Chapter 6
Predicting secondary structure from protein sequences

Jaap Heringa
National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK

1 Introduction

Protein structure is intrinsically hierarchic in its internal organization. The highest level in this hierarchy is constituted by complete proteins or assemblies of such proteins, which become subdivided through domains via super-secondary structure to secondary structure at the lowest hierarchical level.

At higher levels within in this hierarchy, especially from the domain level upwards, the connectivity of the polypeptide backbone between substructures becomes less important. A protein thus can retain a stable structure irrespective of the sequential arrangement of domains and presence of fragments linking them together. Such linker regions often constitute exposed surface loops that do not disrupt the folds of the domains they connect (1).

At the level of protein secondary structure, however, the elements are not only crucially dependent on their amino acid compositions, but, unlike domain and higher-order structures, are also very much context dependent; i.e. they rely critically on the substructures in their environment. It is because of this context dependency, that predicting protein secondary structure is a very difficult task, which after three decades of research has not attained the accuracy on which further prediction of tertiary structure can be based. It must be stressed, however, that some successful predictions of higher-order structure, based on a knowledge of the secondary structure, have been achieved (e.g. ref. 2).

This chapter covers some background aspects of secondary structure prediction and describes recent and successful prediction methods, most of which are available through the World Wide Web and so can be used by virtually every biologist who likes to find out about the secondary structure associated with a particular protein query sequence.

1.1 What is secondary structure?

Perhaps a suitable definition in the context of this chapter for a secondary structure is that it is a consecutive fragment in a protein sequence, which corresponds to a local region in the associated protein structure showing

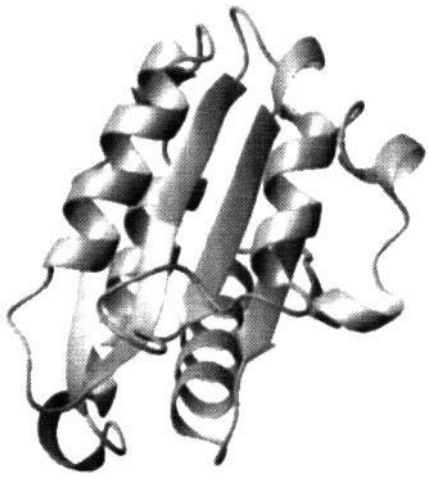

Figure 1 Ribbon diagram of the flavodoxin fold (PDB code 5nul) belonging to the α/β doubly-wound fold family.

distinct geometrical features. The two basic secondary structures are the α-helix and the β-strand. Both show distinct structural features and are easily recognizable in a protein structure (*Figure 1*). Other secondary structure types occurring in protein structures are more difficult to classify as they are less regular than α-helices or β-strands. Such structures are defined in the context of most prediction methods as coil; i.e. leftover secondary structures that cannot be considered in the α-helical or β-stranded conformation.

In general, about 50% of the amino acids fold into α-helices or β-strands, so that roughly half the protein structures are irregularly shaped. The primary reason for the regularity observed for helices and strands is the inherent polar nature of the protein backbone, which contributes a polar nitrogen and oxygen atom for each amino acid. To satisfy energetical constraints, the parts of the main-chain buried in the internal protein core need to form hydrogen-bonds between those polar atoms. The α-helix and β-strand conformations are optimal as each main-chain nitrogen atom can associate with an oxygen partner (and *vice versa*) whenever they adopt one of these two secondary structure types. It must be stressed that, in order to satisfy their hydrogen-bonding constraints, β-strands need to interact with other β-strands, which they can do in a parallel and anti-parallel fashion, thus forming a β-pleated sheet. β-strands thus depend on crucial long-range interactions between residues remote in sequence. They therefore are more context dependent than α-helices, which would be more able to fold 'on their own'. The fact that the vast majority of prediction methods have greatest difficulty in delineating β-strands correctly is believed to be due to their pronounced context dependency.

1.2 Where could knowledge about secondary structure help?

Experimental evidence on early protein folding intermediates has shown that secondary structural elements form at early stages during the folding process (for a review, see ref. 3). These results support the significance of the so-called 'framework' model of protein folding (4, 5), where two or more secondary struc-

tural elements would associate early during folding to provide a structural frame to which subsequently other substructures could attach. Therefore, knowledge of protein secondary structural regions along the sequence is a prerequisite to model the folding process or kinetics associated with it. Also for tertiary model building, the ability to predict the secondary structure from the sequence alone is crucial, as it allows for docking experiments to be carried out on the predicted α-helices and β-strands.

On the architectural side of protein structure, it is possible to recognize the three-dimensional topology by comparing the successfully predicted secondary structural elements of a query protein with a database of known topologies. Successful prediction here means parts of those helices and strands essential for the topology would have been predicted, without necessarily accurate prediction of the edges of those structures or the detection of non-essential secondary structures. An example of topologically essential secondary structures for the flavodoxin fold is given in *Figure 2*. The figure shows a schematic representation

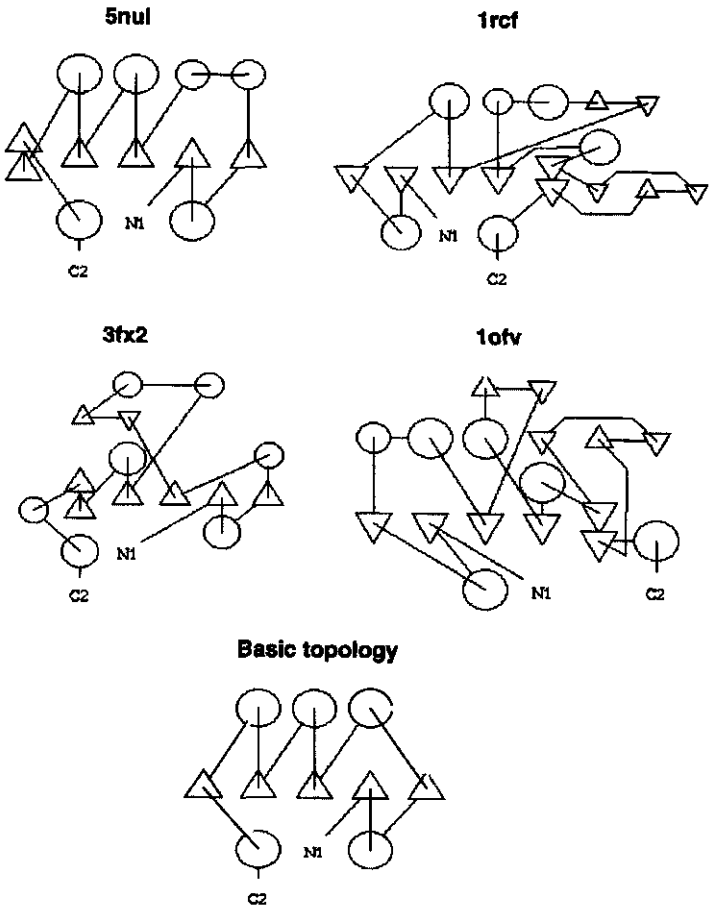

Figure 2 TOPS diagrams for four flavodoxin structures and their basic topology. The essential secondary structures are given in the basic topology diagram.

of the secondary structures as provided by the TOPS server (http://tops.ebi.ac.uk). In a TOPS diagram (6), a α-helix is represented by a circle and a β-strand by a triangle. The flavodoxin fold belongs to the class of α/β-folds with the essential secondary structures distributed over the sequence as $[\beta\alpha]_5$. The five strands fold into a single β-pleated sheet ordered topologically as β_2-β_1-β_3-β_4-β_5, where the numbers indicate their relative position in the sequence and hyphens the hydrogen bonded and spatial interactions between the strands. The five α-helices, each following a β-strand, shield the β-sheet from the solvent and therefore are of an amphipathic nature (see below). From the topologies of a few different flavodoxin structures (*Figure 2*) can be seen that varying substructures can be added on to the basic structure, albeit they do not disrupt the fold of the topologically essential secondary structures. Therefore, proper prediction of the sequential order of the topologically essential helices and sheets often allows the recognition of the fold type associated with the protein sequence considered, thereby conferring the information pertaining to that fold. Furthermore, active sites of enzymes typically comprise amino acids positioned in loops, so that, for example, identically conserved residues at multiple alignment sites predicted to be in loop regions (i.e. not predicted as α-helix or β-strand), could be functional and together elucidate the function of the protein (or protein family) under scrutiny.

1.3 What signals are there to be recognized?

A number of observations on secondary structures as found in the large collection of protein structures deposited in the Protein Data Bank (PDB) (7), could be summed up for each of the secondary structures α-helix, β-strand, and loop as follows.

α-helix:

(a) As the number of residues per turn is 3.6 in the ideal case and helices are often positioned against a buried core, they have one phase contacting hydrophobic amino acids, while the other phase interacts with the solvent. Such amphipathic helices (8) thus show a periodicity of three to four residues in hydrophobicity of the associated sequence stretch (*Figure 3*).

(b) Proline residues do not occur in middle segments as they disrupt the α-helical turn. However, they are seen in the first two positions of α-helices.

β-strand:

(a) β-Strands mostly fold into so-called β-pleated sheets which have two strands forming either edge. Therefore the hydrophobic nature of edge strands is different from that of strands internal to a β-sheet. As side-chains of constituent residues along a β-strand alternate the direction in which they protrude, edge strands of β-sheets can show an alternating pattern of hydrophobic-hydrophilic residues, while buried strands tend to contain merely hydrophobic residues (*Figure 3*).

(b) As β-strand is the most extended conformation (i.e. consecutive Cα atoms are farthest apart), it takes relatively few residues to cross the protein core with a strand. Therefore, the number of residues in a β-strand is usually limited and can be anything from two or three amino acids, whereas helices shielding such strands from solvent comprise more residues.

(c) The β-strands can be disrupted by single residues that induce a kink in the extended structure of the main-chain. Such so-called β-bulges are often comprised of relatively hydrophobic residues.

Coil:

(a) Multiple alignments of protein sequences often display gapped and/or highly variable regions, which would be expected to be associated with loop regions rather than the two basic secondary structures.

(b) Loop regions contain a high proportion of small polar residues like Ala, Gly, Ser, and Thr. Glycine residues are seen in loop regions due also to their inherent *flexibility*.

(c) Proline residues are often seen in loops as well. They are not observed in helices and strands as they kink the main-chain, although they can occur in the N-terminal two positions of α-helices as mentioned above.

