variations by calculating the information difference between the competing hypotheses that residue  $a$  is in structure  $S$ ,  $I(S; a)$ , or that  $a$  is in a different conformation (not  $S$ ),  $I(\text{not } S; a)$ . This difference  $I(\Delta S; a)$  is calculated from Equation 5 with simple substitutions by

$$\begin{aligned} I(\Delta S; a) &= I(S; a) - I(\text{not } S; a) \\ &= \log\{P(S,a)/[1 - P(S,a)]\} + \log\{[1 - P(S)]/P(S)\} \end{aligned} \quad (6)$$

which is derived from the observed amino acid data as

$$I(\Delta S; a) = \log[f_{S,a} / (1 - f_{S,a})] + \log[(1 - f_S) / f_S] \quad (7)$$

where the frequency of finding amino acid  $a$  not in conformation  $S$  is  $1 - f_{S,a}$  and of not finding any amino acid in conformation  $S$  is  $1 - f_S$ . Equation 6 is used to calculate the information difference for a series of  $x$  consecutive positions flanking sequence position  $m$ ,

$$I(\Delta S_m; a_1, \dots, a_x) = \log\{P(S_m, a_1, \dots, a_x) / [1 - P(S_m, a_1, \dots, a_x)]\} + \log\{[1 - P(S)] / P(S)\} \quad (8)$$

from which the following ratio of the joint probability of conformation  $S_m$  given  $a_1, \dots, a_x$  to the joint probability of any other conformation may be calculated

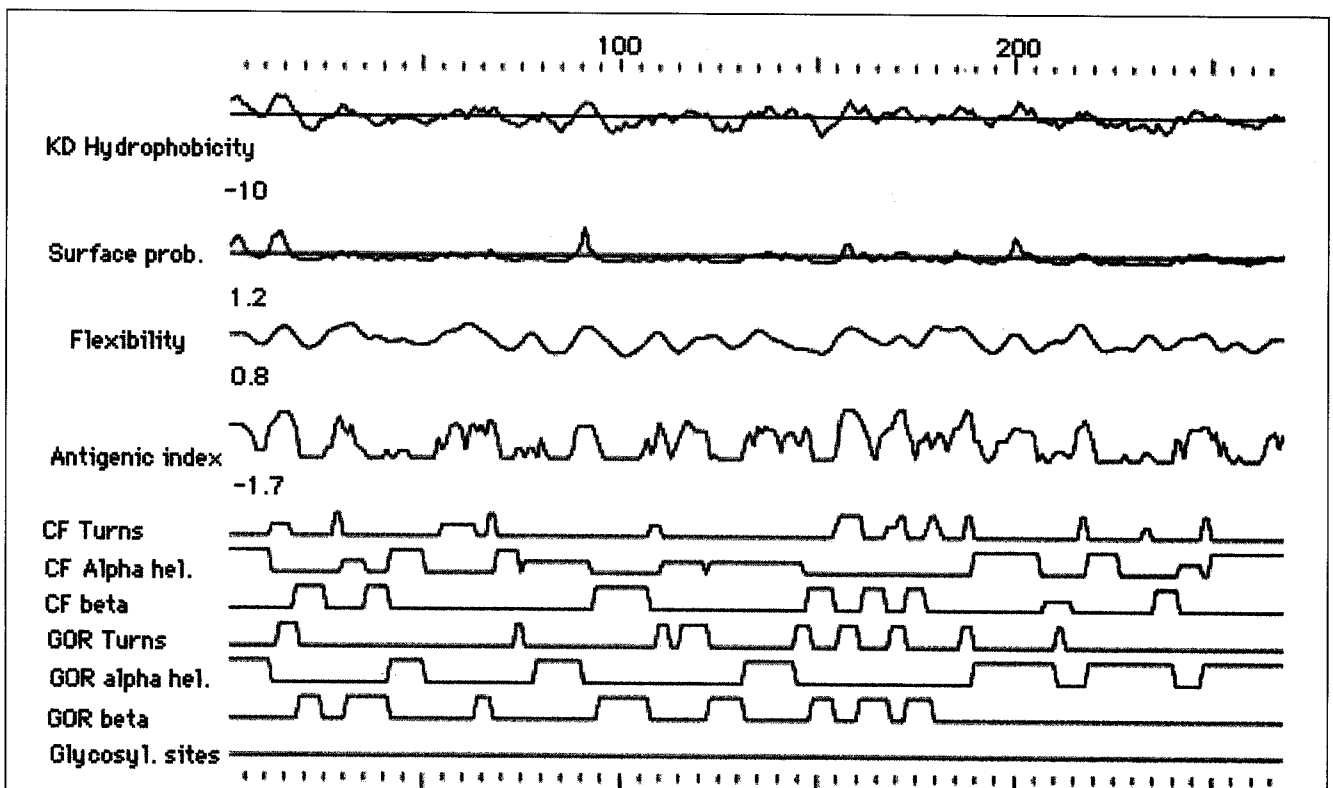
$$P(S_m, a_1, \dots, a_x) / [1 - P(S_m, a_1, \dots, a_x)] = \{P(S) / [1 - P(S)]\} e^{-I(\Delta S_m; a_1, \dots, a_x)} \quad (9)$$

Searching for all possible patterns in the structural database would require an enormous number of proteins. Hence, three simplifying approaches have been taken. First, it was assumed in earlier versions of GOR that there is no correlation between amino acids in any of the 17 positions (both the flanking 8 positions and the central amino acid position), or that each amino acid position had a separate and independent influence on the structural conformation of the central amino acid. The steps are then: (1) Values for  $I(\Delta S; a)$  in Equation 7 are calculated for each of the 17 positions; (2) these values are summed to approximate the value of  $I(\Delta S_m; a_1, \dots, a_x)$  in Equation 8; (3) the probability ratios in Equation 9 are calculated.

The second assumption used in later versions of GOR was that certain pair-wise combinations of an amino acid in the flanking region and central amino acid influence the conformation of the central amino acid. This model requires a determination of the frequency of amino acid pairs between each of the 16 flanking positions and the central one, both for when the central residue is in conformation  $S$  and when the central residue is not in conformation  $S$ . Finally, in the most recent version of GOR, the assumption is made that certain pair-wise combinations of amino acids in the flanking region, or of a

flanking amino acid and the central one, influence the conformation of the central one. Thus, there are  $17 \times 16/2 = 136$  possible pairs to use for frequency measurements and to examine for correlation with the conformation of the central residue. With the advent of a large number of protein structures, it has become possible to assess the frequencies of amino acid combinations and to use this information for secondary structural predictions. The GOR method predicts 64% of the residue conformations in known structures and quite drastically (36.5%) underpredicts the number of residues in  $\beta$  strands.

Use of the Chou-Fasman and GOR methods for predicting the secondary structure of the  $\alpha$  subunit of *Salmonella typhimurium* tryptophan synthase is illustrated in Figure 9.26. In this particular case, the positions of the secondary structures predicted by either of these methods are very similar to those in the solved crystal structure (Branden and Tooze 1991). However, tests of the accuracy of these methods using sequences of other proteins whose structures are known have shown that the Chou-Fasman method is only about 50–60% accurate in predicting the structural domains. The methods are most useful in the hands of a knowledgeable structural biologist, and have been used most successfully in polypeptide design and in analysis of motifs for organelle transport (Branden and Tooze 1991). A useful approach is to analyze each of a series of aligned amino acid sequences and then to derive a consensus structural prediction.



**Figure 9.26.** Example of the secondary structure predictions for the  $\alpha$  subunit of *S. typhimurium* tryptophan synthase by the Chou-Fasman and GOR methods included in the Genetics Computer Group suite of programs. The predictions are shown on the lower panels, labeled as CF for the Chou-Fasman method (Chou and Fasman 1978) and GOR (referred to as GOR I) for the Garnier, Osguthorpe, and Robson method (Garnier et al. 1978). This protein is in the  $\alpha$ - $\beta$  class with an  $\alpha/\beta$  barrel type of structure comprising eight parallel  $\beta$  strands and eight  $\alpha$  helices in an alternating pattern and three additional  $\alpha$  helices, and is shown in Fig. 9.6. The predicted structure is quite accurate and represents the correct pattern of secondary structure.

### ***Patterns of Hydrophobic Amino Acids Can Aid Structure Prediction***

Prediction of secondary structure can be aided by examining the periodicity of amino acids with hydrophobic side chains in the protein chain. This type of analysis was discussed above in the prediction of transmembrane  $\alpha$ -helical domains in proteins. Hydrophobicity tables that give hydrophobicity values for each amino acid are used to locate the most hydrophobic regions of the protein (Table 9.6) (see Lüthy and Eisenberg 1991). As for secondary structure prediction, a sliding window is moved across the sequence and the average hydrophobicity value of amino acids within the window is plotted. A hydrophobicity plot of the  $\alpha$  subunit of *S. typhimurium* tryptophan synthase is included in the first panel of Figure 9.26.

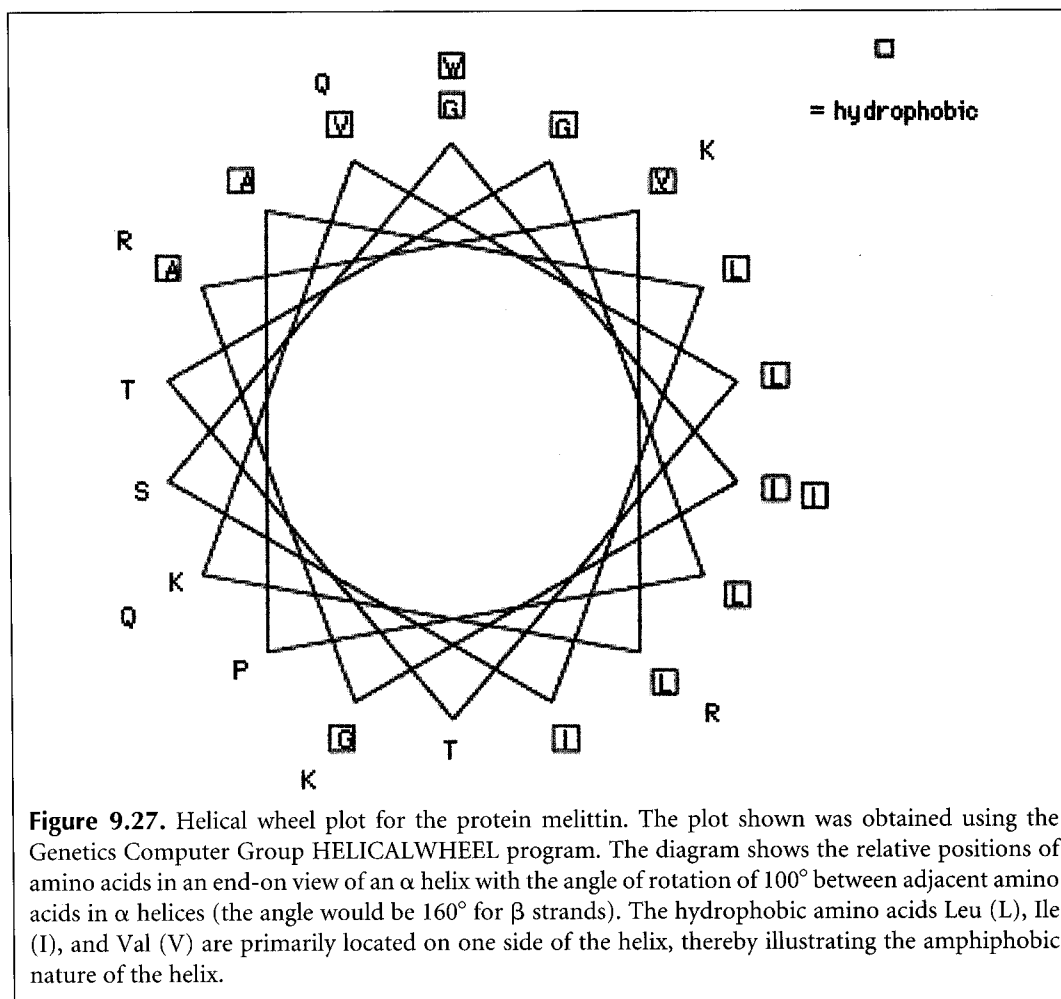
Similar methods for predicting surface peptides including antigenic sites, chain flexibility, or glycosylation sites are also illustrated in Figure 9.26. These methods use the chemical properties of amino acid side chains to predict the location of these amino acids on the surface or buried within the core structure.

The location of hydrophobic amino acids within a predicted secondary structure can also be used to predict the location of the structure. One type of display of this distribution is the helical wheel or spiral display of the amino acids in an  $\alpha$  helix, as shown in Figure 9.27. This use of this display was described above as a way to visualize the location of leucine residues on one face of the helix in a leucine zipper structure. There is also a tendency of hydrophobic residues located in  $\alpha$  helices on the surface of protein structures to face the core of the protein and for polar and charged amino acids to face the aqueous environment on the outside of the  $\alpha$  helix. This arrangement is also revealed by the helical wheel display shown in Figure 9.27. Another type of display, the hydrophobic moment display, is shown in Figure 9.28. The contours in this plot show positions in the amino acid sequence where hydrophobic amino acids tend to segregate to opposite sides of a structure plotted against various angles of rotation from one residue to the next along the protein chain. For  $\alpha$  helices, the angle of rotation is 100 degrees and for  $\beta$  strands, 160 degrees. The analysis in the figure predicts, for example, an  $\alpha$  helix at approximate sequence position 165 that has segregated hydrophobic amino acids on one helix face. Helix  $\alpha$ 5 runs from positions 160 to 168 in the crystal structure of this protein.

