Computational Methods for Protein Secondary Structure Prediction Using Multiple Sequence Alignments

Jaap Heringa*

Division of Mathematical Biology, National Institute for Medical Research (NIMR), The Ridgeway, Mill Hill, London NW7 1AA, United Kingdom

Abstract: Efforts to use computers in predicting the secondary structure of proteins based only on primary sequence information started over a quarter century ago [1-3]. Although the results were encouraging initially, the accuracy of the pioneering methods generally did not attain the level required for using predictions of secondary structures reliably in modelling the three-dimensional topology of proteins. During the last decade, however, the introduction of new computational techniques as well as the use of multiple sequence information has lead to a dramatic increase in the success rate of prediction methods, such that successful 3D modelling based on predicted secondary structure has become feasible [e.g., Ref 4]. This review is aimed at presenting an overview of the scale of the secondary structure prediction problem and associated pitfalls, as well as the history of the development of computational prediction methods. As recent successful strategies for secondary structure prediction all rely on multiple sequence information, some methods for accurate protein multiple sequence alignments will also be described. While the main focus is on prediction methods for globular proteins, also the prediction of trans-membrane segments within membrane proteins will be briefly summarised. Finally, an integrated iterative approach tying secondary structure prediction and multiple alignment will be introduced [5].

INTRODUCTION

Many protein structures show a modular organisation with structural and/or functional units comprised by independently folding substructures called domains. The connectivity of the backbone through the domains is often not important and examples exist of proteins retaining a stable structure irrespective of the sequential arrangement of the domains and presence of linker fragments [6]. Within domains at the level of protein secondary structure, however, the elements rely critically on other counterparts in their environment. It is because of this context dependency, that successful prediction of the secondary structure of proteins from sequence information alone is a difficult and longstanding task.

A secondary structure element is a consecutive fragment of a protein sequence, which corresponds to a local region in the associated protein structure and shows distinct geometrical features. The two basic secondary structure types, the $\alpha$-helix and $\beta$-strand, are regular and easily distinguishable in protein tertiary structures, while other types are sometimes harder to classify. The vast majority of prediction methods therefore are aimed at predicting into three classes, $\alpha$-helix, $\beta$-strand and otherwise, the latter often referred to as coil. Within known protein structures, about 50% of the amino acids are observed to fold into $\alpha$-helices or $\beta$-strands, so that on average roughly half of protein structure is irregular. The primary reason for the regularity observed for helices and strands is the inherent polar nature of the protein backbone, which comprises a polar nitrogen and oxygen atom for each amino acid. For a protein to become foldable with an acceptable internal energy, the parts of the main-chain buried in the internal protein core need to form hydrogen bonds between these polar main-chain atoms. The $\alpha$-helix and $\beta$-
strand conformations are optimal for this, since each main-chain nitrogen atom can associate with an oxygen partner (and vice versa) with both secondary structure types. However, in order to satisfy the hydrogen-bonding constraints, β-strands need to interact with other β-strands, which they can do in a parallel or anti-parallel fashion, to form a β-pleated sheet. β-Strands thus depend on crucial interactions between residues remote in sequence and therefore are believed to have more pronounced context dependencies than α-helices, the latter being more able to fold “on their own”. As a consequence, most prediction methods have greatest difficulty in delineating β-strands correctly.

**Protein Folding**

Experimental evidence on early protein folding intermediates has shown that secondary structural elements form at early stages during the folding process. This bodes well with the so-called "framework" model of protein folding [7,8], where
two or more secondary structural elements would associate early during folding to provide a structural framework to which subsequently other sub-structures could attach. Knowledge of protein secondary structure is important for understanding the folding process or kinetics associated with it, as it allows docking experiments to be carried out on the predicted secondary structural segments.