In addition to the positional requirements in hydrophobicity, there are also general compositional differences between helix, strand and coil conformations and this is the signal used in many of the early prediction methods (see below) for single sequences. Methods that utilize multiple alignments can also exploit the fact that the amino acid exchange patterns are different for the three secondary structure states.

A few additional rules can help in clarifying the structure or function of a protein sequence, once the secondary structure is predicted:

(a) Hydrophobic and particularly conserved hydrophobic residues are normally buried in the protein core.

(b) More than 95% of all so-called β-α-β motifs; i.e. a β-strand followed in sequence by a α-helix and another β-strand, show a right-handed chirality. The aforementioned flavodoxin family (*Figure 2*) indeed shows only right-handed β-α-β motifs. This fact can be used to build a topology for the secondary structures of the sequence(s) considered.

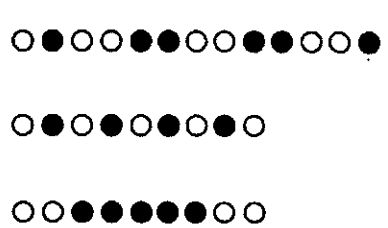

Figure 3 Hydrophobic patterns along secondary structures.

(c) Helices often cover up a core of β-strands. Therefore, if both α-helices and β-strands are predicted, an attempt should be made to distribute the helices evenly at either phase of a tentative β-sheet in topology modelling.

(d) As mentioned, strictly conserved residues in different regions of a multiple alignment can be predicted with great confidence to be responsible for the catalytic functions, particularly if they are polar and predicted to be in loop structures hence unlikely to be buried. As active site residues are positioned together in a protein 3-D structure, the coil structures they constitute should be brought together in a topology model.

2 Assessing prediction accuracy

The most widely used way to assess the quality of an alignment is by calculating the overall per residue three-state accuracy, called the Q_3:

$$Q_3 = [(PH + PE + PC)/N] \times 100\%,$$

where N is the total number of residues predicted and PS is the number of correctly predicted residues in state S ($S = H$, E, or C). Some researchers use the so-called Matthews' correlation coefficient as it more stringently estimates the prediction accuracy for each structural state:

$$C_S = \frac{(P_S + N_S) - (\sim P_S \times N_S)}{\sqrt{(P_S + \sim P_S) \times (P_S + \sim N_S) \times (N_S + \sim P_S) \times (N_S + \sim N_S)}}$$

where P_S and N_S are respectively the number of positive and negative cases correctly predicted for the structural state considered, and $\sim P_S$ and $\sim N_S$ the numbers of false positives and negatives, respectively. Three-state predictions would thus yield three Matthews' correlation coefficients. If overprediction or underprediction occurs for any of the structural states, this is more dramatically reflected in the Matthews' correlations than in the Q_3 percentage. A third way to assess prediction accuracy is by *weights of evidence*, defined for each secondary structural type S as:

$$W_S = \log[(P_S \times N_S) / (\sim P_S \times \sim N_S)].$$

Although this measure is relatively robust to different sampling frequencies of the structural states, the interpretation of the resulting values is not as straightforward as for the other two measures. Because understanding the Q_3 measure is the easiest and its use leads to just one percentage, it is the measure most frequently used in the literature to report prediction accuracy.

A very important issue in assessing performance is the notion of sustained accuracy. Knowledge about the average accuracy of a given method over a set of predicted proteins is not meaningful if unaccompanied by the variance of those predictions. It is important to know what worse case predictions can be expected from a method, even if its mean accuracy is quite high.

A standard scenario to assess prediction accuracy is the *jackknife* test carried out over a large set of test proteins (see *Protocol 1*). This ensures that no infor-

mation about a query sequence or multiple alignment is used in training the method. Nonetheless, unnoticed but systematic tuning of the method to the database might still occur, so that the most rigorous test of any method is the prediction of test cases that have no homologues in the database and have not been seen during the development of the method.

Notwithstanding the importance of the measures for accuracy as listed above, the real success of a secondary structure prediction depends on how the knowledge is being used. An example is the aforementioned fold recognition, where correct prediction of the edges of secondary structural elements is not essential, but missing structures that are crucial for the basic topology is costly. However, all above measures, equally penalize, for example, missing two residues at either side of a seven-amino acid strand or missing a complete topologically essential strand of four residues.

Protocol 1

Jackknife testing

1 Take out one protein of the complete set of N proteins.

2 Train the method on the remaining N–1 proteins (the training set).

3 Predict the secondary structure for the protein taken out.

4 Repeat step 1–3 for all N proteins and assess the average accuracy.

It is possible to test the method by averaging the predictions over all combinations of x proteins ($1<x<N$), each time using the method trained on the remaining N–x proteins. This provides an impression of the influence of different training sets on the sustained accuracy of a single protein being predicted. As the number of combinations grows rapidly with x, the training phase of most methods is too slow for extensive testing using this mode. It can, however, also be used to save computation time if the database is split evenly in test groups of sequences (e.g. 9), as each sequence within a test group is associated with a single training set, thus saving training overhead.

An additional problem in secondary structure prediction is the standard of truth. Most prediction methods are assessed in accuracy by using known tertiary structures from the protein data bank (7) with their secondary structural elements assigned using the DSSP method of Kabsch and Sander (10). Colloc'h *et al.* (11) compared three such secondary structure determination algorithms, among which was the DSSP method, and found significant differences in their secondary structural assignments. This ambiguity in secondary structural assignments can be dramatic for particular proteins where agreement of the methods can be as low as 65% (12, 13). Moreover, in structurally equivalenced sets of homologous proteins with known tertiary structure, the corresponding secondary structural elements can vary in length or show shifts of one to a few residues, and hence a realistic maximum prediction accuracy per residue would be in the range 80–100% (14). Many researchers have suggested that prediction evaluation should be

based on the overlap of predicted and observed segments rather than on individual positions (15–20). A recent secondary structure assignment program that combines many of the features of earlier methods, such as checking hydrogen bonding patterns and stereochemical characteristics, is the knowledge-based method STRIDE (21), claimed to yield assignments in close agreement to those made by crystallographic experts.

3 Prediction methods for globular proteins

3.1 The early methods

Attempts to predict protein secondary structure began more than four decades ago (e.g. 22, 23), while the first computer algorithms appeared a quarter of a century ago (24–26). The algorithms of Nagano (22) and Chou and Fasman (25) were based on statistical information, whereas Lim's method (26) was stereochemically oriented and relied on conserved hydrophobic patterns in secondary structures such as amphipathicity in helices (8). Secondary structure prediction has generally been formulated for three states, helix, strand, and coil. This holds also for recent versions of the early and popular GOR method (27, 28), which considers the influence and statistics of flanking residues on the conformational state of a selected amino acid to be predicted. The popular early methods by Lim (26) and Chou–Fasman (25) as well as the GOR method (27, 28) will be described in more detail.

3.1.1 Lim

Lim (26) developed a set of complicated stereochemical prediction rules for α-helices and β-sheets based on their packing as observed in globular proteins. Apart from being the most successful early method (see below), Lim's stereochemical rules are quite important for understanding protein folding. An example is the set of hydrophobicity rules for α-helices with terminal hydrophobic pairs at sequence positions i and $i + 1$, hydrophobic pairs in middle helical segments positioned at $(i, i + 4)$ and middle hydrophobic triplets positioned at $(i, i + 1, i + 4)$ or $(i, i + 3, i + 4)$ (see also *Figure 3*). The Lim method never gained widespread popularity because a computer implementation has not been available until recently.

3.1.2 Chou–Fasman

The most widely used pioneering method is the one by Chou and Fasman (25), in which predictions are based on differences in residue composition for three states of secondary structure: α-helix, β-strand, and turn (i.e. neither α-helix nor β-strand). Chou and Fasman performed a statistical analysis over a number of crystallographically determined protein tertiary structures and determined the frequency of each amino acid type in the three states. The position of turn residues was included in the frequency calculations given significant positional differences in residue type occurrences at turn sites. The frequencies were

normalized to amino acid type preferences for each of the structural states by dividing each by that found in all positions of the known structures. For helix and strand, effects of neighbouring residues in the protein sequence were taken into account by averaging the preferences over three residues for α-helix predictions and over two for β-strands. Secondary structures were initiated according to the higher preference values and minimum nucleation lengths required for each structural state. Extensions were effected as long as preferences remained high and certain residues were not encountered (e.g. proline in a α-helix). The Chou–Fasman method has owed its early popularity to the straightforward underlying statistics that are easy to understand.

3.1.3 GOR

The GOR method quickly became the standard for a decade after its first appearance. Although the initial versions GOR I and GOR II predicted four states by discriminating between coil and turn secondary structures, GOR III (28) and the most recent version, GOR IV (29) perform the common three-state prediction. The GOR method relies on the frequencies observed for residues in a 17-residue window (i.e. eight residues N-terminal and eight C-terminal of the central window position) for each of the three structural states. The amino acid frequencies are exploited using an information function based on conditional probabilities defined as:

$$I(S; R) = \log[P(S|R)/P(S)],$$

where S one of the structural states H, E, or C, and R is one of the 20 residue types. The factor $P(S|R)$ denotes the conditional probability of a secondary structural state for a sequence position given that it is occupied by residue type R. Rewriting the formula for frequencies gives:

$$I(S; R) = \log[(f_{S,R}/f_R) / (f_S/N)],$$

where $f_{S,R}$ is the frequency of residue type R in state S, f_R the general frequency of residue type R, and $f_{S/N}$ that of structural state S. Significant in this formula is that the information of a particular residue type in one of the structural states is not only based on the normalized frequency, but shows an extra weighting based on the inverse fraction of all residues in that state. In the GOR method, this formula is used to calculate the information difference between the various states defined as $I(\Delta S; R) = I(S; R)—I(!S; R)$ with !S denoting all other states (*not* S). The information difference formula then becomes:

$$I(S; R) = \log[f_{S,R}/f_S] - \log[f_{!S,R}/f_{!S}].$$

The above formula is defined for a single sequence position, but can be easily extended to the GOR 17-residue window by, for example, writing R_{17} instead of R. Unfortunately, it is not feasible to sample all possible 17-residue fragments directly from the PDB (as there are 17^{20} possibilities). The subsequent versions of

the GOR method over the years have explored increasingly detailed approximations of this sampling problem, along with the increase of data in the PDB:

(a) GOR I just treated the 17 positions in the window independently, and so single-position information could be summed over the 17-residue window.