### ***Secondary Structure Prediction by Neural Network Models***

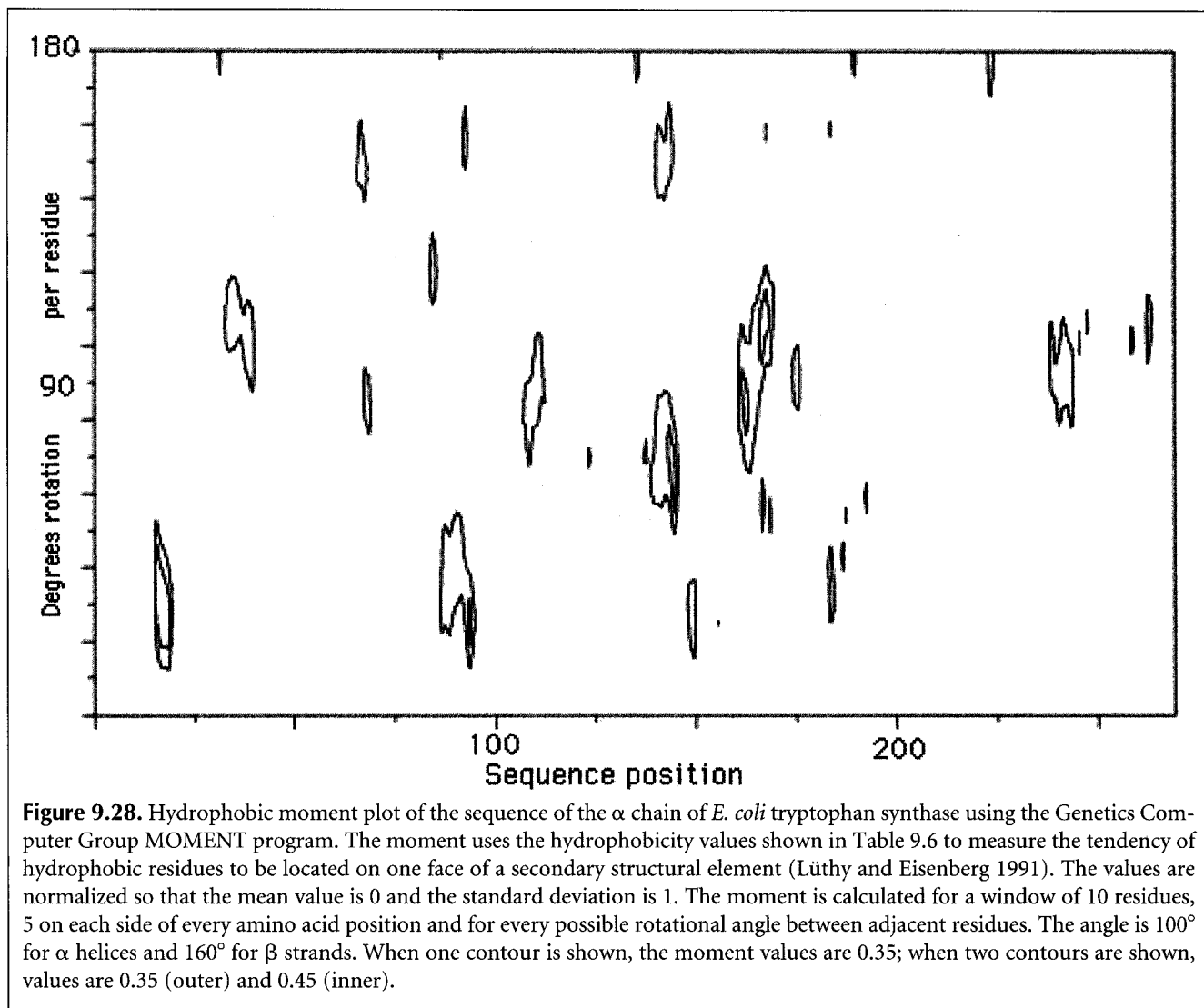
The most sophisticated methods that have been devised to make secondary structural predictions for proteins use artificial intelligence, or so-called neural net algorithms. An earlier method of this type examined patterns that represent secondary structural features like the Chou-Fasman method. However, this method went farther and tried to locate these patterns in a particular order that coincides with a known domain structure. Patterns typical of  $\alpha/\beta$  proteins (Cohen et al. 1983), turns in globular proteins (Cohen et al. 1986), or helices in helical proteins (Presnell et al. 1992) may be located and used to predict secondary structure with increased confidence. The program MACMATCH, which combines these methods with a neural network approach to predict the secondary structure of globular proteins on a Macintosh computer, has been described (Presnell et al. 1993).

In the neural network approach, computer programs are trained to be able to recognize amino acid patterns that are located in known secondary structures and to distinguish these patterns from other patterns not located in these structures. There are many examples of the use of this method to predict protein structures (see, e.g., Qian and Sejnowski 1988; Muggleton et al. 1992; Stolorz et al. 1992; Rost and Sander 1993), which have been reviewed (Holley and Karplus 1991; Hirst and Sternberg 1992). The early methods are reported to be up to 63–64% accurate. These methods have been improved to a level of over 70% for globular proteins by the use of information from multiple sequence alignments (Rost and Sander 1993, 1994). Two Web sites that perform a neural network analysis for protein secondary structure prediction are PHD (Rost and Sander 1993; Rost 1996; <http://www.embl->



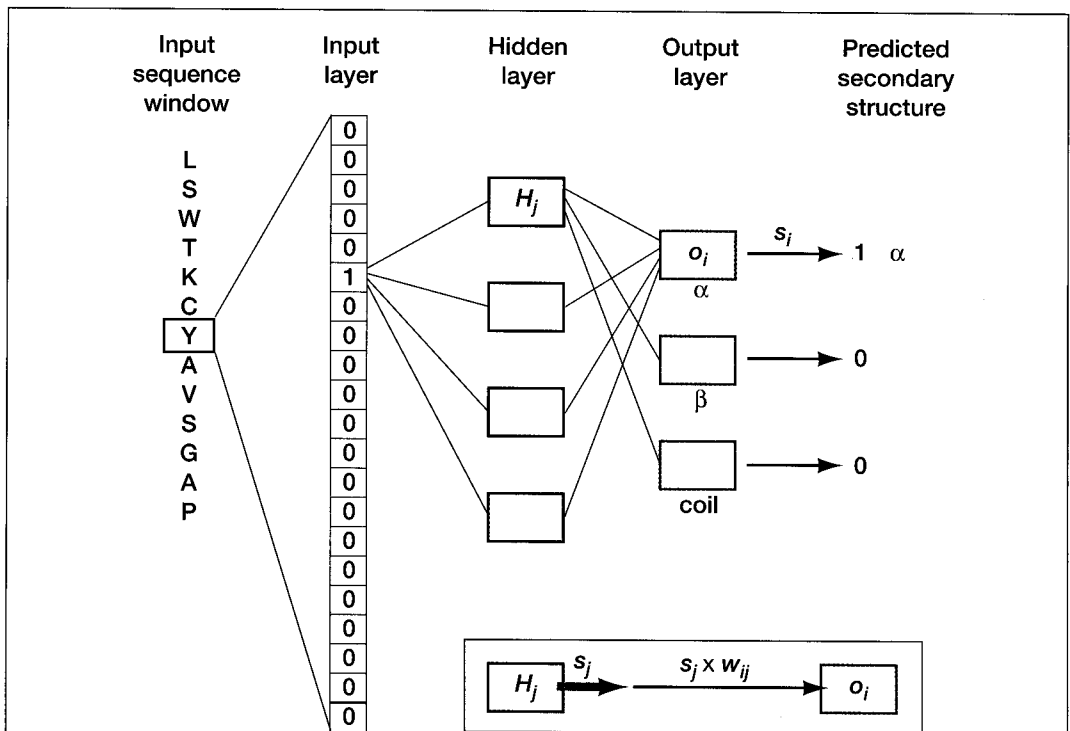
heidelberg.de/predictprotein/predictprotein.html) and NNpredict (Kneller et al. 1990; <http://www.cmp pharm.ucsf.edu/~nomi/nnpredict.html>). These neural network models are theoretically able to extract more information from sequences than the information theory method described above (Qian and Sejnowski 1988). Neural networks have also been used to model translational initiation sites and promoter sites in *E. coli*, splice junctions, and specific structural features in proteins, such as  $\alpha$ -helical transmembrane domains. These applications are discussed elsewhere in this chapter and in Chapter 8.

Neural network models are meant to simulate the operation of the brain. The complex patterns of synaptic connections among a large number of neurons are presumed to underlie the functions of the brain. Some groups of neurons are involved in collecting data as environmental signals, others in processing data, and yet others in providing a response to the signals. Neural networks are an attempt to build a similar kind of learning machine where the input is a 13–17-amino-acid length of sequence and the output is the predicted secondary structure of the central amino acid residue. The object is to train the neural network to respond correctly to a set of such flanking sequence fragments when the secondary structural features of the centrally located amino acid are known. The training is designed to achieve recognition of amino acid patterns associated with secondary structure. If the neural network has sufficient capacity for learning, these patterns may potentially include complex interactions among the flanking amino acids in determining secondary structures. However, two studies with neural networks described below have so far not found evidence for such interactions.



A typical neural network model used for protein secondary structure prediction is illustrated in Figure 9.29. A sliding window of 13–17 amino acid residues is moved along a sequence. The sequence within each window is read and used as input to a neural network model previously trained to recognize the secondary structure most likely to be associated with that pattern. The model then predicts the secondary structural configuration of the central amino acid as  $\alpha$  helix,  $\beta$  strand, or other. Rules or another trained network are then applied that make the prediction of a series of residues reasonable. For example, at least 4 amino acids in a row should be predicted as being in an  $\alpha$  helix if the prediction is to make structural sense.

The model comprises three layers of processing units—the input layer, the output layer, and the so-called hidden layer between these layers. Signals are sent from the input layer to the hidden layer and from the hidden layer to the output layer through junctions between the units. This configuration is referred to as a feed-forward multilayer network. The input layer of units reads the sequence, one unit per amino acid residue, and transmits information on the amino acid at that location. A small window of sequence is read at a time and information is sent as signals through junctions to a number of sequential units in the hidden layer by all of the input units within the window, as shown by the lines joining units in Figure 9.29. These signals are each individually modified by a weighting factor and then



**Figure 9.29.** A typical neural network model for protein secondary structure prediction (after Rost and Sander 1993). Functions of the input (red boxes), hidden (blue boxes), and output (green boxes) layers are described in the text. There is one input unit for each amino acid in the sequence window of 13 (first column). Each input amino acid unit is made up of 21 input positions, one for each amino acid and one for a padding space when the window overlaps the end of the sequence. Other positions may be added to provide additional information. The positions each send information to the hidden unit layer. In a simple input coding system, only one of the 20 components in a given input unit has a value of 1. Shown is an example where the component for Y is turned on while the rest of the components are 0 (second column). When padding for the end of sequence is required, only the padding space is set to 1. When a sequence profile is used as input (not shown), each position is filled with the frequency of the amino acid in the corresponding column of the sequence profile or with a coded form of this frequency, and the numbers of insertions and deletions are added in two extra positions. Another position is used to indicate the amount of information due to the presence of conserved amino acids in the column. Signals from each position in each input unit are weighted as they proceed to units of the hidden layer. A signal from a component of one input unit will receive a different weight for each connection to a hidden unit. Each hidden unit sums the signals ( $s_{in}$ ) received from the input layer and then transforms the sum using the trigger function  $s_{out} = 1/(1 + e^{-ksin})$  to produce an output signal that is between and close to either 0 and 1, simulating the firing of a neuron. Strong signals are transformed by this function to a number approximately equal to 1 and weak or negative values to 0. As the constant  $k$  increases, discrimination between strong and weak signals is increased. The hidden layer output signals are weighted and sent to three output units, representing prediction of an  $\alpha$  helix,  $\beta$  strand, or coil (loop) for the secondary structural configuration of the central amino acid in the window. The sum of these signals is transformed to values between 0 and 1. An output signal close to 1 is a prediction for the amino acid to have the corresponding structural configuration; a weak signal close to zero is no prediction. The example shown predicts an  $\alpha$ -helical configuration for Y. Predictions for a series of adjacent windows are sorted out by applying rules or by additional neural networks. The insert illustrates the operation of the back-propagation algorithm that is used to train the network and is described by an example in the text (p. 455).