Fold Recognition

Another area where a notion of secondary structure is important is fold recognition. In essence, fold recognition or threading techniques test the compatibility of a given query sequence with a database of three-dimensional (3D) structures. Features of each 3D structure in the database are formalised, so that the query sequence can be compared with the structure through a sequence-to-structure alignment, resulting in a pseudo-energy score for the compatibility of the tested structure (Fig. 1). The structure with the lowest pseudo-energy then corresponds to the most likely fold for the query sequence, thereby conferring the information known for that fold. The available threading methods use a variety of tertiary structural features for scoring the threading of the sequence through the structure, often including the predicted secondary structure for the query sequence. It is also possible to recognise the

![Fig. (2). Schematic representation of the flavodoxin topology and secondary structure. Red balls represent α-helices, while green balls designate β-strands, The N-terminal strand is depicted in the yellow. The structure of the chemotaxis protein cheY (PDB code 3chy) was taken as it represents the basic flavodoxin fold.](image)
most likely three-dimensional topology of a query protein only using its predicted secondary structure [9-11]. For such fold recognition to work, it is crucial to predict correctly the helices and strands essential for the topology. Correctly delineating the edges of those structures is less critical, as is the detection of any non-essential secondary structures. An example of the topologically essential secondary structures within the flavodoxin fold is given in Fig (2). The flavodoxin fold belongs to the class of α/β-folds with the essential secondary structures distributed over the sequence as [βα]5. The five strands fold into a single β-pleated sheet ordered topologically as β2β1β3β4β5, where the numbers indicate their sequential position 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). Fig (3) shows a schematic representation of the secondary structure of a few flavodoxin structures as provided by the TOPS server (http://tops.ebi.ac.uk). In a TOPS diagram [12], an α-helix is represented by a circle and a β-strand by a triangle. The included topologies of a few different flavodoxin structures (Fig 3) illustrate that various secondary structural elements can be added onto the basic structure, provided these do not disrupt the basic flavodoxin fold. Proper prediction of the sequential order of the topologically essential helices and sheets, such as [βα]5 for flavodoxins, often allows the recognition of the fold type associated with the protein sequence considered. Further, active sites of enzymes typically comprise amino acids positioned in loops. Therefore, identically conserved residues 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.

**BIOCHEMICAL FEATURES OF SECONDARY STRUCTURES USED IN PREDICTION**

Analyses of secondary structure and related features of the many protein structures deposited in the Protein Data Bank (PDB) [13] have resulted in rules and concepts important for secondary structure prediction. The prediction methods developed over the years differ in their selection of these concepts used for prediction and the way these concepts are implemented in a computer algorithm. However, most prediction methods make use, either implicitly or explicitly, of the following observations respectively for α-helices, β-Strands, and coil structures:

- As the number of residues in a α-helical turn is 3.6 in the ideal case and helices are often positioned against a buried core, they have one phase contacting core hydrophobic amino acids, while the opposite phase interacts with the solvent. This results in so-called amphipathic helices [14], showing a periodicity of three to four residues in hydrophobicity of the associated sequence stretch. As an additional rule, proline residues tend not to occur in middle segments as they disrupt the α-helical turn. However, they are seen in the first two positions of α-helices.

- β-Strands normally constitute a so-called β-pleated sheet with two strands forming either edge. Therefore, the hydrophobic nature of such edge strands is different from that of strands internal to the β-sheet. As side-chains of constituent residues along a β-strand alternate the direction in which they protrude, edge strands of a β-sheet can show an alternating pattern of hydrophobic-hydrophilic residues, while buried strands typically comprise hydrophobic residues only. The β-strand is the most extended conformation (i.e. consecutive Cα atoms are farthest apart), so that 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 would comprise more residues. Further, 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 consist of relatively hydrophobic residues.

- Multiple alignments of protein sequences often display gapped and/or highly variable regions, which would be expected to correspond to loop regions rather than the other two basic secondary structures. Loop
regions contain a high proportion of small polar residues like alanine, glycine, serine and threonine. Glycine and proline residues are also seen in loop regions, the former due to their inherent flexibility, the latter for entropic reasons relating to the observed rigidity in their kinking the main-chain. As proline residues kink the main-chain in an incompatible way for helices and strands, they are normally not observed in these two structures, although, as stated above, they can occur in the N-terminal two positions of α-helices.

The patterns observed for each of the three secondary structural types are reflected in general
compositional differences between helix, strand and coil conformations. These compositional trends are exploited in many of the early prediction methods based on single sequence information. Most modern methods rely on multiple alignments, which means that they can exploit all of the above properties as well as the fact that the amino acid exchange patterns are different for the three secondary structural states (see below).