(b) GOR II did the same but sampled over a larger database.

(c) GOR III (28) refined by including pair frequencies derived from 16 pairs between each non-central and the central residue in the 17-long window. As the PDB at the time was not large enough to provide sufficient data, dummy frequencies were calculated (28).

(d) The current version, GOR IV (29) uses pairwise information over all possible paired positions in a window (there are $17 \times 16/2$ possibilities), albeit with a relatively small weight as compared with the GOR I-type single-position information (a) which is included as well.

The theoretical principles used in the GOR method are statistically sound and no *ad-hoc* rules or artificial variables are invoked, which makes it one of the most elegant methods with a high accuracy given its single sequences prediction. However, as in many other methods (*vide infra*), a post-processing step was introduced for the GOR IV method to refine the predictions. Helices are required to be at least four residues in length and strands should consist of two or more residues. If a shorter helix or strand fragment is initially predicted, the method assesses the probabilities of extending the fragment to the minimum associated length or deleting it (i.e. changing it to coil).

3.2 Accuracy of early methods

The Chou–Fasman, GOR III, and Lim methods were assessed to show accuracies of 50%, 53%, and 56% respectively (30). Version IV of the GOR method, however, raises the single sequence prediction accuracy to 64.4% (29), as assessed through jackknife testing (see *Protocol 1*) over a database of 267 proteins with known structure. Random prediction would yield about 40% correctness given the observed distribution of the three states in globular proteins (with roughly 30% helix, 20% strand, and 50% coil). Although they are significantly beyond the random level, these single-sequence prediction accuracies are not sufficient to allow the successful prediction of the protein topology.

3.3 Other computational approaches

The Chou–Fasman and GOR methods both exploit compositional biases exhibited by the three types of secondary structures. Information derived from single sequences have been explored as well in the form of sequence pattern matching (16, 31–34).

On the algorithmic side, researchers have integrated novel computational concepts to optimize the implementation of observed patterns in mapping the

primary on to the secondary structure and to thus enhance the success rate of prediction. These include:

(a) Neural network applications (9, 35).

(b) Nearest-neighbour methods (36–39).

(c) Linear discriminant analysis (40).

(d) Inductive logic programming (ILP) (41).

Examples of the first three formalisms will be described in the following section. The latter computational concept (ILP) is designed for learning structural relationships between objects. Muggleton *et al.* (41) used the ILP computer program Golem to automatically describe qualitative rules for residues in the α-helix conformation and central in a 9-residue window. The rules made use of the physicochemical amino acid characterizations of Taylor (42) and were established during iterative training steps over a small set of 12 known α/α protein structures. With the thus obtained set of rules, α-helices in four independent α/α proteins were predicted with an accuracy of 81% on a per residue basis (Q_3). The Golem algorithm is of limited use because it is only able to predict helices in all-helical proteins.

3.4 Prediction from multiply-aligned sequences

In 1987, Zvelebil *et al.* (43) for the first time exploited multiple alignments to predict secondary structure automatically by extending the GOR method and reported that predictions were improved by 9% compared to single sequence prediction. Also Levin *et al.* (44) quantified the effect and observed 8% increased accuracies when multiple alignments of homologous sequences with sequence identities of ≥ 25% were used. As a result, the current state-of-the-art methods all use input information from multiple sequence alignments.

3.4.1 Neural network methods

Neural networks are organized as interconnected layers of input and output units, and can also contain intermediate (or 'hidden') unit layers (for a review, see ref. 45). Each unit in a layer receives information from one or more other connected units and determines its output signal based on the weights of the input signals. A neural network can be regarded as a black box, which is trained to optimize the grouping of a set of input patterns into a set of output patterns by adjusting the weights of the internal connections. Therefore, neural nets are learning systems based upon complex non-linear statistics.

PHD

The PHD method (**P**rofile network from **HeiD**elberg) (9) combines the added information from multiple sequence information with the optimization strength of the neural network formalism. The method makes use of three consecutive complete neural networks:

(a) The first network produces the first raw 3-state prediction for each alignment position. It takes as input the fractions of the 20 amino acids at each multiple

alignment position together with the two 6-residue flanking regions; i.e. a 13-residue window ($w = 13$) is used to predict each alignment position with the central residue in the middle position. The output of the first network for each alignment position is three probabilities for three the states (helix, strand, and coil).

(b) A second network refines the raw predictions of the first level by filtering the 3-state probabilities for each alignment position based on the probabilities of the flanking positions. It takes as input the output of the first network and processes the information using a 17-residue window. The output of the second network comprises for each alignment position the three adjusted state probabilities. This post-processing step for the raw predictions of the first network is aimed at correcting unfeasible predictions and would, for example, change (HHH**EE**HH) into (HHH**HH**HH).

(c) The first two networks perform the basic prediction of the secondary structure associated with a query multiple alignment. However, as the networks can be trained in various ways, PHD employs a number of separately trained consecutive network pairs ((a) and (b)) and feeds their predictions (3-state probabilities) into a third network for a so-called jury decision.

The predictions obtained by the jury network undergo a final filtering to delete predicted helices of one or two residues and changing those into coil. The method was trained on a non-redundant set of 130 alignments from the HSSP database (46), each containing one sequence with a known structure. The method showed an overall prediction accuracy of 70.8% in a jackknife test over 126 alignments (4 of 130 alignments were transmembrane protein families), which for computational reasons were divided in 7 groups (see *Protocol* 1). Although this count is not the highest accuracy reported, the PHD method to date shows the most sustained performance as compared with all other methods available on the Web.

If the PHD webserver is given a single sequence for prediction, it performs a BLAST-search to find a set of homologous sequences and aligns those using the MAXHOM alignment program (46). The resulting alignment is then fed into the actual PHD neural net algorithm.

Pred2ary

Another accurate profile and neural net-based prediction method is Pred2ary (35) which was assessed with an accuracy of 74.8% and balanced prediction over the three structural states. The method employs a second neural net to filter the raw predictions of the first net, as does the PHD method (9). A recent extended version, which combines in a jury decision the outputs of a massive number of 120 networks individually trained, is claimed to predict 75.9% ± 7.9% accurately. This is achieved by constructing *a priori* probabilities of correctly predicting the structural state at each query sequence position for all combinations of network output weighs for helix and strand. These probabilities are then used for a final state prediction corresponding to the highest of the *a priori* probabilities for each of the three states.

3.4.2 k-nearest neighbour methods

As with neural network methods, the application of a *k*-nearest neighbour method requires an initial training phase in which a large pool of so-called exemplars is established. In the context of secondary structure prediction, this pool typically consists of sequence fragments of a certain length derived from a database of known structures, so that the central residue of such fragments (exemplars) can be assigned the true secondary structural state as a label. Then, a window of the same length is slid over the query sequence and for each window the *k* most similar fragments from the pool of exemplars are determined using a similarity criterion. The distribution of the *k* secondary structure labels is then used to derive propensities for the three states. In the methods covered below, *k* is in the range 25–100.

Yi and Lander

Yi and Lander (36) were the first to use nearest-neighbour classifiers for prediction of secondary structure. A database of 110 proteins with known tertiary structure was used to derive a large collection of 19-residue exemplars for which only the environmental states were noted; i.e. the residue type information was discarded. As a label for each exemplar the secondary structural state of the central residue was taken. For each 19-residue window of the query protein, 50 nearest neighbour exemplars were identified using the amino acid environmental scoring system of Bowie *et al.* (47), which includes as parameters the secondary structure state, accessible surface area and polarity; and scores the likelihood of a residue type to be in a particular state (or range) over these three parameters. As a score, the average was taken of 19 residues within a query window matched with the 19-position exemplar considered. During training, for each exemplar a cut-off score was determined, which should be met by the query fragment compared to it in order to count the exemplar as a neighbour: The cut-off score can be viewed as a reliability check for the predictive value of the exemplars. The 50 thus obtained nearest neighbours showed a distribution of the associated secondary structure labels, from which probability estimates for the three structural states were derived for the query fragment considered. Yi and Lander explored various scoring systems and found that the best performer included 15 environmental classes (3 secondary structures times 5 different accessibility/polarity classes) combined with an amino acid exchange score from the Gonnet *et al.* matrix (48). Note that for this final scoring system, the amino acid types of the exemplars were taken into account. This scenario resulted in a prediction accuracy of 67.1%. Using a neural network for a jury decision over six different scoring systems led to the final accuracy of 68%, as assessed through jackknife testing (*Protocol 1*).

NNSSP

The NNSSP (Nearest Neighbour Secondary Structure Prediction) (37) method adopts the nearest neighbour approach of Yi and Lander (36) for single sequence prediction. Differences with the Yi and Lander method are:

(a) Predictions are made for multiple alignments.

(b) N- and C-terminal positions of helices and strands; and β-turns are explicitly taken as additional secondary structure types.

(c) When predicting, the database of exemplars (see above) is restricted to sequences similar to the query sequence. This reduces computation and leads to more biologically related nearest neighbours.

(d) Alignment regions with insertions/deletions are explicitly taken into account.

Salamov and Solovyev (37) explored various window lengths and finally choose predictors combining window sizes of 11, 17, or 23; nearest neighbour numbers of 50 or 100, and balanced or non-balanced training (i.e. $3 \times 2 \times 2 = 12$ predictors). A simple majority rule over the 12 predictors increased the accuracy by 0.9%. A few simple filters were effected to refine the thus obtained predictions as follows:

(a) Helices predicted to consist of 1 or 2 residues are deleted (changed to coil), but (EHE) becomes (EEE).

(b) Strands of length 1 or 2 are deleted, but (HEEH) becomes (HHHH).

(c) Helices of length 4 or less are deleted. This rule is applied after a full cycle of rule (a) and (b).

The overall accuracy of the method is 72.2%, which results from a jackknife test over the database of 126 proteins by Rost and Sander (9).