Information content of an alignment is discussed in Chapter 4, page 195.



added together to give a total input signal into each hidden unit. Sometimes a bias is added to this sum to influence the response of the unit. The resulting signal is then transformed by the hidden unit into a number that is very close either to a 1 or to a zero (or sometimes to a  $-1$ ). A mathematical function known as a sigmoid trigger function, simulating the firing or nonfiring states of a neuron, is used for this transformation. Signals from the hidden units are then sent to three individual output units, each output unit representing one type of secondary structure (helix, strand, or other). Each signal is again weighted, the input signals are summed, and each of the three output units then converts the combined signal into a number that is approximately a 1 or a 0. An output signal that is close to 1 represents a prediction of the secondary structural feature represented by that output unit and a signal near to the value 0 means that the structure is not predicted.

When hidden layers are included, a neural network model is capable of detecting higher levels of interaction among amino acids that influence secondary structure. For example, particular combinations of amino acids may produce a particular type of secondary structure. To resolve these patterns, a sufficient number of hidden units is needed (Holley and Karplus 1991); the number varies from 2 to a range of 10–40. An interesting side effect of adding more hidden units is that the neural network memorizes the training set but at the same time is less accurate with test sequences. This effect is revealed by using the trained network to predict the same structures used for training. The number correct increases by over 20% as the number of hidden units increases from 0 to 10. In contrast, accuracy of prediction of test sequences not used for training decreased 3% (Holley and Karplus 1991).

Without hidden layers, the neural network model is known as a perceptron, and has a more limited capacity to detect such combinations. In two studies, networks with no hidden units were as successful in predicting secondary structure as those with hidden units. In addition, the number of hidden units was increased to as many as 60 in one study (Qian and Sejnowski 1988) and 20 in another (Holley and Karplus 1991) without significantly changing the level of success. These observations imply that the influence of local sequence on secondary structure is the additive influence of individual residues and that there is no higher level of interaction among these residues. To detect such interactions, however, requires a large enough training set to provide a significant number of examples, and these conditions may not have been met. These same studies examined the effect of input window size and found that a maximum information for secondary structure prediction seems to be located within a window of 13–17 amino acids, as larger windows do not increase accuracy. However, small windows were less effective, suggesting that they have insufficient information, and below a window size of 5, success at predicting  $\beta$  strands was decreased.

Training the neural network model is the process of adjusting the values of the weights used to modify the signals from the input layer to the hidden layer and from the hidden layer to the output layer. The object is to have these weights balance the input signals so that the model output correctly identifies the known secondary structure of the central amino acid in a sequence window of a protein of known structure. Because there may be thousands of connections between the various units in the network, a systematic method is needed to adjust these values. Initially, the weights are assigned a constant or random value (typical range  $-0.1$  to  $+0.1$ ). The sliding window is then positioned along one of the training sequences. The predicted output for a given sequence window is then compared to the known structure of the central amino acid residue. The model is adjusted to increase the chance of predicting the correct residue. The adjustment involves changing the weighting of propagated signals by a method called the back-propagation algorithm. This procedure is repeated for all windows in all of the training sequences. The better the model, the more predicted structures that will be correct. Conversely, the worse the model, the more predictions that will be incorrect. The object then becomes to minimize this incorrect number. The error  $E$  is expressed as the square of the total number of incorrect predictions by the output units.

*Use of a perceptron for analyzing regulatory DNA sequences is illustrated in Figure 8.11 (p. 363).*

When the back-propagation algorithm is applied, the weights are adjusted by a small amount to decrease errors. A window of a training sequence is used as input to the network, and the predicted and expected (known) structures of the central residue are compared. A set of small corrections is then made to the weights to improve an incorrect prediction, or the weights are left relatively unchanged for a correct prediction. This procedure is repeated using another training sequence until the number of errors cannot be reduced further. A large number of training cycles representing a slow training rate is an important factor for training the network to produce the smallest number of incorrect predictions. Not all of the training sequences may be used—a random input of training patterns may be used and sometimes these may be chosen from subsets of sequences that represent one type of secondary structure to balance the training for each type of structure. The back-propagation algorithm examines the contribution of each connection in the network on the subsequent levels and adjusts the weight of this connection, if needed to improve the predictions. The following example illustrates the operation of the algorithm.

**Example: Back-propagation Algorithm Used to Train the Neural Network (Rost and Sander 1993)**

Consider an output unit  $O_i$  as shown in Figure 9.29. Let us assume that this unit predicts whether or not the central residue in the scanned sequence window is an  $\alpha$ -helical secondary structure. The output signal from this unit is  $s_i$  which, if close to 1, predicts an  $\alpha$ -helical structure or, if close to 0, does not predict an  $\alpha$  helix. The network has been provided with a training sequence and it is known whether or not the central amino acid actually is found in an  $\alpha$  helix. If the structure is an  $\alpha$  helix, then the output of  $O_i$  should be close to 1, and if not, then close to zero.  $d_i$  is the expected or desired output of  $O_i$  and  $d_i = 1$  if a helix is expected and 0 if not. The output of  $O_i$  is determined by the sum of the inputs received from each of the hidden units with which  $O_i$  is connected. The hidden units each emit a signal close to 0 or 1, and each signal is separately weighted as it passes from the hidden unit to  $O_i$ . Focus on one of the hidden units  $H_j$  that is connected to  $O_i$  and emits a signal  $s_j$  that is modified by weight  $w_{ij}$ . The signal arriving at  $O_i$  is thus  $s_j \times w_{ij}$ , as illustrated in the insert in Figure 9.29. The problem at hand is to adjust or not to adjust  $w_{ij}$  so that the output of  $O_i$  ( $s_i$ ) is close to the desired value,  $d_i$ . The value of  $w_{ij}$  is adjusted according to a procedure known as gradient descent that is given by the formula

$$\Delta w_{ij} = w_{ij} - n \partial E / \partial w_{ij} + m \quad (10)$$

where the partial derivative of the error  $E$  with respect to  $w_{ij}$ ,  $\partial E / \partial w_{ij}$ , is calculated by

$$\partial E / \partial w_{ij} = (s_i - d_i) s_i (1 - s_i) s_j \quad (11)$$

and where  $n$  is the rate of training (typical value 0.03) and  $m$  is a smoothing factor that allows a carryover of a fraction of previous values of  $w_{ij}$  (typical value 0.2). Suppose, for example, that  $s_j$  was sent from  $H_j$  to  $O_i$  as 0.2 and that  $d_i$  is 1, so that  $s_j$  is not contributing correct information. Then  $\partial E / \partial w_{ij} = (0.2 - 1) \times 0.2 \times 0.8 \times 0.2 = -0.0256$ .  $w_{ij}$  will then be increased in Equation 10 by the rate of training times this value adjusted by  $m$  for contributions from any previous value of  $w_{ij}$ . Adjusting the weights of connections between the input and hidden layers uses a more detailed formula that takes into account the effects of both the signal sent from the input unit to the hidden units and that of the hidden unit on each of the output units.





structure of the middle amino acid in each of these matching fragments ( $f_\alpha$ ,  $f_\beta$ , and  $f_{\text{coils}}$ ) are then used to predict the secondary structure of the middle amino acid in the query window. As with other secondary structure prediction programs, the predicted secondary structure of a series of residues in the query sequence is subjected to a set of rules or used as input to a neural network to make a final prediction for each amino acid position.

Although not implemented in the most available programs, a true estimate of probability of the above set of frequencies may be obtained by identifying sets of training sequences that give the same value of  $(f_\alpha + f_\beta + f_{\text{coils}})^{1/2}$ . The frequencies of the secondary structures predicted by this group then give true estimates for  $p_\alpha$ ,  $p_\beta$ , and  $p_{\text{coils}}$  for the targeted amino acid in the query sequence (Yi and Lander 1993). Predictions based on the highest probabilities have been shown to be the most accurate, with the top 28% of the predictions being 86% accurate and the top 43% being 81% accurate. In addition, this method of calculating probability possesses more information than single-state predictions. Using this method, therefore, a substantial proportion of protein secondary structures can be predicted with high accuracy (Yi and Lander 1993, 1996).

The several nearest-neighbor programs that have been developed for secondary structure prediction (see Table 9.7) differ largely in the method used to identify related sequences in the training set. Originally, an amino acid scoring matrix such as a BLOSUM scoring matrix was used (Zhang et al. 1992). Distances between sequences based on a statistical analysis of the training sequences have also been proposed (Salzberg and Cost 1992). Use of a scoring matrix (Bowie et al. 1991, 1996) based on a categorization of amino acids into local structural environments, discussed below, in conjunction with a standard amino acid scoring matrix increased the success of the predictions (Yi and Lander 1993; Salamov and Solovyev 1995, 1997). Yet further increases in success have been achieved by aligning the query sequence with the training sequences to obtain a set of nonintersecting alignments with windows of the query sequence (as described in Chapter 3, p. 75), and of using a multiple sequence alignment as input with amino-terminal and carboxy-terminal positions of  $\alpha$  helices and  $\beta$  strands and  $\beta$  turns treated as distinctive types of secondary structure (Salamov and Solovyev 1997).

The program PREDATOR (Table 9.7) is based on an analysis of amino acid patterns in structures that form H-bond interactions between adjacent  $\beta$  strands ( $\beta$  bridges) and between amino acid  $n$  and  $n + 4$  on  $\alpha$  helices (Frishman and Argos 1995, 1996). The H-bond pattern between parallel and antiparallel  $\beta$  strands is different (Fig. 9.3) and two types of antiparallel patterns have been recognized. By utilizing such information combined with substitutions found in sequence alignments, the prediction success of PREDATOR has been increased to 75% (Frishman and Argos 1997). Examples of the NNSSP (Salamov and Solovyev 1997) and PREDATOR (Frishman and Argos 1997) program outputs are given on page 459.

**Example: NNSSP and PREDATOR Output**

Two of the most accurate nearest-neighbor prediction programs are (1) NNSSP (accuracy to 73.5%) shown is the program output from <http://dot.imgen.bcm.tmc.edu:9331/seq-search/struc-predict.html>, choosing the PSSP/NNSSP option. PredSS is the predicted secondary structure by NNSSP (a =  $\alpha$ ; b =  $\beta$ ; c = COILS). The output probabilities Prob a and Prob b give a normalized score by converting the values of  $f_{\alpha}$ ,  $f_{\beta}$ , and  $f_{\text{coils}}$  to a scale of 0–9. (2) PREDATOR (accuracy 75%) applies the FSSP assignments of secondary structure to the training sequences. PREDATOR does not provide a normalized score. PREDATOR predictions from [http://www.embl-heidelberg.de/argos/predator/predator\\_info.html](http://www.embl-heidelberg.de/argos/predator/predator_info.html) are shown below NNSSP prediction on each line (H =  $\alpha$ ; E =  $\beta$ ). The input sequence was the  $\alpha$  subunit of *S. typhimurium* tryptophan synthase (SwissProt ID TRPA\_SALTY, accession P00929), which is in the training sequences because the three-dimensional structure is known.

```

nnssp Sat Mar 13 15:49:19 CST 1999
TS_subunit_alpha
L= 268 SS content: a= 0.56 b= 0.08 c= 0.36
                10      20      30      40      50
PredSS          aaaaaaaaaa   bbbbbb   aaaaaaaaaaaaaaaaaa
AA seq          MERYESLFAQLKERKEGAFVFPVTLGDPGIEQSLKIIDLIEAGADALEL
Prob a          9999999999997421110000001000168889999999974578863
Prob b          00000000000000012777887410001000000000000001122
Predator        HHHHHHHHHHHH EEEEE HHHHHHHHHH

                60      70      80      90      100
PredSS          aaaaaaaaaa   aaaaaaaaaa   bbba
AA seq          GIPFSDPLADGPTIQNATLRAFAAGVTPAQCPEMLALIRQKHTIPIGILL
Prob a          111111110012456889988731105889999999852000111133
Prob b          232211011000121100000011110000000000000002335544
Predator        HHHHHHHHHH HHHHHHHHHH HHHH

                110     120     130     140     150
PredSS          aaaaaaa   aaaaaaaaaa   bbbbbb   aaaaaaa
AA seq          MYANLVFNKGIDEFYAQCEKVGVDVSLVADVPVEESAPFRQAALRHNVPAP
Prob a          54554453447899999988400100000111222234788998731111
Prob b          32112211000000000000011168986322110100000000000123
Predator        HHHHH HHHHHHHHH EEEEE HHHHHHHH E

                160     170     180     190     200
PredSS          bbb   aaaaaaaaaa   bbbb   aaaaaaaaaaaaaaaaaa
AA seq          IFICPPNADDDLRLQIASYGRGYTYLLSRAGVTGAENRAALPLNHLVAKL
Prob a          00000000158999999731111212235211125556654388899999
Prob b          8985200000000000011011367753111221110011220000000
Predator        EEE HHHHHHHH EEEEE HHHHH HHHHHH

                210     220     230     240     250
PredSS          aaa   aaaaaaaaaa   aaaaaaaaaa   aaa
AA seq          KEYNAAPPLQGFGISAPDQVKAIDAGAAGAISGSAIVKIIIEQHINEPEK
Prob a          88632100111101114789999987453122226878888997542588
Prob b          0000000013343320000000000000012211101000000000000
Predator        HHH HHHHHHH HHHHHHH HHHHHHHH HHH

                260
PredSS          aaaaaaaaaaaaaaaaaa
AA seq          MLAALKVVFVQPMKAATRS
Prob a          989999998878898663
Prob b          0000000000000000011
Predator        HHHHHHHH
    
```