Once the secondary structure is predicted for a given protein sequence or multiple alignment, a few additional heuristics can shed light onto its tertiary structure and/or function:

- Hydrophobic residues, particularly when conserved, are normally buried in the protein core.

- A common structural motif is the so-called \(\beta\)-\(\alpha\)-\(\beta\) motif, which denotes a \(\beta\)-strand followed in sequence by an \(\alpha\)-helix and then another \(\beta\)-strand. More than 95% of the observed \(\beta\)-\(\alpha\)-\(\beta\) motifs show a right-handed chirality (Fig. 4). Also for the aforementioned flavodoxin fold (Fig. 2), all five \(\beta\)-\(\alpha\)-\(\beta\) motifs are right handed. This likely right-handed chirality can be used in modelling the topology if one or more of these motifs would be predicted.

- Helices often cover up a core of \(\beta\)-strands. Therefore, if both \(\alpha\)-helices and \(\beta\)-strands are predicted, it should be attempted to distribute the helices evenly at either side of (layers of) tentative \(\beta\)-sheets, in topology modelling.

- 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. Since active site residues are spatially close in a protein structure, the coil structures they constitute should be placed in proximity in a topology model, which in turn provides inside in the possible placement of the tentative helices and strands.

**PREDICTION METHODS FOR GLOBULAR PROTEINS**

Using computers in predicting protein secondary has its onset over a quarter century ago [1]. The accuracy of the computational methods devised early on is in the range 50-54% [15]. Random prediction would yield about 40% correctness given the observed distribution of the three states in globular proteins (with generally about 30% helix, 20% strand and 50% coil). Although significantly beyond the random level, the accuracy of the early single-sequence prediction methods have not been sufficient to allow the successful prediction of protein

![Fig. (4). Handedness of the \(\beta\)-\(\alpha\)-\(\beta\) motif. The usual right-handed as well as the rare left-handed motif are depicted. In both images, the sequence runs through the helix from right to left.](attachment:image.png)
topology based on docking secondary structure. During the last decade, however, the introduction of novel computational techniques as well as the use of multiple sequence information has lead to a significant increase in the success rate of prediction methods. In the remainder of this section, the most significant early and recent methods will be discussed.

THE EARLY APPROACHES – SINGLE SEQUENCE PREDICTION


Nagano

Nagano [1] explored the likelihood of residue pairs within specific secondary structures to show short-range interactions up to six residues separated in sequence \( (m \leq 6) \). The contributions of these interactions for each \( m \) value were used in a linear combination to obtain a propensity for each amino acid to reside in helix, loop or \( \beta \)-structure. Nagano performed a statistical analysis over 95 crystallographically determined protein tertiary structures and determined the frequency of each pair of interacting residues. Although few structures were available at the time, Nagano realised the importance of non-redundant database information in training his method, as the weighting factors for the \( m \) values leading to best prediction were derived from 30 representative protein structures. Also, Nagano [1] touched upon the notion of the jackknife test (see below), as predicted protein examples were included with propensities derived using a training database from which the actually predicted protein was deleted.

Chou-Fasman

The most widely used pioneering method is the one by Chou and Fasman [2], in which predictions are based on differences in residue composition for three states of secondary structure: \( \alpha \)-helix, \( \beta \)-strand and turn (i.e., neither \( \alpha \)-helix nor \( \beta \)-strand). Chou and Fasman performed a statistical analysis over a number of protein tertiary structures and determined the frequency of each amino acid type in four states. The position of turn residues was also included in the frequency calculations since significant positional differences were observed for residue types occurring at different turn positions. The frequencies were normalised 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 \( \alpha \)-helix predictions and over two for \( \beta \)-strands. Secondary structures were initiated according to the higher preference values and minimum nucleation lengths required for each structural state. An initiated secondary structure was extended as long as preferences remained high and certain residues were not encountered (e.g., proline in a \( \alpha \)-helix). The Chou-Fasman method has been assessed independently to have an accuracy of 50% [15].