PREDATOR

The PREDATOR method of Frishman and Argos (38, 39) owes its accuracy mostly to the incorporation of long-range interactions for β-strand prediction and attains 68% prediction accuracy for single sequence prediction which was assessed using a one-at-a time jackknife test (see *Protocol 1*) over the protein set of Rost and Sander (RS) (9). Using a *k*-nearest neighbour approach (with $k = 25$ and 13-residue windows), propensities for the general three states (P^H, P^E, and P^C) were determined for each residue. Using pairwise potentials involving long-range interactions, two more propensities for β-strand were determined. This was done by assessing the likelihood for all pairwise 5-residue fragments (separated by more than six amino acids) to form parallel or anti-parallel β-bridges, based on summing residue hydrogen bonding propensities obtained from known structures (two sets of propensities for anti-parallel and one for parallel bridges). As the final parallel and anti-parallel β-strand propensity for each residue (P^{Par} and $P^{Antipar}$), the maximum scoring window pair was taken with the residue considered at the N-terminal position in one of the windows. Pairwise hydrogen bonding potentials were also determined for α-helical residues at a sequence separation of four residues. Their sum was calculated over a 7-residue window to arrive at an extra helix propensity for the residue N-terminal in the window (P^{Helix}). The last additional propensity concerned β-turns (P^{Turn}) and was obtained by summing single-residue propensities in classic β-turn positions 1–4 (49) using

a four-residue window. For each of the thus obtained seven independent propensities, threshold values (T) were calculated and used in the following five rules applied consecutively to arrive at a three-state prediction for each residue:

1. If ($P^{Par} > T^{Par}$ or $P^{Antipar} > T^{Antipar}$) and $P^{Helix} < T^{helix}$, then predict β-strand; otherwise, if $P^{Helix} > T^{helix}$, then predict α-helix, otherwise predict coil.
2. If $P^{C} > T^{C}$, then predict coil.
3. If $P^{E} > T^{E}$, then predict β-strand.
4. If $P^{H} > T^{H}$, then predict α-helix.
5. If $P^{Turn} > T^{Turn}$, then predict coil.

Apart from the novel scheme to employ long-range interaction to aid strand prediction, the PREDATOR method can also use information from multiple sequences to enhance predictions. However, PREDATOR does not use or construct a multiple alignment, but rather compares the sequences using pairwise local alignments (50). The current method is not able to extract local alignments from a multiple alignment provided by the user, while leaving the multiple alignment intact, but it is planned to realize this option in a future release (Frishman, personal communication). As predictions by PREDATOR are carried out for a single base sequence, a set of highest scoring local alignments is compiled through matching the base sequence with each of the other sequences. A weight is then compiled for each matched local fragment based on the alignment score and length of the local alignment. For each residue in the base sequence, the weighted sum over all stacked fragments (see *Figure* 4) is compiled independently for the seven propensities and subjected to the above five rules to arrive at a three-state prediction. The extra information conferred by the multiple sequences resulted in a Q_3 of 74.8% (39), as assessed using one-at-a-time jackknife testing over the RS protein set. As for the Pred2ary method showing identical accuracy, this Q_3 is the highest reported in the literature.

3.4.3 Linear discriminant analysis: the DSC method

The DSC method combines the compositional propensities from multiple alignments with a set of concepts important for secondary structure prediction

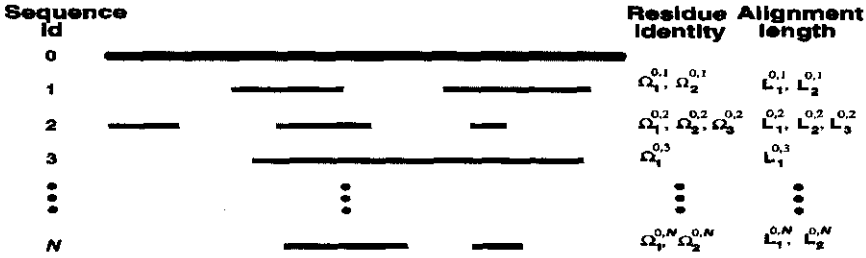

Figure 4 Usage of local alignments in the PREDATOR algorithm. For details, see text.

(see Section 1.3). This information is processed using linear statistics. Apart from the conformational propensities, the following concepts are used:

- N-terminal and C-terminal sequence fragments are normally coil.
- Moments of hydrophobicity (see *Figure* 3).
- Alignment positions comprising gaps are indicative for coil regions.
- Moments of conservation.
- Autocorrelation.
- Residue ratios in the alignment.
- Feedback of predicted secondary structure information.
- Simple filtering.

The relative importance of these concepts was determined in five runs, which successively relied on increased information as follows:

(a) **Run 1**: The GOR method was used on each of the aligned sequences and the average GOR score for each of the three states was compiled for each alignment position.

(b) **Run 2**: For each position in the query multiple alignment, a so-called attribute vector was compiled, consisting of 10 attributes: three averaged GOR scores for H, E, and C; distance to alignment edge; hydrophobic moment assuming helix; hydrophobic moment assuming strand; number of insertions; number of deletions; conservation moment assuming helix and that assuming strand.

(c) **Run 3**: Positional 20-attribute vectors were determined consisting of the above 10 attributes and those in a smoothed fashion.

(d) **Run 4**: Positional 27-attribute vectors were compiled comprising the 20 attributes of the preceding round, combined with fractions of predicted α-helix and β-strand, and fractions of the five most discriminating residue types; His, Glu, Gln, Asp, and Arg.

(e) **Run 5**: A set of 11 filter rules were employed for a final prediction, such as, for example, ([E/C]CE[H/E/C][H/C]) → C. These filter rules were found automatically using machine learning.

For run (b) to (d), a linear discrimination function was determined for each of the three secondary structural states. A linear discrimination function is effectively a set of weights for the attributes in the positional vector, so that the secondary structure associated with the highest scoring discrimination function is assigned to the alignment position considered.

The DSC predictions are based on the information arising from the five above runs. The Q_3 was assessed for successively increasing numbers of runs (run 1, runs 1 and 2, runs 1–3, 1–4, 1–5) for the five runs based on the Rost–Sander protein set and comprised 63.5%, 67.8%, 68.3%, 69.4%, and 70.1% (DSC), respectively. The DSC method performs especially well for moderately sized proteins in the range 90–170 residues. A special feature of the DSC technique is that it

accepts predictions by the PHD algorithm as input and attempts to refine those using the above concepts. The Q3 of this PHD-DSC combinatorial procedure was evaluated at 72.4% (40).

3.4.4 SSPRED: a secondary structure specific exchange method

The SSPRED method (51) exploits an alternative aspect of the positional information provided by multiple alignment, in that it uses the amino acid pairwise exchanges observed for each multiple-alignment positions. Using the 3D-ALI database (52) of combined structural and sequence alignments of distantly homologous proteins, three amino acid exchange matrices were compiled for helix, strand, and coil, respectively. Each matrix contains preference values for amino acid exchanges associated with its structural state as observed in the 3D-ALI database. They are used to predict the secondary structure of a query alignment through listing the unique observed residue exchanges for each alignment position and adding the corresponding preference values over each of the three exchange matrices. The fact that each exchange type is counted only once for each alignment position provides implicit weighting of the sequences, thus avoiding predominance of related sequences. The secondary structure associated with the matrix showing the highest sum is then assigned to the alignment position. Following these raw predictions, three simple cleaning rules are applied and completed in three successive cycles:

(a) **Single position interruptions**: if a sequence site is predicted in one structural state and the two flanking positions in another, the position is changes into that of the consistent flanking sites, for example (H[E/C]H) becomes (HHH) where [E/C] indicates E or C.

(b) **Double position interruptions**: if in five consecutive positions two middle sites are of another type than the three flanking sites, the middle positions are changed to the flanking types. For instance, (HH[E/C][E/C]H) or (H[E/C][E/C]HH) becomes (HHHHH).

(c) **Short fragments**: helices predicted less than or equal to 4 and strands less than or equal to 2 in length are changed into coil predictions.

The accuracy of the method was assessed over one-at-a-time jackknife testing and amounted to 72%, albeit over a relatively small test set of 38 protein families.

3.5 A consensus approach: JPRED

The JPRED server at the EMBL-European Bioinformatics Institute (Hinxton, UK) conveniently runs state-of-the-art prediction methods such as PHD (9), PREDATOR (38, 39), DSC (40), and NNSSP (37), while also ZPRED (43) and MULPRED (Barton, unpublished) are included. The NNSSP method has to be activated explicitly, as it is the slowest of the ensemble. The server accepts a multiple alignment and predicts the secondary structure of the sequence on top of the alignment: Alignment positions showing a gap for the top sequence are deleted. A single sequence can also be given to the server. In the latter case, a BLAST-search is performed to