### ***Hidden Markov Model (Discrete-Space Model)***

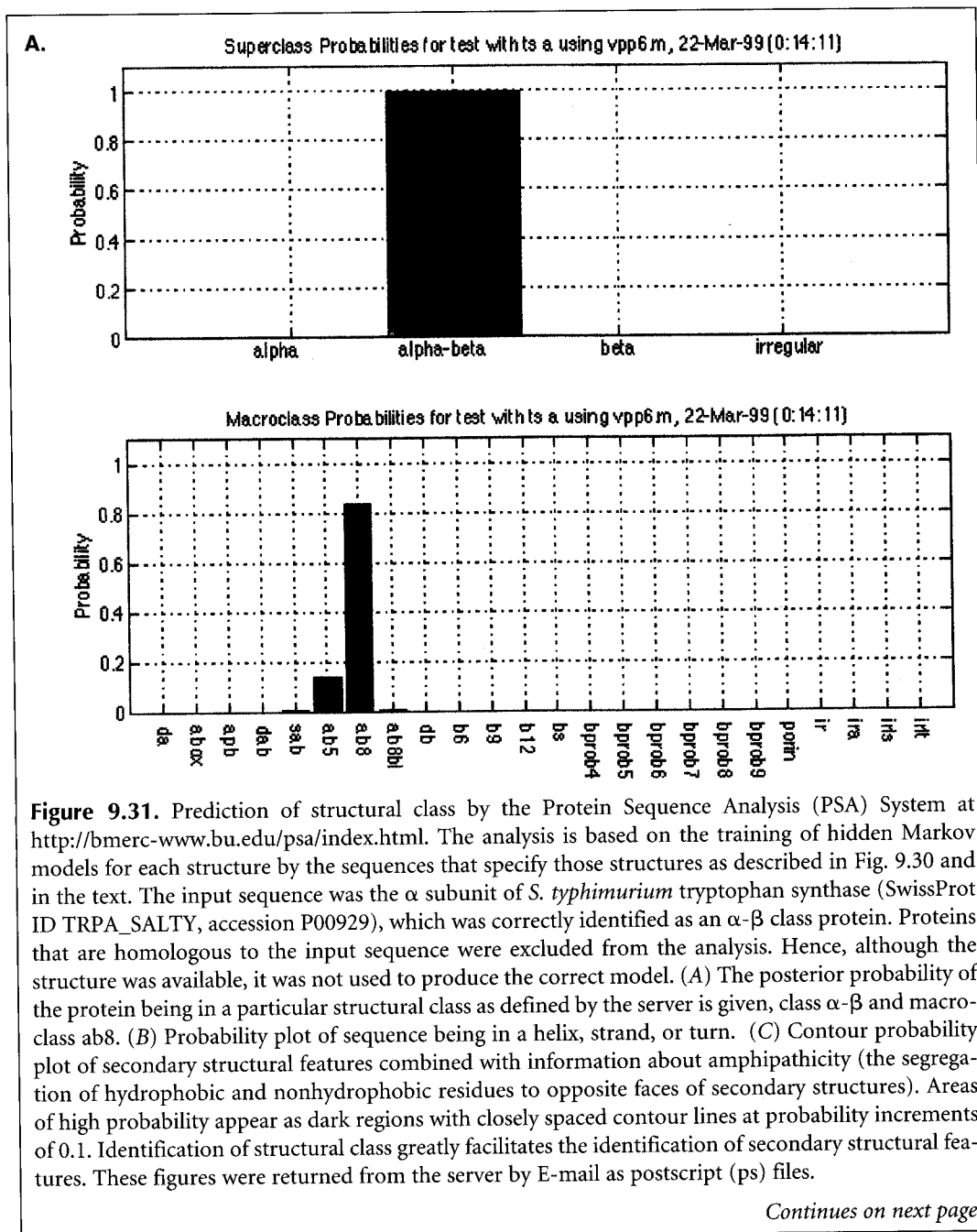
HMMs have been used to model alignments of three-dimensional structure in proteins (Stultz et al. 1993; Hubbard and Park 1995; Di Francesco et al. 1997, 1999; FORREST Web server at <http://absalpha.dcrn.nih.gov:8008/>). In one example of this approach, the models are trained on patterns of  $\alpha$  helices,  $\beta$  strands, tight turns, and loops in specific structural classes (Stultz et al. 1993, 1997; White et al. 1994), which then may be used to provide the most probable secondary structure and structural class of a protein. The manner by which protein three-dimensional domains can be modeled is illustrated in Figure 9.30. An example of the class prediction by the Protein Sequence Analysis (PSA) server at Boston University is shown in Figure 9.31.

## **Prediction of Three-dimensional Protein Structure**

Because the number of ways that proteins can fold appears to be limited, there is considerable optimism that ways will be found to predict the fold of any protein, just given its amino acid sequence. Structural alignment studies have revealed that there are more than 500 common structural folds found in the domains of the more than 12,500 three-dimensional structures that are in the Brookhaven Protein Data Bank. These studies have also revealed that many different sequences will adopt the same fold. Thus, there are many combinations of amino acids that can fit together into the same three-dimensional conformation, filling the available space and making suitable contacts with neighboring amino acids to adopt a common three-dimensional structure. There is also a reasonable probability that a new sequence will possess an already identified fold. The object of fold recognition is to discover which fold is best matched. Considerable headway toward this goal has been made.

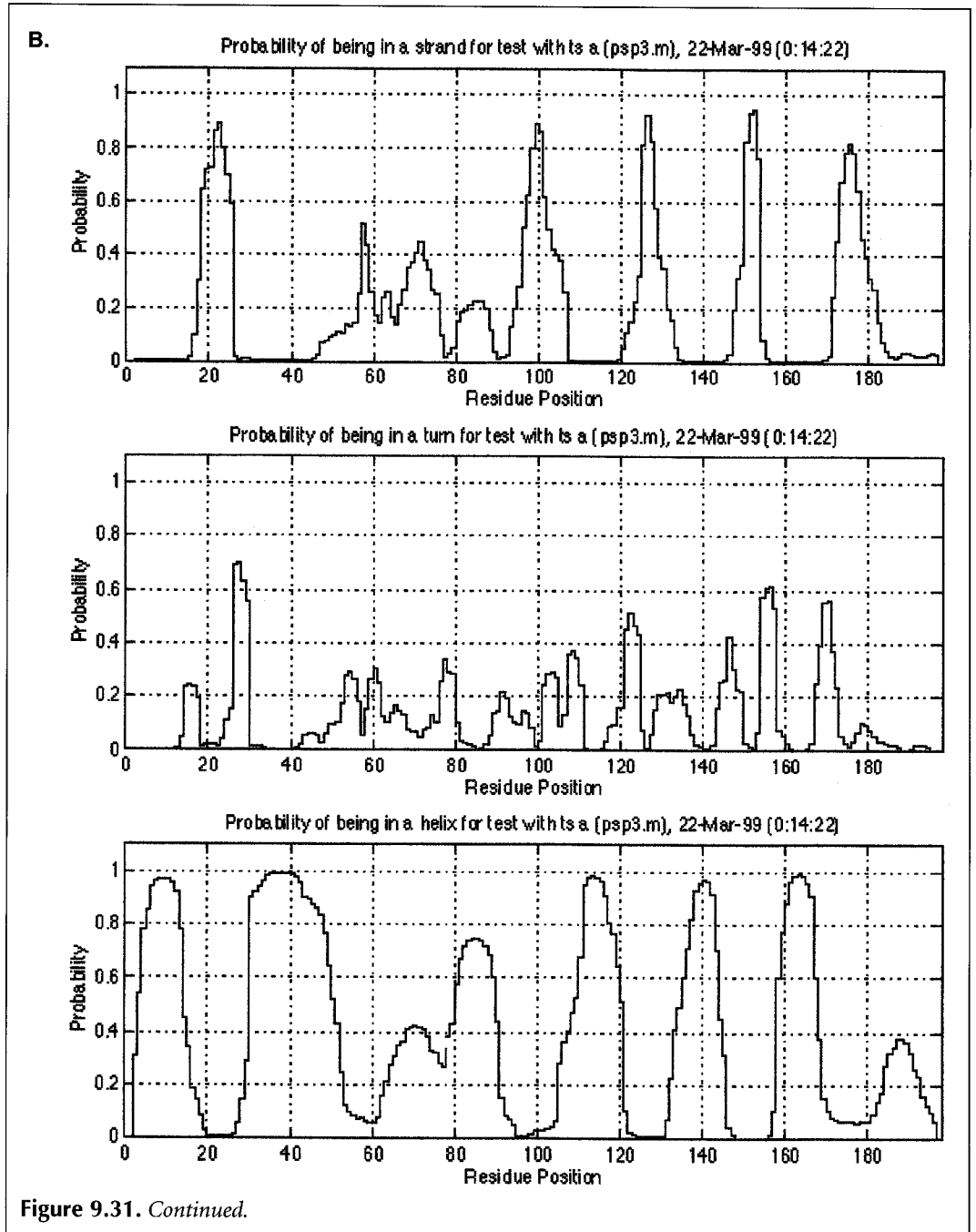
Sequence alignment can be used to identify a family of homologous proteins that have the same sequence, and presumably a similar three-dimensional structure. As discussed above, there are many databases that link sequence families to the known three-dimensional structure of a family member. The structure of even a remote family or superfamily member can be predicted through such sequence alignment methods. When the sequence of a protein of unknown structure has no detectable similarity to other proteins, other methods of three-dimensional structure prediction may be employed. One such method is sequence threading.

In threading, the amino acid sequence of a query protein is examined for compatibility with the structural core of a known protein structure. Recall that the protein core is made up of  $\alpha$  helices,  $\beta$  strands, and other structural elements folded into a compact structure. The environment of the core is strongly hydrophobic with little room for water molecules, extra amino acids, or amino acid side chains that are not able to fit into the available space. Side chains must also make contact with neighboring amino acid side chains in the structure, and these contacts are needed for folding and stability. Threading methods examine the sequence of a protein for compatibility of the side groups with a known protein core. The sequence is “threaded” into a database of protein cores to look for matches. If a reasonable degree of compatibility is found with a given structural core, the protein is predicted to fold into a similar three-dimensional configuration. Threading methods are undergoing a considerable degree of evolution at the present time. An excellent description of algorithms for threading is found in Lathrop et al. (1998). Presently available methods require considerable expertise with protein structure and with programming. However, there are some sites where the analysis may be performed on a Web server, as shown in Table 9.8.