Lim

Lim (3) developed a set of complicated stereochemical prediction rules for \( \alpha \)-helices and \( \beta \)-sheets based on their packing as observed in globular proteins. Apart from being the most accurate early method, Lim’s stereochemical rules are quite important for understanding protein folding. For example, Lim devised a number of rules for hydrophobic residue interactions within \( \alpha \)-helices, including terminal hydrophobic amino acid pairs at sequence positions \( i \) and \( i+1 \) of a helix, hydrophobic pairs in middle helical segments positioned at \( (i, i+4) \) and middle hydrophobic triplets positioned at \( (i, i+1, i+4) \) or \( (i, i+3, i+4) \). The Lim method never gained widespread popularity due to the fact that a computer implementation has not been developed until recently, with more accurate automatic methods now available (see below). Nonetheless, the Lim method shows the highest early accuracy of 56% [15].
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 [17] and the most recent version, GOR IV [18] 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 converted to secondary structure propensities for the central window position using an information function based on conditional probabilities. As it is not feasible to sample all possible 17-residue fragments directly from the PDB (there are $17^{20}$ possibilities), in subsequent versions of the GOR method over the years increasingly detailed approximations have been used to alleviate the sampling problem, in conjunction with the growth of data in the PDB. In GOR I and GOR II, the 17 positions in the window were treated as being independent, and so single-position information could be summed over the 17-residue window. In GOR III, this approach was refined by including pair frequencies derived from 16 pairs between each non-central and the central residue in the 17-residue window [17]. The current version, GOR IV combines pair-wise information over all possible paired positions in a window (there are $17 \times 16/2$ possibilities), albeit with a relatively small weight, compared with the single-position information of GOR I [18].

The theoretical principles used in the GOR method are statistically sound and no ad-hoc rules or artificial variables are invoked. It therefore is one of the most elegant methods with a high accuracy given its single sequence prediction. However, as in many other recent methods (see below), a post-processing step was introduced in 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). The GOR III method was evaluated independently to predict on average 53% of the residue states correctly [15]. Version IV of the GOR method was reported by the authors to perform single sequence prediction accuracy with an accuracy of 64.4% [18], as assessed through jackknife testing over a database of 267 proteins with known structure. The jackknife test represents a test scenario for those methods that need to be tuned using a training database. In its simplest form for a database containing $N$ sequences with known tertiary structure, each time a prediction is made for one sequence after training the method on a database containing the $N-1$ remaining sequences (one-at-a-time jackknife testing). A complete jackknife test would involve $N$ such predictions. If $N$ is large enough, meaningful statistics can be derived from the observed performance. For example, the mean prediction accuracy and associated standard deviation give a good indication of the sustained performance of the method tested.

RECENT COMPUTATIONAL APPROACHES – PREDICTION BASED ON MULTIPLE SEQUENCE ALIGNMENTS

The early methods by Nagano [1], Chou and Fasman [2] and Garnier, Osguthorpe and Robson [16] (GOR) exploit compositional biases exhibited by the three types of secondary structure. Following these pioneering prediction achievements, attempts have been made to enhance secondary structure prediction for single sequences by using sequence information at a more abstract and general level, such as the hydrophobicity rules applied in Lim’s method [3], through pattern matching techniques [95,19-22]. To define such sequence patterns, a number of researchers have used the physicochemical amino acid characterisations of Taylor [23], who utilised Venn diagrams to represent partially overlapping residue type groups, classified according to physicochemical features such as polarity, charge, size and the like.

In 1987, Zvelebil et al. [24] for the first time exploited multiple sequence alignments to predict secondary structure automatically by extending the GOR method and reported that predictions were improved by 9% compared to single sequence prediction. Multiple alignments, as opposed to single sequences, offer a much improved means to recognise positional physicochemical features such as hydrophobicity patterns. Moreover, they provide better insight into the positional
constraints of the amino acid composition. Finally, the placement of gaps in the alignment can be indicative for loop regions (see above). Levin et al. [25] also quantified the effect and observed 8% increased accuracy when multiple alignments of homologous sequences with sequence identities of ≥25% were used. As a consequence, the current state-of-the-art methods all use input information from multiple sequence alignments.