CLUSTALX:
3chy ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMPNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM
FLAV_MEGEL MVEIVYWSGTGNTEAMANEIEAAVKAAGVESVRFEDTNVDDVALLGCPAMGSEELEDSVVEPFFTDLAPKLKGKGSYGWGSGEWMDAWKQRTEDTGATVIGTAIVNEMPDNAPECKE.LGEAAAKA..
4fxn .MKIVYWSGTGNTEKMAELIAKGIIESGVNTINVSDVNIDELLILGCSAMGDEVLEESEFEPFIEEISTKISGKGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVNEPDEAEQDCIEFGKKIANI..
FLAV_DESGI KALIVYGSTTGNTEGVAEAIAKTLNSEGTTVVNVADVTAPGLALLGCSTWGDDELQEDFVPLYEDLDRAGLKDKGCGDSSYTYAVDVIEKKAEELGATLVASSLKIGEPDSAE..VLDWAREVLARV.
FLAV_DESSA KSLIVYGSTTGNTETAAEYVAEAFENKEVELKNVTDVSVADLGLFGCSTWGEEELQDDFIPLYDSLENADLKGKGCGDSDYTYAVDAIEEKLEKMGAVVIGDSLKIGDPERDE..IVSWGSGIADKI.
FLAV_DESDE KVLIVFGSSTGNTESIAQKLEELIAAGGVTLLNAADASAENLALFGCSAWGMEEMQDDFLSLFEEFNRFGLAGRASGDQEYEHAVPAIEERAKELGATIIAEGLKMGDASNDPEAVASFAEDVLKQL.
1fx1 KALIVYGSTTGNTEYTAETIARQLANAGVDSRDAASVEAGGLFLLGCSTWGDDELQDDFIPLFDSLEETGAQGRGCGDSSYEYAVDAIEEKLKNLGAEIVQDGLRIGDPRAARDDIVGWAHDVRGAI.
FLAV_DESVH KALIVYGSTTGNTEYTAETIARELADAGVDSRDAASVEAGGLFLLGCSTWGDDELQDDFIPLFDSLEETGAQGRGCGDSSYEYAVDAIEEKLKNLGAEIVQDGLRIGDPRAARDDIVGWAHDVRGAI.
FLAV_CLOAB KISILYSSKTGKTERVAKLIEEGVKRSGVKTMNLDAVDKKFLQIFGTPTYYAN.ISWEMKKWIDESSEFNLEGKSTANSIAGGALLTILNHLMVKGMLVYSGCGVHIEIQENEDENARIFGERIANKVK
FLAV_ANASP KIGLFYGTQTGKTESVAEIIRDEFGNDV..LHDVSQAEVTDLNIIGCPTWNIGQ....WEGLYSELDDVDFNGKGTGDQIGYAAIGILEEKISQRGGKTVGYWGKFGLALDEDNQSDLTDDRIKSWVA
FLAV_AZOVI KIGLFFGSNTGKTRKVAKSIKKRFDDET..ALNVNRVSAEDFAILGTPTLGEGPENESWEEFLPKIEGLDFSGKGLGDQVGYPALGELYSFFKDRGAKIVGSWGKFGLALDLDNQSGKTDERVAAWLA
2fcr KIGIFFSTSTGNTTEVADFIGKTLGAKA..IDVDDVTDPQALKFLGAPTWNTGT...SWDEFLYKLPEVDMKDLGLGDAEGYPAIEEIHDCFAKQGAKPVGFSGKFGLPLDMVNDQIPMEKRVAGWVE
FLAV_ENTAG TIGIFFGSDTGQTRKVAKLIHQKLDGIA..LDVRRATREQFLSLLGTPTLGDGPQYDSWQEFTNTLSEADLTGKGLGDQLNYSAMRILYDLVIARGACVVGNWNEFGLPLDQENQYDLTEERIDSWLE
FLAV_ECOLI ITGIFFGSDTGNTENIAKMIQKQLGKDV..VHDIAKSSKEDLELLGIPTWYYGQ....WDDFFPTLEEIDFNGKGCGDQEDVAALGTIRDIIEPRGATIVGHWDHFGLAIDEDRQPELTAERVEKWVK
dsc --EEEE------HHHHHHHHHHHHHH---HHHHHHHHHHHHH----------HHHHHHHHHH----------HHHHHHHHHHHHHHHHHHHHHHHH----------HHHHHH---
mul EEEEEE---------HHHHHHHHHHH----HH-------HHHEE--------------HHHHHHHHH----------HHHHHHHHHH--HHEE--------------HHHHHHHHHH--
phd -EEEEEE-----HHHHHHHHHHH----------HHHHH--EEEE-E------HHHHHHH----HHHH------HHH--HHHHHHHHHH--EEEE-------------HHHHHHHHHH-
pred ----EEEEEE----HHHHHHHHHH----------HHHHH------EEEE---------HHHHHHH------HHHHHHHHHHHHHHHHHH-----EEEE--HHHHHHHHHHHHH----
zpred EEEEEEE-----HHHHHHHHHHH-----HHHHHHHHHHHHEEEE----------HHHHHHHHHH----------HHHHHHHHHHHHHHEEEEE--------HHHHHHHHHHHHHHHH-
PRALINE:
3chy ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMPNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM
1fx1 ..PKALIVYGSTTGNTEYTAETIARQLANAGYEDSRDAASVEAGGLFEGFDLVLSTWGDDS.DFIPLFDSLEETGAQGRKVACFCCGDS.SEYFCGAKNLGAEGLRIDRAARDDIVGWAHDVRGAI..
FLAV_DESDE .MSKVLIVFGSSTGNTESIAQKLEELIAAGGHETLLNAADASAENLADGYDAVLSAWGMED.DFLSLFEEFNRFGLAGRKVAAFASGDQ.EEHFCGAKELGATGLKMESNDPEAVASFAEDVLKQL..
FLAV_DESVH .MPKALIVYGSTTGNTEYTAETIARELADAGYEDSRDAASVEAGGLFEGFDLVLSTWGDDS.DFIPLFDSLEETGAQGRKVACFGCGDS.SEYFCGAKNLGAEGLRIDRAARDDIVGWAHDVRGAI..
FLAV_DESGI .MPKALIVYGSTTGNTEGVAEAIAKTLNSEGMETVVNVADVTAPGLAEGYDVVLSTWGDDE.DFVPLYEDLDRAGLKDKKVGVFGCGDS.STYFCGAEELGATSLKIDD..SAEVLDWAREVLARV..
FLAV_DESSA .MSKSLIVYGSTTGNTETAAEYVAEAFENKEIDSLKNVTDVSVADLGNGYDIVLSTWGEEE.DFIPLYDSLENADLKGKKVSVFGCGDS.DTYFCGAEKMGAVSLKIDE..RDEIVSWGSGIADKI..
4fxn ..MK..IVYWSGTGNTEKMAELIAKGIIESGKDNTINVSDVNIDELLNE.DILISAMGDEV.EFEPFIEEIS.TKISGKKVALFG.....SGWGDGKNGYGCVPLIVQDEAEQDCIEFGKKIANI...
FLAV_MEGEL .MVE..IVYWSGTGNTEAMANEIEAAVKAAGADESVRFEDTNVDDVASK.DVILPAMGSEE.VVEPFFTDLA.PKLKGKKVGLFG.....SGWGSGEEDTGAT.AIVNDNA.PECKELGEAAAKA...
2fcr ...KIGIFFSTSTGNTTEVADFIGKTLGAK.ADPIDVD.DVTDPQALKDYDLLFPTWNTG.WD.EFLYDKLPEVDMKDLPVAIFGLGDAEGDNFCDAAKQGAKGLPLDIPMEKRVAGWVEAVVSETGV
FLAV_ANASP .SKKIGLFYGTQTGKTESVAEIIRDEFGND.VVLHDVS.QAE.VTDLNDYQYLIPTWNIGEWE.GLY.SELDDVDFNGKLVAYFGTGDQIGDNFQDASQRGGKGLALDDLTDDRIKSWVAQLKSEFGL
FLAV_ECOLI ..AITGIFFGSDTGNTENIAKMIQKQLGKD.VAVHDIA.KSS.KEDLEAYDILLPTWYYGEWD.DFF.PTLEEIDFNGKLVALFGCGDQEDEYFCDAEPRGATCLAIDELTAERVEKWVKQISEELHL
FLAV_AZOVI ..AKIGLFFGSNTGKTRKVAKSIKKRFDDETMSALNVN.RVS.AEDFAQYQFLIPTLGEGEWE.EFL.PKIEGLDFSGKTVALFGLGDQVGENYLDAKDRGAKGLALDGKTDERVAAWLAQIAPEPGL
FLAV_ENTAG .MATIGIFFGSDTGQTRKVAKLIHQKLDGI.ADPLDVR.RAT.REQFLSYPVLLPTLGDGEWQ.EFT.NTLSEADLTGKTVALFGLGDQLNKNPVSAIARGACGLPLDDLTEERIDSWLEKLKPAV.L
FLAV_CLOAB ..MKISILYSSKTGKTERVAKLIEEGVKRSGNIVKTMNLDAVDKKFLQESEGIIPTYYAN.WEMKKWIDESSEFNLEGKLGAAPSTANSIASDI..AMVKGMLEIQENRIFGERIANKVKQIF.....
dsc ---EEEEEE-------HHHHHHHHHHHH---HHHH------HHHH---EEEEE---------HH----HH--------EEE----------HHHHH---HHHH---HHHHHHHHHHHH---
mul ---EEEEEE---------HHHHHHHH------HH--------H-HH----EEE----------HHH-HHHHHH-----HEEEE-----------H-------HHH---HHHHHHHHHHH-H---
phd ---EEEEEEE-----HHHHHHHHHHH-------------HHHH---EE--EE-------HHHHHHHHH------EEEEEE------HHH----------E------HHHHHHHHHHHHH--
pred ----EEEEEE----HHHHHHHHHH-------------HHHHH------EEEE---------HHHHHHH-------HHHHHHHHHHHHHHHHH-----EEEE--HHHHHHHHHHHHH----
zpred ----EEEEEE------HHHHHHHHHHH---HHHHHHHHHHHHHHHH--EEEEE--------HHHHHHHHHHHHHHHHEEEEE---------HHHHHHHH-EEEHHHHHHHHHHHHHHH-----
JPRED CONSENSUS PREDICTION:
3chy ADKELKFLVVDDFSTMRRIVRNLLKELGFNNVEEAEDGVDALNKLQAGGYGFVISDWNMPNMDGLELLKTIRADGAMSALPVLMVTAEAKKENIIAAAQAGASGYVVKPFTAATLEEKLNKIFEKLGM
cons CLUSTALX -EEEEEE-----HHHHHHHHHHHH----------HHHHHHHHEEEE-E------HHHHHHHH--HHHHHH------HHHHHHHHHHHHHHHHHHHHEE----------HHHHHHHHHH-
53.9%
cons PRALINE ---EEEEEEE-----HHHHHHHHHHHH-------------HHHHH---EEEEEE------HHHHHHHHH------EEEEEE------HHH--HHH-----E------HHHHHHHHHHHHH--
85.1%
cons HOMOLOGS -----EEEEE----HHHHHHHHHHH----EEEHH---HHHHHHHH----EEEEE-------HHHHHHHHH------EEEEEE----HHHHHHHHH-----E-----HHHHHHHHHHHHH--
88.3%
DSSP EEEE HHHHHHHHHHHHH EEEE HHHHHHHHH EEEEE HHHHHHHHH EEEEE HHHHHHHHH EEEE HHHHHHHHHHHHH

Figure 5 Secondary structure prediction for chemotaxis protein cheY (3chy). The top alignment block represents the multiple alignment of the 3chy sequence with 13 distant flavodoxin sequences by the method PRALINE. The middle block is the same sequence set aligned by CLUSTALX. Under each of the alignments are given the alignments by five secondary structure prediction methods. The bottom block depicts consensus secondary structures determined by Jpred over the five methods used, respectively for the PRALINE and CLUSTALX alignments, as well as for a set of 32 homologous sequences aligned by CLUSTALX (cons HOMOLOGS). Vertical bars (|') under each of the consensus predictions indicate correct predictions. The bottom line identifies the standard of truth as obtained from the 3chy tertiary structure by the DSSP program. (10) The secondary structure states assigned by DSSP other than 'H' and 'E' were set to ' '(coil) for clarity.

find homologous sequences, which are subsequently multiply aligned using CLUSTALX and then processed with the user-provided single sequence on top in the alignment. If sufficient methods predict an identical secondary structure for a given alignment position, that structure is taken as the consensus prediction for the position. In case no sufficient agreement would be reached, the PHD prediction is taken. This consensus prediction is somewhat less accurate when the NNSSP method is not invoked or completed in the computer time slot allocated to the user. An example of output by the JPRED server for the signal transduction protein cheY (PDB code 3chy) is given in *Figure 5* (*vide infra*).