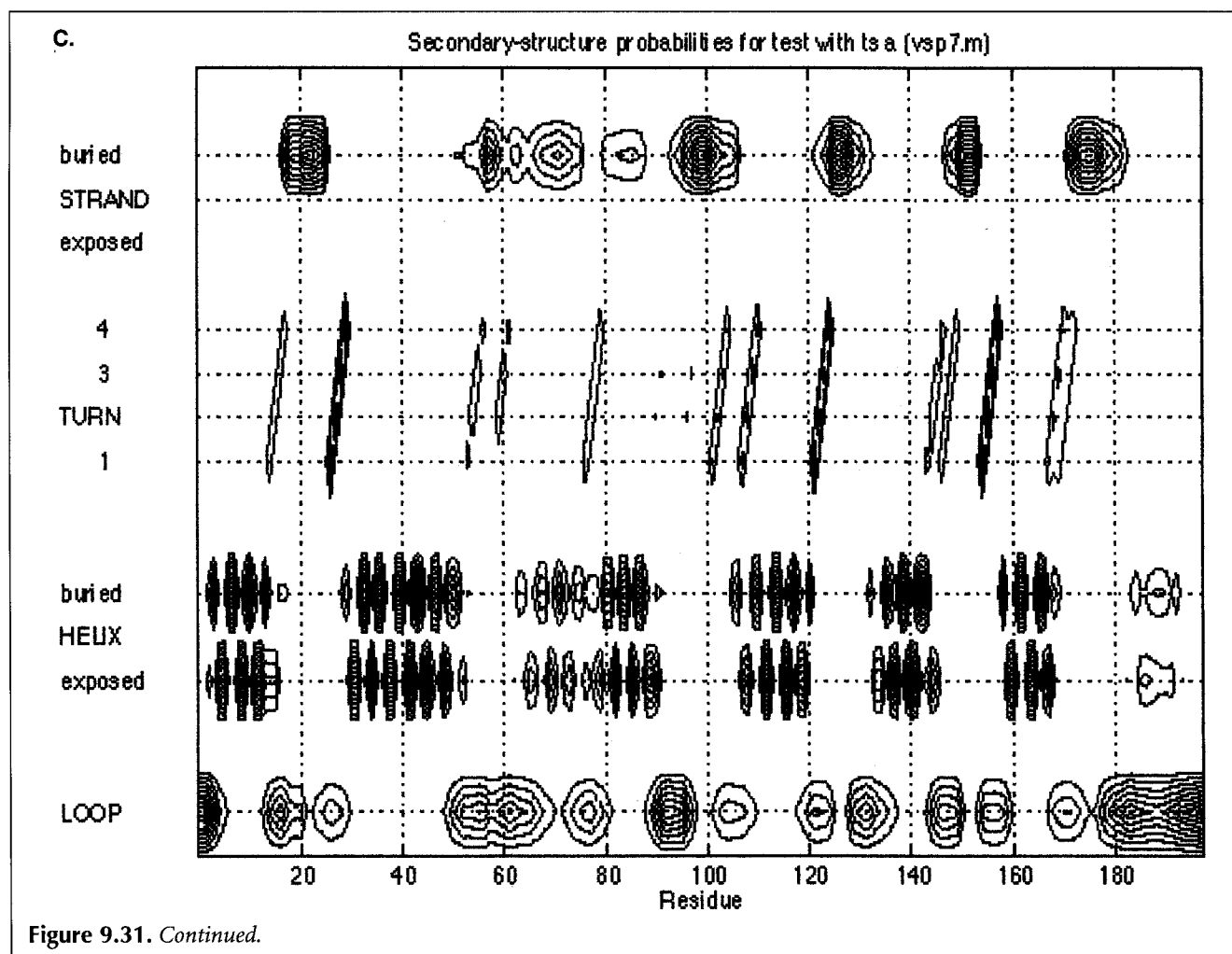


There are two methods in common use for deciding whether or not a given protein sequence is compatible with a known structural core, the environmental template (or structural profile) method and the contact potential method. In the environmental template method (Bowie et al. 1991, 1996; see also Ouzounis et al. 1993; Johnson et al. 1996), the environment of each amino acid in each known structural core is determined, including the secondary structure, the area of the side chain that is buried by closeness to other atoms, types of nearby side chains, and other factors. On the basis of these descriptions at each site, the position is classified into one of 18 types, 6 representing increasing levels of residue burial and fraction of surface covered by polar atoms combined with three classes of secondary structure. Each amino acid is then assessed for its ability to fit into that type





of site in the structure. For example, if the side group is buried, another amino acid with a hydrophobic side chain may fit best into the structure at that position. The sequence of the protein is then aligned with a series of such environmentally defined positions in the structure to see whether a series of amino acids in the sequence can be aligned with the assigned structural environments of a given protein core. The procedure is then repeated for each core in the structural database, and the best matches of the query sequence to the core are identified. In the residue-residue contact potential method, the number and closeness of contacts between amino acids in the core are analyzed (Sippl 1990; Jones et al. 1992; Sippl and Weitckus 1992; Bryant and Lawrence 1993). The query sequence is evaluated for



amino acid interactions that will correspond to those in the core and that will contribute to the stability of the protein. The most energetically stable conformations of the query sequence thereby provide predictions of the most likely three-dimensional structure.

### ***Structural Profile Method***

In the structural profile method, predictions as to which amino acids might be able to fit into a given structural position are in the form of a sequence profile. This method assumes that if the query protein folds the same way as a target structure, the environments of the amino acids will be in the same linear order as they are in the target. In the normal scoring matrix, it is assumed that a given amino acid substitution always has the same likelihood of every occurrence of the substitution. However, in protein three-dimensional structures, a given substitution may have quite different effects depending on where in the structure and in which structure the substitution occurs. In a loop, where there are not many chemical and physical constraints, the substitution may usually not have any deleterious effects on the overall structure of the protein. In contrast, the same substitution in protein cores, where there are many restraints, may sometimes be possible without deleterious effects, but in other cases may be extremely deleterious. Thus, a sequence profile giving values for substitutions at each amino acid position is made for each core in the PDB.

**Table 9.8.** *Threading servers and program sources*

Program	Web address	Method	Reference
123D	<a href="http://www-lmmb.ncifcrf.gov/~nicka/123D.html">http://www-lmmb.ncifcrf.gov/~nicka/123D.html</a>	contact potentials between amino acid side groups	Alexandrov et al. (1996)
3D-PSSM	<a href="http://www.bmm.icnet.uk/~3dpssm">http://www.bmm.icnet.uk/~3dpssm</a>	sequence-structure using position-specific scoring matrices	Russell et al. (1997)
Honig lab	<a href="http://honiglab.cpmc.columbia.edu/">http://honiglab.cpmc.columbia.edu/</a>	threading methods using biophysical properties	see Web site
Libra I	<a href="http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/LIBRA_I.html">http://www.ddbj.nig.ac.jp/htmls/E-mail/libra/LIBRA_I.html</a>	target sequence and 3D profile are aligned by dynamic programming	Ota and Nishikawa (1997)
NCBI structure site	<a href="http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.html">http://www.ncbi.nlm.nih.gov/Structure/RESEARCH/threading.html</a>	Gibbs sampling algorithm used to align sequence and structure <sup>a</sup>	Bryant (1996)
Profit	<a href="http://lore.came.sbg.ac.at/home.html">http://lore.came.sbg.ac.at/home.html</a>	fold recognition by the contact potential method	M. Sippl (see Web site)
Threader 2	<a href="http://insulin.brunel.ac.uk/threader/threader.html">http://insulin.brunel.ac.uk/threader/threader.html</a>	prediction by recognition of the correct fold from a library of alternatives	Jones et al. (1995)
TOPITS	<a href="http://www.embl-heidelberg.de/predictprotein/doc/help05.html#P5_adv_prd_topits">http://www.embl-heidelberg.de/predictprotein/doc/help05.html#P5 adv prd topits</a>	detects similar motifs of secondary structure and accessibility between a sequence of unknown structure and a known fold	Rost (1995a,b)
UCLA-DOE structure prediction server	<a href="http://www.doe-mpi.ucla.edu/people/frsvr/frsvr.html">http://www.doe-mpi.ucla.edu/people/frsvr/frsvr.html</a>	fold-recognition using 3D profiles and secondary structure prediction methods	Fischer and Eisenberg (1996)

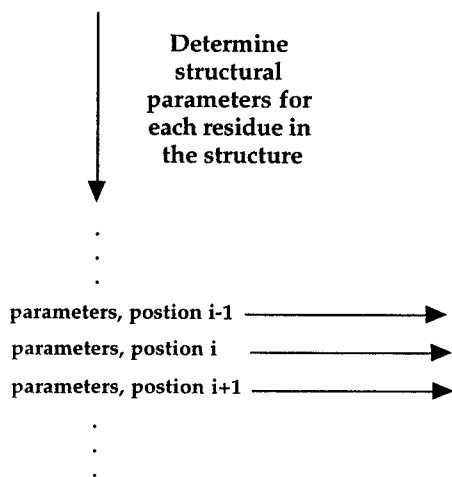
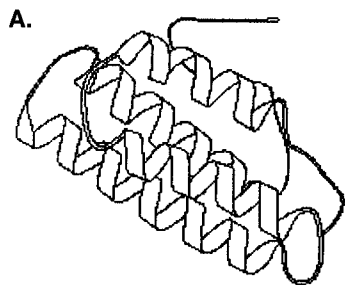
Information on the research groups that work on structure prediction may be found at the CASP2 Web sites accessible at <http://predictioncenter.llnl.gov/>.

<sup>a</sup>Program has to be set up on a UNIX server.

These profiles, one for each core in the database, are then used to score the query sequence to be modeled for compatibility with that core.

The structural three-dimensional profile is a table of scores with one row for each amino acid position in the core and a column for each possible amino acid substitution at that position plus two columns for deletion penalties at that site, as shown in Figure 9.32. Each position in the core is assigned to one of 18 classes of structural environment. The scores in each row reflect the suitability of a given amino acid for that particular environment. The penalty at each core position reflects the acceptability of an insertion or deletion of one or more amino acids at that position in the structure. If the position is within the core, these penalties are generally high to reflect incompatibility with the structure, but lower for positions on the surface of the core and within loop regions. The dynamic programming algorithm is used to identify an optimal, best-scoring alignment, much as in aligning sequences by dynamic programming (discussed in Chapter 3). If a target structure is found to have a significantly high score, the new sequence is predicted to have a fold similar to that of the target core.

An entire database of sequences may be matched to a given structural profile to find the most compatible, a procedure called inverse folding. The alignment score for each protein is determined and then converted to a *Z* score, the number of standard deviations from the mean score for all of the sequences. The highest scoring sequences are the most compatible with a given structure (Bowie et al. 1996).



### 3D Profile

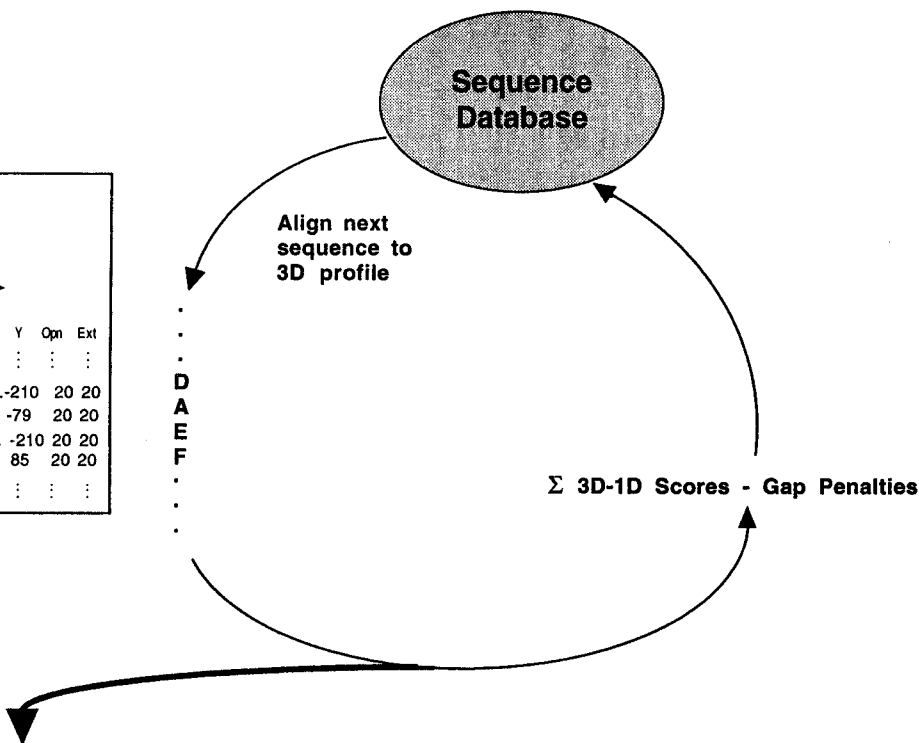
Position in Structure ↓	Amino Acid Type →								Gap Penalties	
	A	C	D	E	...	W	Y	OPN	EXT	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	
10	-50	101	121	...	-92	-75	100	10		
-24	87	-132	-95	...	182	167	100	10		
22	34	-11	-5	...	54	76	100	10		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮		

Scores reflecting compatibility with structural parameters

B.

### 3D Profile

Position in Structure ↓	Amino Acid Type →							
	A	C	D	E	F	Y	Opn	Ext
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
E <sub>α</sub>	46	-46	44	59	-220	...	-210	20 20
P <sub>1</sub> α	6	-93	28	55	-143	...	-79	20 20
B <sub>1</sub> α	46	-44	44	59	-220	...	-210	20 20
E <sub>α</sub>	-89	10	-162	-71	70	...	85	20 20
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



List of Top Scoring Sequences

The above three-dimensional profile provides a discrete list of scores for matching one-dimensional sequence to a three-dimensional structure. This profile undergoes sharp transitions in values as the structural environment changes. Improved performance has been achieved by smoothing the values in these transitional regions to give a more gradual change using a Fourier analysis. Another improvement in the profile representation of protein three-dimensional structures, known as the *residue pair preference profile* (R3P) method, has been introduced (Wilmanns and Eisenberg 1993, 1995; Bowie et al. 1996).

R3P takes into account the amino acid neighbors, main-chain conformations, and secondary structure of each residue in the structure. Recall that to make amino acid scoring matrices for sequence–sequence comparisons, the frequency of amino acid substitutions in alignments is counted in sequence alignments. These frequencies are then divided by the expected frequency of finding the amino acids together in an alignment by chance. The ratio of the observed to expected counts is an odds score, and this score is usually converted to a log odds score for convenience in combining likelihood scores by adding their logarithms. Similarly, in the R3P method of making a three-dimensional scoring profile, the frequency of finding a particular pair of interacting amino acids, each with a particular structural feature, is calculated from the number of occurrences in known structures. For example, how often does amino acid *a* in an  $\alpha$  helix interact with amino acid *b* in a  $\beta$  strand? This observed frequency of interaction in a specific structural configuration is then divided by the frequency of finding *a* and *b* interacting in any configuration, and the result is converted to a log odds score.