Novel computational concepts, partly coming from unrelated fields, have also been used to optimise the implementation of observed patterns in mapping the primary onto the secondary structure, thus enhancing the success rate of prediction. The most important examples include neural network applications [35,26], nearest-neighbour methods [27-30], linear discriminant analysis [31,32] and hidden Markov modelling (HMM) approaches [33]. Various recent methods relying on these concepts will be described in the next section. For most of the methods, a WWW server is available, of which the addresses are given in Table 1.

Neural Network Methods

Neural networks are learning systems based upon complex non-linear statistics. They are organised as interconnected layers of input and output units, and can also contain intermediate (or "hidden") unit layers (for a review, see ref. [34]). 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 has to be trained, which is done by adjusting the weights of the internal connections to optimise the grouping of a set of input patterns into a set of output patterns. A neural network has to be regarded as a black box, as it is normally difficult to understand the internal functioning of the network. Although a neural network is a powerful learning tool, there is a risk of overtraining the network, which leads to proper recognition of those patterns the network has been confronted with during training, but much less successful recognition of patterns that have not been seen.

PHD

The PHD method (Profile network from HeiDelberg) [35] combines the information from multiple sequence alignments with the optimisation strength of the neural network formalism. The PHD method makes use of three consecutive complete neural networks:

1. 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 those of the two 6-residue flanking regions; i.e., a 13-residue window \( w=13 \) is slid over the multiple alignment and each time secondary structure 3-state propensities are predicted for the residue in the central window position.

2. A second network refines the first level predictions by taking in information from more flanking residues. It takes as input the 3-state propensities of the first network and processes the information using a slightly longer 17-residue window. The output of the second network then comprises for each alignment position 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 \((HHH \, HH \, HH)\).

3. 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 network pairs 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 simple filtering step to delete predicted helices of one or two residues and changing those into coil. The PHD method was trained on a non-redundant set of 130 alignments from the HSSP database [36], 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 involved transmembrane protein families, which were deleted from the test). To avoid prohibitive computation time, the alignments were divided in 7 non-overlapping groups (each containing 18 test structures) for jackknife testing.
Although 70.8% is not the highest accuracy reported, the PHD method has been refined since this assessment and is among the best performers in the field. Moreover, it shows a sustained performance as compared with other methods available on the web, and thus is likely to come up with a reasonable prediction for most cases. A Webserver is available for the PHD method (Table 1). If 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 [36]. The resulting alignment is then fed into the actual PHD neural net algorithm. It is recommended to also attempt alternative multiple alignment methods to optimise the accuracy of the PHD predictions (see below).

**Pred2ary**

Another accurate profile and neural net-based prediction method is Pred2ary (26) 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 (35). 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 increased accuracy is achieved by converting each possible pair of network output weights for helix and strand into an *a priori* probability for the pair to predict the true structural state. These probabilities are then used for a final prediction corresponding to the highest of the *a priori* probabilities for each of the three states. The Pred2ary method is accessible through the Web (Table 1).

**PSIPRED**

A recent method that incorporates multiple sequence information and neural nets is PSIPRED [37]. However, the method exploits position specific scoring matrices (PSSMs) as generated by the PSI-BLAST algorithm [38] and feeds those to a two-layered neural network. The PSSMs used are not global alignments (see below) as with the PHD or Pred2ary methods, but contain information from local fragments of sequences that are deemed homologous to the query sequence (1 most similar fragment per homologous sequence is included). Since the method invokes the PSI-BLAST database search engine to gather information from related sequences, the method only needs a single sequence as input. Further, the inherent complexities of multiple training and use of a third neural net, as in both the PHD and Pred2ary methods have been avoided. Nonetheless, the accuracy of the PSIPRED method is 76.5%, as evaluated by the author [37]. This top-accuracy was confirmed by blind tests at

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDATOR¹</td>
<td>[29,30]</td>
<td><a href="http://www.embl-heidelberg.de/cgi/predator_serv.pl">http://www.embl-heidelberg.de/cgi/predator_serv.pl</a></td>
</tr>
<tr>
<td>PSIPRED</td>
<td>[37]</td>
<td><a href="http://insulin.brunel.ac.uk/psipred/">http://insulin.brunel.ac.uk/psipred/</a></td>
</tr>
</tbody>
</table>

¹Method can also be run using the JPRED server.
the CASP3 meeting, where assessments of the state-of-the-art prediction methods have been made. Moreover, the method is fast and can be easily ported to any common computer system. A webserver for the method is also available (Table 1).