3.6 Multiple-alignment quality and secondary-structure prediction

Multiple-alignment protocols use heuristics to overcome the combinatorial explosion that arises when all possible alignments would be tested exhaustively. Most global alignment methods therefore establish an order in which the sequences are aligned progressively based on the alignment scores of all possible pairwise alignments (the number is $N \times (N - 1)/2$ with N the number of sequences). Although most methods show a comparable overall quality in alignment construction for sequences showing residue identities of 30% or higher, significant differences can arise in individual cases, particularly when evolutionary distant sequences are included. As the currently most successful secondary structure prediction methods all employ positional information from multiple alignments, it is clear that alignment quality is crucial for accurate prediction. As an example, the popular multiple alignment program CLUSTALX and the recently developed method PRALINE (53) (see below) were used to automatically construct a multiple alignment for the signal transduction protein cheY (PDB code 3chy) and 13 distant flavodoxin sequences. The 3chy structure adopts a flavodoxin fold (see *Figure 2*) despite very low sequence similarities with genuine flavodoxins. *Figure 5* shows both alignments with secondary structure predictions by the JPred server as well as JPred consensus predictions for the two alignments. The difference in accuracy of the two consensus predictions amounts to more than 30%, an order of magnitude more than the increase in prediction accuracy obtained over the last five years. It must be stressed that the flavodoxin sequences are evolutionary distant from the cheY sequence, such that the alignments were

only included to illustrate their crucial importance for secondary structure prediction rather than to argue in favour of any of the two used alignment programs based on this single example. The Jpred server was also given the single 3chy sequence, after which the Jpred server constructed an evolutionary related set of homologs through a BLAST search and aligned the 32 resulting sequences using CLUSTALX. The accuracy of the consensus secondary structure prediction by Jpred was 3% higher than that obtained for the PRALINE alignment of the cheY-flavodoxin set (*Figure 5*). Moreover, it successfully delineated the second β-strand of the 3chy structure, which was missed by the predictions based on both the CLUSTALX and PRALINE alignments.

3.7 Iterated multiple-alignment and secondary structure prediction

As mentioned, most reliable secondary structure prediction methods utilize sequence information in multiple alignments and their prediction accuracy is crucially dependent on the quality of a multiple alignment used. If in turn a multiple alignment would be guided by the predicted secondary structure, an iterative scheme would be possible that optimizes both the multiple-alignment quality and secondary structure prediction. This procedure is implemented in the PRALINE multiple alignment method (53). A multiple alignment is constructed initially without information about the secondary structure (*Figure 6a*). Then, the secondary structure is predicted (for which any of the aforementioned methods could be used) and iteratively a new alignment is constructed, now using the predicted secondary structure. PRALINE employs dynamic programming to progressively construct a multiple alignment for a query set of sequences and therefore relies on an amino acid exchange weights matrix and a pair of gap penalties (for a review, see ref. 54). The initial alignment is constructed using a default residue exchange matrix (e.g. the BLOSUM62 matrix) and gap penalties. After secondary structure prediction, resulting in a tentative secondary structure for each sequence if a single sequence-based method is used or in a single secondary structure if a method reliant on a multiple alignment is effected (*Figure 6a*), PRALINE utilizes the thus obtained secondary structure information as illustrated in *Figure 6b*. At each alignment step during the progressive alignment, pairs of sequences (and/or profiles representing already aligned sequence blocks) are matched using three secondary structure-specific residue exchange matrices (55) and associated gap penalties. As shown in *Figure 6b*, the residue exchange weights for matched sequence positions with identical secondary structure states is taken from the corresponding residue exchange matrix. Sequence positions with inconsistent secondary structure states are treated with the default exchange matrix. The secondary structure information is thus used in a conservative manner based upon the assumption that consistent secondary structure predictions are indicative for their reliability when performed for each individual sequence (*Figure 6a*). In this way, the multiple alignments guide the secondary structure predictions, which in turn guide the alignment.

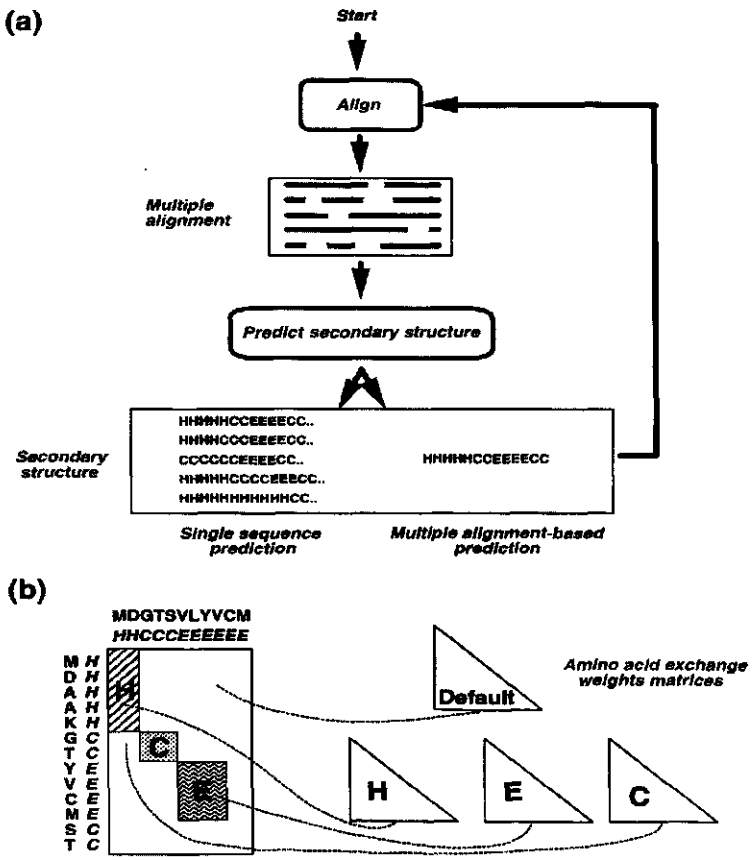

Figure 6 Iterative multiple-alignment and secondary-structure prediction by the PRALINE method.

4 Prediction of transmembrane segments

Membrane proteins (MP) form a distinct topological class due to the presence of one or more transmembrane (TM) sequence segments. In contrast to globular proteins where all possible mutual orientations of individual structural elements are in principle possible, MP transmembrane segments are subjected to severe restrictions imposed by the lipid bilayer of the cell membrane.

There is a considerable lag in structures available for MPs relative to the large and vastly growing numbers of soluble proteins, as little X-ray or NMR data regarding the tertiary structure of MPs have been available until recently (56). The most frequently observed secondary structure in transmembrane segments is the α-helix, but also transmembrane structures based on β-strands that constitute a β-barrel have been encountered. The initial idea that TM segments are either completely α-helical or consist of β-strands exclusively, was disrupted by electron microscopy data for the nicotinic acetylcholine receptor (57), which was interpreted as a central five-helix bundle surrounded by β-strands, albeit based on preliminary data with low resolution.

Fortunately, the location of the transmembrane segments in primary structure of the MP is relatively easy to predict due to the rather strong tendency of certain hydrophobic amino acid types with their special physico-chemical properties to occur in membrane spanning regions. Thus, efforts concerning the theoretical analysis of MPs over the past two decades have been focused on the determination of the membrane sequence segment boundaries and their tentative orientation with respect to the membrane, although mostly assuming α-helical structures.

4.1 Prediction of α-helical TM segments

The following considerations form the basis of transmembrane subsequence prediction.

(a) Amino acids immersed in the lipid phase are likely to be hydrophobic. Therefore, any physical measure of amino acid hydrophobicity derived from physical calculations and/or experimental data can serve as a measure of likeliness for a residue type to occur in a membrane-spanning segment.

(b) The propensities of amino acids to reside in the lipid bilayer can be inferred from abundant but not very precise experimental data on the boundaries of the transmembrane segments acquired from site-directed mutagenesis, enzymatic cleavage, immunological methods and the like. This contrasts with the standard secondary structure prediction methods for soluble proteins where statistical propensities of different amino acids to form one of the major secondary structure elements are derived from more accurate protein tertiary structural data from X-ray crystallography and NMR spectroscopy.

(c) Transmembrane segments are believed to adopt in most cases α-helical conformation. An α-helix is the most suitable local arrangement because, in the absence of water molecules inside the membrane, all main-chain polypeptide donors and acceptors must mutually satisfy each other through formation of hydrogen bonds as occur in an α-helix. This energetic argument is supported by experimental evidence where polypeptide chain tends to adopt helical conformation in a non-polar medium (58). Therefore, α-helical propensities of amino acids derived from the analysis of globular proteins can be considered in MP structure prediction.

Although the globular protein interior is less apolar than the lipid bilayer, extensive usage of these data has been made for MP structure prediction, particularly with the classical hydrophobicity scale of Kyte and Doolittle (59). Other techniques are more specifically aimed at searching MP transmembrane regions (60–63). Hydrophobic scales can be used to build a smoothed curve, often called a hydropathic profile, by averaging over a sliding window of given length along the protein sequence to predict transmembrane regions. Stretches of hydrophobic amino acids likely to reside in the lipid bilayer appear as peaks with lengths corresponding to that expected for a transmembrane segment, typically 16–25 residues. The choice of window length should correspond to the expected

length of a transmembrane segment. Given that the average membrane thickness is about 30 Å, approximately 20 residues form a helix reaching from one lipid bilayer surface to another. A further threshold is also required to determine the exact boundaries of a membrane spanning segment. Kyte and Doolitle (59) early on set their limit by examining the hydropathic character of just a few membrane proteins. Later, a much larger learning set was used by Klein *et al.* (64) through discriminant analysis. Rao and Argos (65) suggested a minimum value for the peak hydrophobicity and a cut-off value at either end of the peak to terminate the helix.