The pair preference log odds score  $S(aa_i, s_i, aa_j, s_j)$  for the amino acids  $aa_i$  and  $aa_j$  having properties  $c_i$  and  $c_j$ , respectively, is given by

$$S(aa_i, c_i, aa_j, c_j) = \ln [ P(aa_i, c_i, aa_j, c_j) / P(aa_i, aa_j) ] \quad (12)$$

where  $P(aa_i, c_i, aa_j, c_j)$  is the frequency of amino acids  $aa_i$  and  $aa_j$  having properties  $c_i$  and  $c_j$ , respectively, and  $P(aa_i, aa_j)$  is the frequency of finding an amino acid pair  $aa_i$  and  $aa_j$ . The score for position  $aa_i$  is then given by a weighted sum of all scores for the interacting pairs with  $aa_i$ .

$$S(aa_i) = \sum w_j S(aa_i, c_i, aa_j, c_j) / \sum w_j \quad (13)$$

where  $w_j$  is a weight representing the compatibility of the environment residue with its own local environment.

Amino acid interactions of a given amino acid residue in a particular core are then analyzed. To determine the neighbors of a given amino acid in the structure, a sphere of radius 12 Å is drawn centered on the  $C_\beta$  atom (see Fig. 9.2). If the  $C_\beta$  atom of another residue in the structure falls within this sphere, they may be interacting. A cylinder of radius 1.6 Å is then drawn between the  $C_\beta$  atoms and, if no H bonds or any other

**Figure 9.32.** The structural three-dimensional profile. (A) Generation of a three-dimensional profile of a structural core. (B) Screening sequences for compatibility with 3D profile. The methods of analysis are described in the text. (A and B: Redrawn, with permission, from Bowie et al. 1996 [copyright Academic Press].)

residue falls within this cylinder, the amino acid pair is considered to be interacting. This procedure is repeated for each amino acid that falls within the sphere, resulting in a defined list of approximately 8 amino acid pairs that are close enough without barriers to prevent interaction with the given residue. The amino acid type and one of several structural properties for the residue in question and for each interacting residue are then obtained. For example, the secondary structure of the two residues ( $\alpha$  helix,  $\beta$  strand, or other) may be taken into account, giving  $20 \times 3$  possible combinations of amino acid and secondary structure. Structural properties of the interacting residue may instead include the backbone dihedral angles  $\Phi$  and  $\Psi$  (see Fig. 9.2) and the number of neighboring residues.

The structural configurations of the given residue and each interacting neighbor are determined. From this information, a score for this interaction can be found from the above analysis. Scores for all of the remaining interacting residues in their particular configurations can be found and then added to give a log odds score for the given amino acid site in the core. This score represents the likelihood of finding such a set of amino acid neighbors in their respective configurations in known protein structures. A value is then determined for various amino acid substitutions or for placing an insertion or deletion at that site. A similar set of scores is then obtained for each position in the protein core to generate a three-dimensional profile matrix based on the neighboring interactions. Three such profiles have been generated, one for each type of structural property in the amino acid pairs—backbone angles, secondary structure, and the number of neighboring residues for the interacting amino acid. A combined three-dimensional profile using elements of these residue pair preference profiles and those of the neighborhood three-dimensional profiles has also been used.

Sequence–structure alignments produced by the R3P method can be improved by an iterative procedure. In the initial alignment between a sequence and the three-dimensional profile of a core, predictions are made as to which residues will interact in the modeled three-dimensional structure. This feature provides information for improving the alignment. Likelihood scores for the predicted interactions can be calculated in the same way as described above for the amino acid interactions in the core. The scores for these interactions may then be summed, as before. In this case, these scores are weighted before summing to reduce the influence of those neighboring amino acids that are not in a compatible environment (Bowie et al. 1996). In evaluations of the R3P method with known three-dimensional structures, alignments are 50% or more correct on average for sequences whose three-dimensional structure pairs superimpose with a root mean square (rms) deviation of 1.97 Å or less (Wilmanns and Eisenberg 1995). Sequence–structure alignments may be further improved by including in the analysis the predicted secondary structure of the input sequence, with further improvements in fold assignment of 25% (Fischer and Eisenberg 1996).

One disadvantage of the structural profile method and the use of environmental variables is that these properties are statistically associated with the original sequence. Hence, the method retains a preference for matching the original sequence of the core protein. On the other hand, the success of present methods of three-dimensional structure prediction depends on a certain minimal level of similarity. The sequence of environmental patterns in the query sequence and the structure must also be in the same order throughout the sequence for the method to work. However, as discussed above for the SSAP alignment program, this problem may be circumvented by using local alignments.

### Contact Potential Method

In this method, each structural core is represented as a two-dimensional contact matrix. The method is very similar to that used by the distance matrix method of the program DALI and illustrated in Figure 9.15. A simple matrix is produced with the amino acids in the structure listed across the rows and down the columns. In each matrix position, the distance between the corresponding pair of amino acids in the structure is placed. The amino acids in closest contact are immediately recognizable, and a group produces recognizable patterns. The object is to superimpose sets of amino acid pairs in the query sequence on to the distance matrix of the core. As shown in Figure 9.15, part B, sequences that fold into a similar structure should show similar contacts, although the amino acids that make up each structural feature do not have to be in the same linear order in both sequences. However, a large number of contacts must be analyzed to find the correct alignment.

To find the best combinations, the approximate conformational energies of each predicted pair are summed to predict the conformational stability of the predicted structure. Contacts have been extensively analyzed, and lookup tables with energies associated with these contacts have been produced. Hence, the energetic contributions of many possible combinations of pairs can be tested in a relatively short period of time. Computer experiments have revealed that contact energies can be used to choose the correct core in a structural database. Supporters of this method claim that the method can detect structural similarity in proteins that do not share any detectable sequence similarity. However, as shown in the next section, in truly blind experiments, the reliability of predictions drops when there is less than 25% sequence identity. A possible limitation to this analysis is that the energy associated with an isolated amino acid pair is assumed to be similar to that found in known protein structures. Recent experiments have suggested that the conformational energy of groups of amino acids larger than two may provide a more reliable prediction.

#### **Example: Structure Prediction by Web Servers That Provide a Threading Service**

These results were sent by E-mail. (A) Structure prediction by Libra (Table 9.8) and (B) UCLA-DOE structure prediction server. This server also provides a Web page for each match giving the results of other types of sequence database searches, secondary structure analyses, and the TOPITS server results. This analysis is of the  $\alpha$  subunit of *S. typhimurium* tryptophan synthase (SwissProt ID TRPA\_SALTY, accession P00929). The Web addresses and methods used by these servers are given in Table 9.8.

A. -----  
 Forward Folding Search by LIBRA I  
 -----

LIBRA I was written by M. Ota in 1994-97

gap1= 2.400 gap9= 4.800 gapE= 0.400 npdb= 1389  
 gap1: Gap opening penalty for exposed sites  
 gap9: Gap opening penalty for buried sites  
 gapE: Gap extension penalty  
 npdb: Number of the structural templates

[Input sequence was the a subunit of *S. typhimurium* tryptophan synthase, Swiss-prot ID TRPA\_SALTY, accession P00929]

Compatible structures are:

Rk	StrC	Protein	Lsr	Lal	Rsc	SD	Rs/N	ID%
1	2tsya	TRYPTOPHAN SYNTHASE;	262	268	-154.2	-5.57	-0.576	83.2
2	1tlfa	TRYPTIC CORE FRAGMENT OF THE L	296	284	-99.4	-2.94	-0.350	12.0
3	2liv-	LEUCINE(SLASH)*ISOLEUCINE(SLAS	344	275	-95.4	-2.75	-0.347	9.1

.

Key:

Rk : Rank position  
 StrC: Structural code  
 Lsr : Length of the structural template  
 Lal : Length of the aligned region  
 Rsc : Raw score of the structural template  
 SD : Standardized score  
 Rs/N: Raw score (Rsc) normalized by the alignment length (Lal)  
 ID% : Sequence identity

3D-ID alignments are:

1 2tsya structure vs your tsa sequence

```

3177238623 7322149997 9899888615 1775499469 6466899979
IAAAAAAAAA AAAGleeBBB BBBBelegeA AAAAAAAAAA AAAllegeBBB
MERYENLFAQ LNDRRREGAFV PFVTLGDPGI EQSLKIIDL IDAGADALEL
::::: :::: : : ::::: ::::::::::: ::::::::::: : ::::::::::
MERYESLFAQ LKERKEGAFV PFVTLGDPGI EQSLKIIDL IEAGADALEL

```

.

Key:

1st line: Accessibility of the aligned site;  
 1(exposed to water)-9(buried in protein)  
 2nd line: Local conformation of the aligned site;  
 A(alpha), B(beta), gel(coils classified by the dihedral angles)  
 3rd line: Sequence of the template structure  
 4th line: Match site of the alignment  
 5th line: Query sequence

B.

Most similar fold: lwsya  
 TRYPTOPHAN SYNTHASE (E.C.4.2.1.20)

RANK	Z-SCORE	FOLD	LENGTH	ALI	%ID
1	77.23	lwsya	248		85
2	5.21	1aj0	232		24
3	4.64	1adla	222		18.

.

LEGEND:

COL. 1: RANK. The ranks are obtained by sorting the fold library, by Z-SCORES, in decreasing order. Only the 15 structures that are most compatible to your sequence are shown.  
 COL. 2: Z-SCORE. The z-scores are computed using the distribution of raw scores (not shown) of all folds.



COL. 3: FOLD. Protein Data Bank codes for the coordinates of the 3D structures.

COL. 4: LENGTHALI. The number of residues from your sequence that were aligned to the fold.

COL. 5: % ID. Percentage of identical residues in the alignment. [Description of each match not shown]

**RELIABILITY OF THIS PREDICTION:**

With this method the confidence threshold is a z-score of 4.8 +/- 1.0.

**YOUR HIGHEST SCORING FOLD IS ABOVE THIS THRESHOLD**

Below is the alignment of your sequence with the top hit structure. In the near future, all the alignments in a more readable format will be made available.

```

hhhhhhhhhhhhhh  bbbbbb  hhhhhhhhhhhhh  bbb
MERYESLFAQLKERKEGAFVPPFVTLGDPGIEQSLKIIDTLIEAGADALEL
|||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||  |||||
MERYENLFAQLNDRREGAFVPPFVTLGDPGIEQSLKIIDTLIDAGADALEL
hhhhhhhhhhhh  bbbbbb  hhhhhhhhhhhhh  bbb

```

## EVALUATING THE SUCCESS OF STRUCTURE PREDICTIONS

As the above methods were developed, they were tested for ability to predict a structure that was already known. The structure to be predicted may be left out of the learning step so that the method has not been trained in any recognizable way to identify the correct structure. However, when the result is already known, there is always a possibility that the method was helped in some unintended way to identify the correct structure. A totally blind test of prediction accuracy provides a more objective test. A series of contests called CASP (critical assessment of structure prediction) was conceived in which structural biologists who were about to publish a structure were asked to submit the corresponding sequence for structure prediction by the contestants. The predictions were then compared with the newly determined structures. The newest CASP3 competition is given on a Web site (<http://predictioncenter.llnl.gov/casp3/results/access.cgi>). The results of earlier projects are given at <http://predictioncenter.llnl.gov/casp1/> and <http://predictioncenter.llnl.gov/casp2/>. The contest involved a large number of research groups using a variety of methods including threading techniques. In one report of CASP3, the authors suggest that although there was overall progress from CASP1 to CASP2, there was little additional progress from CASP2 to CASP3. However, some improvement can be argued in CASP3 since the targets were more difficult (Sippl et al. 1999).

In the CASP2 conference, 32 groups made a total of 369 predictions on 15 different targets. There were two goals for each group: (1) to predict the correct three-dimensional fold of the target protein as the most similar known structures and (2) to predict the alignment of the sequence to the fold accurately. Once the structures of the prediction targets became available, the structure was aligned by DALI, SSAP, and VAST with all entries in the structural database to determine the closest matching structures that should have been found and also the sequence-structure alignment. The predictions were then compared to these alignments and evaluated for accuracy by specific criteria (Levitt 1997; Marchler-Bauer et al. 1997). This task was a most difficult one because different groups of investigators made predictions for different groups of proteins, some proteins much more difficult to predict than others. The range of sequence identity of target sequences to a known structure varied from 20% to 85%. The most difficult to predict and also the least successfully predict-

ed were those that have less than 25% identity to any other protein of known structure. The easiest and most successfully predicted were those with sequence similarity above 25% (Martin et al. 1997).

The results of the CASP2 contest have been published by the participants in a special issue of *PROTEINS: Structure, Function and Genetics*, Suppl. 1, 1997, which provides details of the threading methods used. A similar volume discusses progress of CASP3 (*PROTEINS*, Suppl. 3, 1999). Threading methods improved considerably in performance in the 2-year period between the CASP1 and CASP2 meetings. A large number of groups using threading methods recognized the easier targets and performed much better than using simple sequence alignments (Levitt 1997). The advantages of using distant sequence homology and human knowledge of protein structure to predict three-dimensional structure was demonstrated by Murzin and Bateman (1997), who made the largest number of correct predictions. Their method uses the SCOP database, which organizes all known protein folds according to their structural and evolutionary relationships, for manual predictions. Their approach correctly assigned into an existing SCOP superfamily all six targets that were attempted, and found a homologous protein with a very similar structure. Local alignments between the target sequence and the corresponding protein superfamily were also among the most accurate. Several threading groups that were among the best performers are given in Table 9.8. At the present time, these methods are most suitable for modeling sequences that are recognizably similar to a known structure. These results confirm an earlier analysis that threading algorithms are quite disappointing in performance (Lemer et al. 1995). Improvements have been achieved by using a set of multiply aligned sequences instead of a single sequence (Defay and Cohen 1996; Ortiz et al. 1998).