Although many of the above techniques constitute an essential part of all major sequence analysis packages, the relatively simple physical considerations forming the basis of these methods do not exhaust the whole variety of possible situations.

If a membrane protein has more than one transmembrane helix, the relative orientation of the helices and the interaction of the corresponding sidechains are also important for structure prediction. The structure of the membrane proteins determined to date and also some theoretical evidence (66) support the view that α-helices in membranes form compact clusters. While the residues facing the lipid environment conform to the preferences described above, the interface residues between different helices do not necessarily have contact with the membrane, and, can therefore, behave differently. It is possible that charged residues occur in the helices in a coordinated fashion such that positively charged sidegroups on one helix will have their negatively charged counterparts on another helix. These charges could, for instance, constitute a membrane channel. Intra-membrane α-helices can thus have an amphipathic character (see above). In such cases hydropathic profiles can work poorly in detecting transmembrane segments. In certain cases where the number of transmembrane segments is large (more than 20 as in some channel proteins), the inner helices of the transmembrane helical bundle can completely avoid contact with the lipid bilayer and, therefore, any restrictions on their amino acid content—or even length—could be artificial.

Eisenberg *et al.* (67) introduced a quantitative measure of helix amphipathicity called the 'hydrophobic moment', and defined as a vector sum of the individual amino acid hydrophobicities radially directed from the helix axis. The hydrophobic moment provides sufficient sensitivity to distinguish amphipathic α-helices of globular, surface, and membrane proteins. Many methods for amphipathic analysis were developed based on Fourier analysis of the residue hydrophobicities (68, 69) and the average hydrophobicity on one helix face (70).

Several prediction methods have emerged which utilize multiple factors, complex decision rules, and large learning sets. Von Heijne (63) proposed a synthetic technique in which a standard hydrophobicity analysis is supplemented by charge bias analysis (see Section 4.2). Other methods include the joint usage of several selected hydrophobicity scales (71) or the application of optimization techniques with membrane segments as defined by X-ray analysis serving as reference examples (72).

Persson and Argos (73) incorporated sequence information from multiple alignments to aid TM prediction. The propensities of amino acids to reside in either the central or the flanking regions of a transmembrane segment were calculated using more then 7500 individual helical assignments contained in the SWISS-PROT sequence databank. These values were then used to build a prediction algorithm wherein, for each segment of a multiple sequence alignment, and for each sequence, average values of the central and flanking propensities are calculated over windows of respectively 15 and 4 residues long. If the peak values for central transmembrane regions exceed a certain threshold, this region is considered as a possible candidate to be membrane spanning. The algorithm then attempts to expand this region in both sequence directions until a flanking peak is reached or the central propensity averages fall below a certain value. Additional restraints are imposed on the possible length of the tentative transmembrane segments. The optimal window length was found to be about 15 residues. Due to the increased amount of information utilized by the technique, more accurate prediction results were achieved as compared with earlier methods. The gain in sensitivity is due to the usage of multiple alignments as well as the introduction of a second propensity for flanking regions.

Neural networks (see Section 3.4.1) have also been applied to the TM prediction problem. Early attempts involved training on secondary structural elements of globular proteins (74). Rost *et al.* (75) trained the PHD method on multiple alignments for 69 protein families with known transmembrane helices and achieved 95% prediction accuracy using the jackknife test.

4.2 Orientation of transmembrane helices

Another aspect of transmembrane segment prediction is prediction of membrane sidedness, or orientation. For bacterial membrane proteins it was found that intracellular loops in between transmembrane helices contain arginine and lysine residues much more frequently than the extracellular exposed loops (76, 77). This pattern has been shown to apply also to eukaryotic membrane proteins, but to a lesser extent (78). An additional observation, made for eukaryotic proteins, is that the difference in the total charge of the approximately 15 residues flanking the transmembrane region on both sides of the membrane also coincides with the orientation of the protein (79). If the C-terminal portion of the protein adjacent to this segment is more positive in charge then the N-terminal portion, the C-terminus will reside in the cytosol, and *vice versa*. Non-random charge distribution may also play an important role in membrane insertion of the protein. These findings, collectively known as the 'positive inside rule', aid prediction schemes for MP topology. However, the positive inside rule is only applicable to α-helical TM regions.

4.3 Prediction of β-strand transmembrane regions

As the methods described above all predict TM segments assuming the α-helical conformation, transmembrane segments constituted by β-strands are not likely

to be predicted successfully. Four different families of β-barrel membrane proteins are known to date (porins, OmpA, FhuA, and FepA). For example, porins form voltage-dependent membrane channels and have a β-barrel fold constituted by 16 β-strands (80). Hydrogen bonds are formed only between adjacent β-strands. Most of the outer surface of the barrel faces the lipid environment whereas the internal part serves as an aqueous pore. Each individual β-strand could therefore be expected to be amphipathic with a period of two residues. However, while every second residue facing the lipid bilayer is hydrophobic, those side chains protruding towards the interior of the barrel display no definitive tendency, thus lowering the amphipathic signal. Another complication is the fact that the number of amino acid residues in extended conformation needed to span the membrane is much smaller than that for the helical conformation, typically about 10. Consequently, smoothed hydropathic profiles are likely to miss such short stretches.

5 Coiled-coil structures

If a protein is predicted to contain α-helices, higher-order information as well as increased confidence in predictions made could be gained from testing the possibility that a pair of helices adopts a superhelical twist resulting in a coiled-coil conformation. The left-handed coiled-coil interaction involves a repeated motif of seven helical residues (*abcdefg*). The *a* and *d* positions are normally occupied by non-polar residues constituting the hydrophobic core of the helix–helix interface, whereas the other positions display a high likelihood to comprise hydrophilic residues. The *e* and *g* positions in addition are often charged and can form salt-bridges to each other. The program COILS2 (81, 82) exploits this information and compares a query sequence with a database of known parallel two-stranded coiled-coils. A similarity score is derived and compared to two score distributions, one for globular proteins (without coiled-coils) and one for known coiled-coil structures, and a probability is then calculated for the query sequence to adopt a coiled-coil conformation. As the program assumes the presence of heptad repeats, the probabilities are derived using windows of 14, 21, and 28 amino acids. However, the program offers the option to include user-defined window lengths two allow the handling of cases with extreme coiled-coil lengths. A recently updated scoring matrix which includes new structures with known coiled-coils and contains amino acid type propensities at the various positions in the heptad repeats, led to increased recognition of coiled-coils elements. The COILS2 method accurately recognises left-handed two-stranded coiled coils but loses sensitivity for coiled-coil structures composed of more than two strands. It is not able to recognize right-handed or buried coiled–coil helices and therefore is not applicable to transmembrane coiled-coil structures known to basically show the similar coiled-coil conformations as soluble proteins, albeit with dramatically different and more hydrophobic constituent amino acids (56).

6 Threading

If a homologous protein with known structure is available for a query sequence, this structure can then be aligned to the query sequence using the threading technique (83). Treading methods test the feasibility for a given sequences to adopt a particular fold, based on assessing the likelihood for the amino acids in the query sequence to occur in the local residue environments within the known tertiary structure. The optimal fit of the query sequence through the tertiary structure effectively leads to an alignment, which can be used to copy the secondary structure of the known fold to the query sequence. Although the incorporation of tertiary structure information should lead to better alignment and recognition of related sequences, the increased sensitivity of available threading methods as compared with conventional sequence alignments is not always clear. Jones *et al.* (84) discuss various threading methods available and also how their results should be interpreted.

7 Recommendations and conclusions

Table 1 lists WWW addresses of some of the available prediction methods discussed in this chapter and in *Protocol 2* some recommendations are given to maximize the chances of an accurate prediction of the secondary structure associated with a protein query sequence. In cases where a multiple alignment is used, it is generally important to test the consistency and quality of the alignment constructed, as this can have dramatic consequences for the prediction accuracy of multiple-alignment-based methods. In testing the consistency of the currently most accurate prediction methods and determining a consensus prediction, the positional reliability indices offered by some of prediction methods should be included. Furthermore, the general accuracies for predicting each of the three secondary structural states that are published for a number of the methods can be used to weight the contribution of their positional predictions

Table 1 Websites of various secondary structure prediction methods and related services

Service	Reference	URL
GOR4	27, 28, 29	http://absalpha.dcrt.nih.gov:8008/gor.html
PHD[a]	9	http://dodo.cpmc.columbia.edu/predictprotein/[b]
Pred2ary	35	http://yuri.harvard.edu/~jmc/2ary.html
NNSSP[a]	37	http://dot.imgen.bcm.tmc.edu:9331/pssprediction/pssp.html
PREDATOR[a]	38, 39	http://www.embl-heidelberg.de/cgi/predator_serv.pl
DSC[a]	40	http://bonsai.lif.icnet.uk/bmm/dsc/dsc_read_align.html
Zpred[a]	43	http://kestrel.ludwig.ucl.ac.uk/zpred.html
Jpred	-	http://barton.ebi. ac.uk/servers/jpred.html
COILS2	81, 82	http://www.isrec.isb-sib.ch/coils/COILS_doc.html

[a] Method can also be run using the Jpred server.

[b] Mirror websites for PHD can be found here as well.

in a consensus. Specifically, the PREDATOR method should be included in the trials as it is the only method relying on multiple local rather than global alignment of the query sequences. It is important to realize that there is no single best prediction method so that the degree of consistency over a variety of methods is crucial for getting an idea about the prediction accuracy. Attempts to recognize higher-order structure, such as the fold the protein might adopt or the likelihood for coiled-coil structures, could enhance the confidence in predictions made or help correcting possible mispredictions. Easily recognizable errors might be disruptions in alternating α-helix/β-strand predictions in a likely α/β protein fold or the occurrence of a single β-strand within a tentative α-helical protein. In general, it is likely that the accuracy of computerized prediction methods can be enhanced further if such reasoning with higher order structure is formalized and incorporated in the prediction mechanisms. Some easy benefits will come from the steadily increasing structural protein data that can be used to better train and tune the statistical methods. The current availability of the prediction methods optimizes the chance for development of sensitive consensus methods. It is clear from the ongoing interest and activity in both the application and development of secondary structure prediction methods that the end of the three decades of research efforts is not in sight.

Protocol 2

Predicting secondary structure

1. Get a balanced and non-redundant set of homologous sequences for a given protein query sequence.
2. Try a number of multiple alignment routines to obtain a consistent multiple alignment.
3. Check the alignment carefully by eye using any additional information (e.g. active site residues, disulfide bridges, etc.).
4. Use as many good secondary structure prediction methods as possible and construct a consensus prediction (a convenient aid is the Jpred server).
5. Try to recognize super-secondary or higher-order structural features from the predicted secondary structure elements and try to interpret and correct prediction results (e.g. the missed second β-strand in the flavodoxin example) (see Section 3.6).

References

1. Heringa, J. and Taylor, W. R. (1997). *Curr. Opin. Struct. Biol.*, **7**, 416.
2. Springer, T. A. (1997). *Proc. Natl. Acad. Sci. USA*, **94**, 65.
3. Baldwin R. L. and Roder H. (1991). *Curr. Biol.*, **1**, 218.
4. Goldenberg, D. P., Frieden, R. W., Haack, J. A., and Morrison, T. B. (1989). *Nature*, **338**, 127.
5. Baldwin, R. L. (1990). *Nature*, **346**, 409.

6. Flores, T. P., Moss, D. S., and Thornton, J. M. (1994). *Protein Eng.*, **7**, 31.
7. Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., *et al.* (1977). *J. Mol. Biol.*, **112**, 535.
8. Schiffer, M. and Edmundson, A. B. (1967). *Biophys. J.*, **7**, 121.
9. Rost, B. and Sander, C. (1993). *J. Mol. Biol.*, **232**, 584.
10. Kabsch, W. and Sander, C. (1983). *Biopolymers*, **22**, 2577.
11. Colloc'h, N., Etchebest, C., Thoreau, E., Henrissat, B., and Mornon, J.-P. (1993). *Protein Eng.*, **6**, 377.
12. Sklenar, H., Etchebest, C., and Lavery, R. (1989). *Proteins: Struct. Funct. Genet.*, **6**, 46.
13. Woodcock, S., Mornon, J.-P., and Henrissat, B. (1992). *Protein Eng.*, **5**, 629.
14. Russell, R. B. and Barton, G. J. (1993). *J. Mol. Biol.*, **234**, 951.
15. Taylor, W. R. (1984). *J. Mol. Biol.*, **173**, 512.
16. Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J. (1986). *Biochemistry*, **25**, 266.
17. Cohen, F. E. and Kuntz, I. D. (1989). In *Prediction of protein structure and the principles of protein conformation* (ed. G. D. Fasman), pp. 647–706. Plenum, New York, London.
18. Sternberg, M. J. E. (1992). *Curr. Opin. Struct. Biol.*, **2**, 237.
19. Benner, S. A., Cohen, M. A., and Gerloff, D. (1993). *J. Mol. Biol.*, **229**, 295.
20. Rost, B., Sander, C., and Schneider, R. (1994). *J. Mol. Biol.*, **235**, 13.
21. Frishman, D. and Argos, P. (1995). *Proteins: Struct. Funct. Genet.*, **25**, 633.
22. Szent-Györgyi, A. G. and Cohen, C. (1957). *Science*, **126**, 697.
23. Periti, P. F., Quagliarotti, G., and Liquori, A. M. (1967). *J. Mol. Biol.*, **24**, 313.
24. Nagano, K. (1973). *J. Mol. Biol.*, **75**, 401.
25. Chou, P. Y. and Fasman, G. D. (1974). *Biochemistry*, **13**, 211.
26. Lim, V. I. (1974). *J. Mol. Biol.*, **88**, 857.
27. Garnier, J., Osguthorpe, D. J., and Robson, B. (1978). *J. Mol. Biol.*, **120**, 97.
28. Gibrat, J.-F., Garnier, J., and Robson, B. (1987). *J. Mol. Biol.*, **198**, 425.
29. Garnier, J. G., Gibrat, J.-F., and Robson, B. (1996). In *Methods in enzymology* (ed. R. F. Doolittle), Vol. 266, pp. 540-53. Academic Press.
30. Schultz, G. A. (1988). *Annu. Rev. Biophys. Chem.*, **17**, 1.
31. Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J. (1983). *Biochemistry*, **25**, 4894.
32. Taylor, W. R. and Thornton, J. M. (1983). *Nature*, **354**, 105.
33. Rooman, M. J., Wodak, S., and Thornton, J. M. (1989). *Protein Eng.*, **3**, 23.
34. Presnell, S. R., Cohen, B. I., and Cohen, F. E. (1992). *Biochemistry*, **31**, 983.
35. Chandonia, J.-M. and Karplus, M. (1998). *Proteins: Struct. Funct. Genet.*, **35**, 293.
36. Yi, T.-M. and Lander, E. S. (1993). *J. Mol. Biol.*, **232**, 1117.
37. Salamov, A. A. and Solovyev, V. V. (1995). *J. Mol. Biol.*, **247**, 11.
38. Frishman, D. and Argos, P. (1996). *Protein Eng.*, **9**, 133.
39. Frishman, D. and Argos, P. (1997). *Proteins: Struct. Funct. Genet.*, **27**, 329.
40. King, R. D. and Sternberg, M. J. E. (1996). *Protein Sci.*, **5**, 2298.
41. Muggleton, S., King, R., and Sternberg, M. J. E. (1992). *Protein Eng.*, **5**, 647.
42. Taylor, W. R. (1986). *J. Theor. Biol.*, **119**, 205.
43. Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E. (1987). *J. Mol. Biol.*, **195**, 957.
44. Levin, J. M., Pascarella, S., Argos, P., and Garnier, J. (1993). *Protein Eng.*, **6**, 849.
45. Minsky, M. and Papert, S. (1988). *Perceptrons*. MIT Press, Cambridge, MA.
46. Sander, C. and Schneider, R. (1991). *Proteins: Struct. Funct. Genet.*, **9**, 56.
47. Bowie, J. U., Lüthy, R., and Eisenberg, D. (1991). *Science*, **253**, 164.
48. Gonnet, G. H., Cohen, M. A., and Benner, S. A. (1992). *Science*, **256**, 1443.

49. Hutchinson, E. G. and Thornton, J. M. (1994). *Protein Sci.*, **3**, 2207.
50. Smith, T. F. and Waterman, M. S. (1981). *J. Mol. Biol.*, **147**, 195.
51. Mehta, P. K., Heringa, J., and Argos, P. (1995). *Protein Sci.*, **4**, 2517.
52. Pascarella, S. and Argos, P. (1992). *Protein Eng.*, **5**, 121.
53. Heringa, J. (1999). *Comp. Chem.*, **23**, 341.
54. Heringa, J., Frishman, D., and Argos, P. (1997). In *Proteins: a comprehensive treatise, Vol. 1, principles of protein structure* (ed. G. Allen), pp. 171–277. JAI Press, New York.
55. Lüthy, R., McLachlan, A. D., and Eisenberg, D. (1991). *Proteins: Struct. Func. Genet.*, **10**, 229.
56. Langosch, D. and Heringa, J. (1998). *Proteins: Struct. Func. Genet.*, **31**, 150.
57. Unwin, N. (1993). *J. Mol. Biol.*, **229**, 1101.
58. Singer, S. J. (1962). *Adv. Protein Chem.*, **17**, 1.
59. Kyte, J. and Doolittle, R. F. (1982). *J. Mol. Biol.*, **157**, 105.
60. Argos, P., Rao, M. J. K.,and Hargrave, P. A. (1982). *Eur. J. Biochem.*, **128**, 565.
61. Sweet, R. M. and Eisenberg, D. (1983). *J. Mol. Biol.*, **171**, 479.
62. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C. (1987). *J. Mol. Biol.*, **195**, 659.
63. Von Heijne, G. (1992). *J. Mol. Biol.*, **225**, 487.
64. Klein, P., Kanehisa, M. I., and DeLisi, C. (1985). *Biochim. Biophys. Acta*, **815**, 468.
65. Rao, M. J. K. and Argos, P. (1986). *Biochim. Biophys. Acta*, **869**, 197.
66. Wang, J. and Pullman, A. (1991). *Biochim. Biophys. Acta*, **1070**, 493.
67. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1982). *Nature*, **299**, 371.
68. Eisenberg, D., Weiss, R. M., and Terwilliger, T. C. (1984). *Proc. Natl. Acad. Sci. USA*, **81**, 140.
69. Finer-Moore, J. and Stroud, R. M. (1984). *Proc. Natl. Acad. Sci. USA*, **81**, 155.
70. Vogel, H., Wright, J. K., and Jähnig, F. (1985). *EMBO J.*, **4**, 3625.
71. Esposti, M. D., Crimi, M., and Venturoli, G. (1990). *Eur. J. Biochem.*, **190**, 207.
72. Edelman, J. (1993). *J. Mol. Biol.*, **232**, 165.
73. Persson, B. and Argos, P. (1994). *J. Mol. Biol.*, **237**, 181.
74. Fariselli, P., Compiani, M., and Casadio, R. (1993). *Eur. Biophys. J.*, **22**, 41.
75. Rost, B., Casadio, R., Fariselli, P., and Sander, C. (1995). *Protein Sci.*, **4**, 521.
76. von Heijne, G. (1986). *EMBO J.*, **5**, 3021.
77. Boyd, D. and Beckwith, J. (1989). *Proc. Natl. Acad. Sci. USA*, **86**, 9446.
78. Sipos, L. and von Heijne, G. (1993). *Eur. J. Biochem.*, **213**, 1333.
79. Hartmann, E., Rapoport, T. A., and Lodish, H. F. (1989). *Proc. Natl. Acad. Sci. USA*, **86**, 5786.
80. Schirmer, T. and Rosenbusch, J. P. (1991). *Curr. Opin. Struct. Biol.*, **1**, 539.
81. Lupas, A., van Dyke, M., and Stock, J. (1991). *Science*, **252**, 1162.
82. Lupas, A. (1996). In *Methods in enzymology* (ed. R. F. Doolitle), Vol. 266, pp. 513–25. Academic Press.
83. Jones, D. T., Taylor, W. R., and Thornton, J. M. (1992). *Nature*, **358**, 86.
84. Jones, D. T., Orengo, C. A., and Thornton, J. M. (1996). In *Protein structure prediction: a practical approach* (ed. M. J. E. Sternberg), pp. 173–206. Oxford University Press, Inc., New York.

This page intentionally left blank