## STRUCTURAL MODELING

In the above section, detecting sequence similarity between a query sequence and a sequence of known structure plays an important role in successful structure prediction. Database searches as described in Chapter 6 provide alignments of a query sequence with a database of sequences, and can be used to search a database of protein sequences restricted to those of known structure. Hence, any alignment provides an indication as to which amino acids in the query may occupy a particular position in a structure. A search of this kind may be enhanced by superimposing the query sequence onto the molecular backbone of the matched sequence to produce a PDB file suitable for analysis by a three-dimensional viewer. An example of this type of analysis is provided by the Swiss-model Web site (Table 9.9). Molecular distances, angle, and energies of the superimposed sequence may then be analyzed and manipulated by the SPDBV viewer (Table 9.4). Additional Web sites for molecular modeling are listed in Table 9.9.

**Table 9.9.** *Web sites for predicting structural features of a query sequence*

Site	Web address	Description	Reference
Modeller	<a href="http://guitar.rockefeller.edu/modeller/modeller.html">http://guitar.rockefeller.edu/modeller/modeller.html</a>	dynamic programming alignment of sequences and structures and molecular dynamics methods	Sali et al. (1995)
Swiss-model	<a href="http://www.expasy.ch/swissmod/SWISS-MODEL.html">http://www.expasy.ch/swissmod/SWISS-MODEL.html</a>	sequence alignment of query with sequences of known structure	Peitsch (1996)
Whatif	<a href="http://www.cmbi.kun.nl/whatif/">http://www.cmbi.kun.nl/whatif/</a>	flexible molecular graphics rendering of models	Rodriguez et al. (1998)

## SUMMARY AND FUTURE PROSPECTS

This chapter has described a number of methods for predicting protein structure from amino acid sequence. The best approach is to locate a link by sequence analysis between a new protein and a protein of known structure. Even a marginal sequence alignment with a protein of known structure can provide a feasible structural model. Databases that organize proteins into clusters and families with links to known protein structure are also a valuable resource for structure prediction. Proteins that represent new structural folds and domains can be readily identified in these databases, and these proteins can then be targeted for structural analysis by laboratory methods. Meanwhile, the methods for secondary structure and threading analysis (fitting a sequence to a structure) can provide useful predictions, although with variable levels of reliability. Increased confidence should come when several methods give a similar prediction.

The analysis of genomes described in Chapter 10 offers an additional opportunity for protein analysis. Functions of proteins can be discovered through conserved patterns of gene regulation and organization on the chromosomes of related organisms. The function and structure of a protein in one organism can then be predicted based on the function and structure of a functionally similar protein in a second organism.

## REFERENCES

- Alexandrov N.N. and Fischer D. 1996. Analysis of topological and nontopological structural similarities in the PDB: New examples with old structures. *Proteins* **25**: 354–365.
- Alexandrov N.N., Nussinov R., and Zimmer R.M. 1996. Fast protein fold recognition via sequence to structure alignment and contact capacity potentials. *Pac. Symp. Biocomput.* 1996: 53–72.
- Attwood T.K., Flower D.R., Lewis A.P., Mabey J.E., Morgan S.R., Scordis P., Selley J., and Wright W. 1999. PRINTS prepares for the new millennium. *Nucleic Acids Res.* **27**: 220–225.
- Bairoch A. 1991. PROSITE: A dictionary of sites and patterns in proteins. *Nucleic Acids Res. (suppl.)* **19**: 2241–2245.
- Barker W.C., Pfeiffer F., and George D.G. 1995. Superfamily and domain. In *Methods in protein structure analysis* (ed. M.Z. Atassi and E. Appella), pp. 473–481. Plenum Press, New York.
- . 1996. Superfamily classification in the PIR-international protein sequence database. *Methods Enzymol.* **266**: 59–71.
- Berger B., Wilson D.B., Wolf E., Tonchev T., Milla M., and Kim P.S. 1995. Predicting coiled coils by use of pairwise residue correlations. *Proc. Natl. Acad. Sci.* **92**: 8259–8263.
- Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., and Bourne P.E. 2000. The protein data bank. *Nucleic Acids Res.* **28**: 235–242.
- Blundell T.L. and Johnson M.S. 1993. Catching a common fold. *Protein Sci.* **2**: 877–883.
- Bowie J.U., Lüthy R., and Eisenberg D. 1991. A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.
- Bowie J.U., Zhang K., Wilmanns M., and Eisenberg D. 1996. Three-dimensional profiles for measuring compatibility of amino acid sequence with three-dimensional structure. *Methods Enzymol.* **266**: 598–616.
- Branden C. and Tooze J. 1991. *Introduction to protein structure*. Garland Publishing, New York.
- Brenner S.E., Chothia C., Hubbard T.J., and Murzin A.G. 1996. Understanding protein structure: Using Scop for fold interpretation. *Methods Enzymol.* **266**: 635–643.
- Brown N.P., Leroy C., and Sander C. 1998. MView: A web compatible database search or multiple alignment viewer. *Bioinformatics* **14**: 380–381.
- Bryant S.H. 1996. Evaluation of threading specificity and accuracy. *Proteins* **26**: 172–185.
- Bryant S.H. and Lawrence C.E. 1993. An empirical energy function for threading protein sequence through the folding motif. *Proteins Struct. Funct. Genet.* **16**: 92–112.
- Chothia C. 1992. Proteins. One thousand families for the molecular biologist. *Nature* **357**: 543–544.

- Chou P.Y. and Fasman G.D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol. Relat. Areas Mol. Biol.* **47**: 45–147.
- Cohen F.E., Abarbanel R.M., Kuntz I.D., and Fletterick R.J. 1983. Secondary structure assignment for  $\alpha/\beta$  proteins by a combinatorial approach. *Biochemistry* **22**: 4894–4904.
- . 1986. Turn prediction in proteins using a pattern-matching approach. *Biochemistry* **25**: 266–275.
- Corpet F., Gouzy J., and Kahn D. 1998. The ProDom database of protein domain families. *Nucleic Acids Res.* **26**: 323–326.
- Cuff J.A. and Barton G.J. 2000. Application of multiple sequence alignment profiles to improve protein secondary structure prediction. *Proteins* **40**: 502–511.
- Cuff J.A., Clamp M.E., Siddiqui A.S., Finlay M., and Barton G.J. 1998. JPred: A consensus secondary structure prediction server. *Bioinformatics* **14**: 892–893.
- Dayhoff M.O., Barker W.C., Hunt L.T., and Schwartz R.M. 1978. Protein superfamilies. In *Atlas of protein sequence and structure* (ed. M.O. Dayhoff), vol. 5, suppl. 3, pp. 9–24. National Biomedical Research Foundation, Georgetown University, Washington, D.C.
- Defay T.R. and Cohen F.E. 1996. Multiple sequence information for threading algorithms. *J. Mol. Biol.* **262**: 314–323.
- Di Francesco V., Munson P.J., and Garnier J. 1999. FORESST: Fold recognition from secondary structure predictions of proteins. *Bioinformatics* **15**: 131–140.
- Di Francesco V., Geetha V., Garnier J., and Munson P.J. 1997. Fold recognition using predicted secondary structure sequences and hidden Markov models of protein folds. *Proteins* (suppl. 1): 123–128.
- Dodge C., Schneider R., and Sander C. 1998. The HSSP database of protein structure-sequence alignments and family profiles. *Nucleic Acids Res.* **26**: 313–315.
- Fischer D. and Eisenberg D. 1996. Fold recognition using sequence-derived predictions. *Protein Sci.* **5**: 947–955.
- . 1999. Finding families for genomic ORFans. *Bioinformatics* **15**: 759–762.
- Frishman D. and Argos P. 1995. Knowledge-based protein secondary structure assignment. *Proteins* **23**: 566–579.
- . 1996. Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9**: 133–142.
- . 1997. Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **27**: 329–335.
- Garnier J., Gibrat J.-F., and Robson B. 1996. GOR method for predicting protein secondary structure from amino acid sequence. *Methods Enzymol.* **266**: 540–553.
- Garnier J., Osguthorpe D.J., and Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**: 97–120.
- Geourjon C. and Deleage G. 1994. SOPM: A self-optimized method for protein secondary structure prediction. *Protein Eng.* **7**: 157–164.
- . 1995. SOPMA: Significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comput. Appl. Biosci.* **11**: 681–684.
- Gibrat J.-F., Madej T., and Bryant S.H. 1996. Surprising similarities in structure comparison. *Curr. Opin. Struct. Biol.* **6**: 377–385.
- Guermeur Y., Geourjon C., Gallinari P., and Deleage G. 1999. Improved performance in protein secondary structure prediction by inhomogeneous score combination. *Bioinformatics* **15**: 413–421.
- Guex N. and Peitsch M.C. 1997. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. *Electrophoresis* **18**: 2714–2723.
- Hansen J.E., Lund O., Rapacki K., and Brunak S. 1997. O-GLYCBASE version 2.0: A revised database of O-glycosylated proteins. *Nucleic Acids Res.* **25**: 278–282.
- Henikoff J.G. and Henikoff S. 1996. Blocks database and its applications. *Methods Enzymol.* **266**: 88–105.
- Henikoff S., Pietrokovski S., and Henikoff J.G. 1998. Superior performance in protein homology detection with the Blocks database servers. *Nucleic Acids Res.* **26**: 309–312.
- Henikoff S., Greene E.A., Pietrokovski S., Bork P., Attwood T.K., and Hood L. 1997. Gene families: The taxonomy of protein paralogs and chimeras. *Science* **278**: 609–614.
- Hirst J.D. and Sternberg M.J. 1992. Prediction of structural and functional features of protein and nucleic acid sequences by artificial neural networks. *Biochemistry* **31**: 7211–7218.
- Hofmann K., Bucher P., Falquet L., and Bairoch A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**: 215–219.

- Hogue C.W. 1997. Cn3D: A new generation of three-dimensional molecular structure viewer. *Trends Biochem. Sci.* **22**: 314–316.
- Hogue C.W. and Bryant S.H. 1998a. Structure databases. *Methods Biochem. Anal.* **39**: 46–73.
- . 1998b. Structure databases. In *Bioinformatics: A practical guide to the analysis of genes and proteins* (ed. A.D. Baxevanis and B.F. Ouellette), pp. 46–73. Wiley-Liss, New York.
- Hogue C.W., Ohkawa H., and Bryant S.H. 1996. A dynamic look at structures: WWW-Entrez and the molecular modeling database. *Trends Biochem. Sci.* **21**: 226–229.
- Holley L.H. and Karplus M. 1991. Neural networks for protein structure prediction. *Methods Enzymol.* **202**: 204–224.
- Holm L. and Sander C. 1993. Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**: 123–138.
- . 1994. Searching protein structure databases has come of age. *Proteins* **19**: 165–173.
- . 1996. Mapping the protein universe. *Science* **273**: 595–603.
- . 1998. Touring protein fold space with Dali/FSSP. *Nucleic Acids Res.* **26**: 316–319.
- Hubbard T.J. and Park J. 1995. Fold recognition and ab initio structure predictions using hidden Markov models and  $\beta$ -strand pair potentials. *Proteins* **23**: 398–402.
- Johnson M.S., May A.C., Ridionov M.A., and Overington J.P. 1996. Discrimination of common protein folds: Application of protein structure to sequence/structure comparisons. *Methods Enzymol.* **266**: 575–598.
- Jones D.T., Miller R.T., and Thornton J.M. 1995. Successful protein fold recognition by optimal sequence threading validated by rigorous blind testing. *Proteins* **23**: 387–397.
- Jones D.T., Taylor W.R., and Thornton J.M. 1992. A new approach to protein fold recognition. *Nature* **358**: 86–89.
- Kabsch W. and Sander C. 1983. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**: 2577–2637.
- . 1984. On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations. *Proc. Natl. Acad. Sci.* **81**: 1075–1078.
- Kelley L.A., MacCallum R.M., and Sternberg M.J. 2000. Enhanced genome annotation using structural profiles in the program 3D-PSSM. *J. Mol. Biol.* **299**: 499–520.
- King R.D., Saqi M., Sayle R., and Sternberg M.J. 1997. DSC: Public domain protein secondary structure prediction. *Comput. Appl. Biosci.* **13**: 473–474.
- Kneller D.G., Cohen F.E., and Langridge R. 1990. Improvements in protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* **214**: 171–182.
- Krause A., Stoye J., and Vingron M. 2000. The SYSTERS protein sequence cluster set. *Nucleic Acids Res.* **28**: 270–272.
- Kyte J. and Doolittle R.F. 1982. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**: 105–132.
- Laskowski R.A., Hutchinson E.G., Michie A.D., Wallace A.C., Jones M.L., and Thornton J.M. 1997. PDBsum: A web-based database of summaries and analyses of all PDB structures. *Trends Biochem. Sci.* **22**: 488–490.
- Lathrop R.H., Rogers R.G., Jr., Bienkowska J., Bryant B.K.M., Buturović L.J., Gaitatzes C., Nambudripad R., White J.V., and Smith T.F. 1998. Analysis and algorithms for protein sequence-structure alignment. *New Compr. Biochem.* **32**: 237–283.
- Lemer C.M., Rooman M.J., and Wodak S.J. 1995. Protein structure prediction by threading methods: Evaluation of current techniques. *Proteins* **23**: 337–355.
- Levin J.M. 1997. Exploring the limits of nearest neighbour secondary structure prediction. *Protein Eng.* **10**: 771–776.
- Levin J.M., Robson B., and Garnier J. 1986. An algorithm for secondary structure determination in proteins based on sequence similarity. *FEBS Lett.* **205**: 303–308.
- Levitt M. 1997. Competitive assessment of protein fold recognition and alignment accuracy. *Proteins Struct. Funct. Genet.* (suppl. 1): 92–104.
- Levitt M. and Chothia C. 1976. Structural patterns in globular proteins. *Nature* **261**: 552–558.
- Lupas A. 1996. Prediction and analysis of coiled-coil structures. *Methods Enzymol.* **266**: 513–525.
- Lupas A., Van Dyke M., and Stock J. 1991. Predicting coiled coils from protein sequences. *Science* **252**: 1162–1164.
- Lüthy R. and Eisenberg D. 1991. Protein. In *Sequence analysis primer* (ed. M. Gribskov and J. Devereux), pp. 61–87. Stockton Press, New York.

- Madej T., Gibrat J.-F., and Bryant S.H. 1995. Threading a database of protein cores. *Protein Struct. Funct. Genet.* **23**: 356–369.
- Marchler-Bauer A., Levitt M., and Bryant S.H. 1997. A retrospective analysis of CASP2 threading predictions. *Proteins* (suppl. 1): 83–91.
- Martin A.C., MacArthur M.W., and Thornton J.M. 1997. Assessment of comparative modeling in CASP2. *Proteins* (suppl. 1):14–28.
- Mathews B. 1975. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **405**: 442–451.
- Minor D.L., Jr. and Kim P.S. 1996. Context-dependent secondary structure formation of a designed protein sequence. *Nature* **380**: 730–734.
- Mizuguchi K., Deane C.M., Blundell T.L., and Overington J.P. 1998a. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7**: 2469–2471.
- Mizuguchi K., Deane C.M., Blundale T.M., Johnson M.S., and Overington J.P. 1998b. JOY: Protein sequence-structure representation. *Bioinformatics* **14**: 617–623.
- Muggleton S., King R.D., and Sternberg M.J. 1992. Protein secondary structure prediction using logic-based machine learning. *Protein Eng.* **5**: 647–657.
- Murzin A.G. and Bateman A. 1997. Distant homology recognition using structural classification of proteins. *Proteins* (suppl. 1): 105–112.
- Murzin A.G., Brenner S.E., Hubbard T., and Chothia C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Nevill-Manning C.G., Wu T.D., and Brutlag D.L. 1998. Highly specific protein sequence motifs for genome analysis. *Proc. Natl. Acad. Sci.* **95**: 5865–5871.
- Nielson H., Engelbrecht J., Brunak S., and von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**: 1–6.
- Orengo C.A. and Taylor W.R. 1996. SSAP: Sequential structure alignment program for protein structure comparison. *Methods Enzymol.* **266**: 617–635.
- Orengo C., Flores T.P., Taylor W.R., and Thornton J.M. 1993. Identification and classification of protein fold families. *Protein Eng.* **6**: 485–500.
- Orengo C.A., Michie A.D., Jones S., Jones D.T., Swindells M.B., and Thornton J.M. 1997. CATH — A hierarchic classification of protein domain structures. *Structure* **5**: 1093–1108.
- Ortiz A.R., Kolinski A., and Skolnick J. 1998. Fold assembly of small proteins using Monte Carlo simulations driven by restraints derived from multiple sequence alignments. *J. Mol. Biol.* **277**: 419–448.
- Ota M. and Nishikawa K. 1997. Assessment of pseudo-energy potentials by the best-five test: A new use of the three-dimensional profiles of proteins. *Protein Eng.* **10**: 339–351.
- Ouzounis C., Sander C., Scharf M., and Schneider R. 1993. Prediction of protein structure by evaluation of sequence-structure fitness. Aligning sequences to contact profiles derived from three-dimensional structures. *J. Mol. Biol.* **232**: 805–825.
- Panchenko A.R., Luthey-Schulten Z., and Wolynes P.G. 1996. Foldons, protein structural modules, and exons. *Proc. Natl. Acad. Sci.* **93**: 2008–2013.
- Panchenko A.R., Luthey-Schulten Z., Cole R., and Wolynes P. 1997. The foldon universe: A survey of structural similarity and self-recognition of independently folding units. *J. Mol. Biol.* **272**: 95–105.
- Park J., Teichmann S.A., Hubbard T., and Chothia C. 1997. Intermediate sequences increase the detection of homology between sequences. *J. Mol. Biol.* **273**: 349–354.
- Pascarella S. and Argos P. 1992. A data bank merging related protein structures and sequences. *Protein Eng.* **5**: 121–137.
- Patthy L. 1987. Detecting homology of distantly related proteins with consensus sequences. *J. Mol. Biol.* **198**: 567–577.
- . 1996. Consensus approaches in detection of distant homologies. *Methods Enzymol.* **266**: 184–198.
- Pearson W.R. 1996. Effective protein sequence comparison. *Methods Enzymol.* **266**: 227–258.
- Pellegrini M., Marcotte E.M., Thompson M.J., Eisenberg D., and Yeatts T.O. 1999. Assigning protein functioning by comparative genome analysis: Protein phylogenetic profiles. *Proc. Natl. Acad. Sci.* **96**: 4255–4288.
- Pennisi E. 1998. Taking a structured approach to understanding proteins. *Science* **279**: 978–979.
- Peitsch M.C. 1996. ProMod and Swiss-Model: Internet-based tools for automated comparative protein modelling. *Biochem. Soc. Trans.* **24**: 274–279.

- Pongor S., Skerl V., Cserzo M., Hatsagi Z., Simon G., and Bevilacqua V. 1993. The SBASE domain library: A collection of annotated protein segments. *Protein Eng.* **6**: 391–395.
- Portugaly E. and Linial M. 2000. Estimating the probability for a protein to have a new fold: A statistical computational model. *Proc. Natl. Acad. Sci.* **97**: 5161–5166.
- Presnell S.R., Cohen B.L., and Cohen F.E. 1992. A segment-based approach to protein secondary structure prediction. *Biochemistry* **31**: 983–993.
- . 1993. MacMatch: A tool for pattern-based protein secondary structure prediction. *Comput. Appl. Biosci.* **9**: 373–374.
- Qian N. and Sejnowski T.J. 1988. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* **202**: 865–884.
- Richardson D.C. and Richardson J.S. 1994. Kinemages — Simple macromolecular graphics for interactive teaching and publication. *Trends Biochem. Sci.* **19**: 135–138.
- Rodriguez R., China G., Lopez N., Pons T., and Vriend G. 1998. Homology modeling, model and software evaluation: Three related resources. *Bioinformatics* **14**: 523–528.
- Rost B. 1995a. In *Protein folds. A distance-based approach* (ed. H. Bohr and S. Brunak), pp. 132–151. CRC Press, Boca Raton, Florida.
- . 1995b. TOPITS: Threading one-dimensional predictions into three-dimensional structures. *Ismb* **3**: 314–321.
- . 1996. PHD: Predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* **266**: 525–539.
- Rost B. and Sander C. 1993. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl. Acad. Sci.* **90**: 7558–7562.
- . 1994. Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**: 55–72.
- Rost B., Casadio R., and Fariselli P. 1996. Refining neural network predictions for helical transmembrane proteins by dynamic programming. *Ismb* **4**: 192–200.
- Rost B., Casadio R., Fariselli P., and Sander C. 1995. Transmembrane helices predicted at 95% accuracy. *Protein Sci.* **4**: 521–533.
- Russell R.B., Saqi M.A., Sayle R.A., Bates P.A., and Sternberg M.J. 1997. Recognition of analogous and homologous protein folds: Analysis of sequence and structure conservation. *J. Mol. Biol.* **269**: 423–439.
- Salamov A.A. and Solovyev V.V. 1995. Prediction of protein secondary structure by combining nearest-neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**: 11–15.
- . 1997. Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **268**: 31–36.
- Sali A., Potterton L., Yuan F., van Vlijmen H., and Karplus M. 1995. Evaluation of comparative protein modeling by MODELLER. *Proteins* **23**: 318–326.
- Salzberg S. and Cost S. 1992. Predicting protein secondary structure with a nearest-neighbor algorithm. *J. Mol. Biol.* **227**: 371–374.
- Sander C. and Schneider R. 1991. Database of homology-derived protein structures. *Proteins Struct. Funct. Genet.* **9**: 56–68.
- Sayle R.A. and Milner-White E.J. 1995. RASMOL: Biomolecular graphics for all. *Trends Biochem. Sci.* **20**: 374.
- Schultz J., Milpetz F., Bork P., and Ponting C.P. 1998. SMART, a simple modular architecture tool: Identification of signaling domains. *Proc. Natl. Acad. Sci.* **95**: 5857–5864.
- Sippl M.J. 1990. Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *J. Mol. Biol.* **213**: 859–883.
- Sippl M.J. and Weitckus S. 1992. Detection of native-like models for amino acid sequences of unknown three-dimensional structure in a data base of known protein conformations. *Proteins* **13**: 258–271.
- Sippl M.J., Lackner P., Domingues F.X., and Koppensteiner W.A. 1999. An attempt to analyse progress in recognition from CASP1 to CASP3. *Proteins (suppl. 3)* **37**: 226–230.
- Sonnhammer E.L., Eddy S.R., Birney E., Bateman A., and Durbin R. 1998. Pfam: Multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.* **26**: 320–322.
- Stolorz P., Lapedes A., and Xia Y. 1992. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* **225**: 363–377.
- Stultz C.M., White J.V., and Smith T.F. 1993. Structural analysis based on state-space modeling. *Protein Sci.* **2**: 305–314.

- Stultz C.M., Nambudripad R., Lathrop R.H., and White J.V. 1997. Predicting protein structure with probabilistic models. *Adv. Mol. Cell Biol.* **22B**: 447–506.
- Sudarsanam S. 1998. Structural diversity of sequentially identical subsequences of proteins: Identical octapeptides can have different conformations. *Proteins* **30**: 228–231.
- Swindells M.B., Orengo C.A., Jones D.T., Hutchinson E.G., and Thornton J.M. 1998. Contemporary approaches to protein structure classification. *BioEssays* **20**: 884–891.
- Tatusov R.L., Koonin E.V., and Lipman D.J. 1997. A genomic perspective on protein families. *Science* **278**: 631–637.
- Taylor W.R. and Orengo C.A. 1989. Protein structure alignment. *J. Mol. Biol.* **208**: 1–22.
- von Heijne G. 1987. *Sequence analysis in molecular biology — Treasure trove or trivial pursuit*, pp. 81–121. Academic Press, San Diego, California.
- Vriend G. and Sander C. 1991. Detection of common three-dimensional substructures in proteins. *Proteins* **11**: 52–68.
- White J.V., Stultz C.M., and Smith T.F. 1994. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math. Biosci.* **119**: 35–75.
- Wilmanns M. and Eisenberg D. 1993. Three-dimensional profiles from residue-pair preferences: Identification of sequences with  $\beta/\alpha$ -barrel fold. *Proc. Natl. Acad. Sci.* **90**: 1379–1383.
- . 1995. Inverse protein folding by the residue pair preference profile method: Estimating the correctness of alignments of structurally compatible sequences. *Protein Eng.* **8**: 627–639.
- Wu C. 1996. Gene classification artificial neural system. *Methods Enzymol.* **266**: 71–88.
- Wu C., Zhao S., and Chen H.L. 1996. A protein class database organized with ProSite protein groups and PIR superfamilies. *J. Comput. Biol.* **3**: 547–561.
- Xenarios I., Rice D.W., Salwinski L., Baron M.K., Marcotte E.M., and Eisenberg D. 2000. DIP: The database of interacting proteins. *Nucleic Acids Res.* **28**: 289–291.
- Yi T.M. and Lander E.S. 1993. Protein secondary structure prediction using nearest-neighbor methods. *J. Mol. Biol.* **232**: 1117–1129.
- . 1996. Iterative template refinement: Protein-fold prediction using iterative search and hybrid sequence/structure templates. *Methods Enzymol.* **266**: 322–339.
- Yona G., Linial N., and Linial M. 1999. ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins* **37**: 360–378.
- Zhang X., Mesirov J.P., and Waltz D.L. 1992. Hybrid system for protein secondary structure prediction. *J. Mol. Biol.* **225**: 1049–1063.