**k-Nearest Neighbour Methods**

As with neural network methods, the application of $k$-nearest neighbour methods requires an initial training phase in which a collection of so-called exemplars is gathered. This collection 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 (or multiple alignment) and for each window the $k$ most similar fragments are determined using a certain similarity criterion. The distribution of the thus obtained $k$ 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**

In the method of Yi & Lander [27], a database of 110 proteins with known tertiary structure was used to derive a large collection of 19-residue fragments (exemplars), of which the environmental states were noted. For each 19-residue window slid over the query protein, 50 nearest neighbour exemplars were identified using the amino acid environmental scoring system of Bowie et al. [39], which includes as environmental parameters the secondary structure state, accessible surface area and polarity. It then scores the likelihood of a query residue type to be in a particular state (or range of values describing the state) over these three environmental parameters. As a score, the average was taken over 19 residues within a query window matched with the 19-position exemplar considered. For each exemplar, a cut-off score was determined, which should be met by a query fragment compared with it to count the exemplar as a neighbour and include its secondary structure label. The 50 thus obtained exemplars for each query window showed a distribution of the associated secondary structure labels, from which probability estimates for the three structural states were derived. Yi and Lander explored various scoring systems and found that the best protocol included 15 environmental classes (3 secondary structures combined with 5 different accessibility/polarity classes) in conjunction with an amino acid exchange score taken from the Gonnet et al. amino acid exchange matrix [40]. This means that the sensitivity of the method was not optimal when only environmental parameters for the exemplars were used, such that sequence information had to be included to arrive at the reported prediction accuracy of 67.1%. By using a neural network for a jury decision over six different scoring systems, the final accuracy was elevated to 68%, as assessed through jackknife testing.

**NNSSP**

The NNSSP (Nearest Neighbour Secondary Structure Prediction) [28] method adopts the nearest neighbour approach of Yi and Lander [27]. Differences with the Yi and Lander method are:

1. N- and C-terminal positions of helices and strands; and β-turns are explicitly taken as additional secondary structure types.
2. Predictions are made for multiple alignments rather than single sequences, so that insertions/deletions within alignment regions are explicitly taken into account.
3. The database of exemplars (see above) is for each prediction restricted to sequences similar to the query multiple alignment. This reduces computation times and leads to biologically related nearest neighbours.

Salamov and Solovyev [28] explored various window lengths and finally combined window sizes of 11, 17 or 23; nearest neighbour numbers ($k$) of 50 or 100, and balanced or non-balanced training, leading to a total number of $3 \times 2 \times 2 = 12$ different prediction set-ups. A simple majority rule over these 12 predictors increased the accuracy by 0.9%, as compared to the individual best of the 12 predictors. A few simple filters were also effected to refine the thus obtained predictions:

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

3. Helices of length \( \leq 4 \) are deleted. This rule is only applied after a full cycle of rule 1 and 2.

The overall accuracy of the method is 72.2\% [28], as assessed through a one-at-the-time jackknife test over the database of 126 proteins by Rost and Sander [35]. A WWW server for the NNSSP method is available (Table 1).

**PREDATOR**

The PREDATOR method of Frishman and Argos [29, 30] owes its accuracy mostly to the incorporation of long-range interactions for \( \beta \)-strand prediction. The method attains 68\% prediction accuracy for single sequence prediction, which was assessed using a one-at-a-time time jackknife test over the protein set of Rost and Sander (RS) [35]. Using a \( k \)-nearest neighbour approach (with \( k = 25 \) and 13-residue windows), propensities for the general three states (\( P^\alpha \), \( P^\beta \) and \( P^\gamma \)) were determined for each residue. Using pair-wise potentials involving long-range interactions, two more propensities for \( \beta \)-strand were determined. This was done by assessing the likelihood for all pair-wise 5-residue fragments (separated by more than six amino acids) to form parallel or anti-parallel \( \beta \)-bridges, based on summing residue hydrogen bonding propensities obtained from known structures. Two sets of propensities, one for anti-parallel and another for parallel bridges, were derived for this from a large collection of \( \beta \)-sheet structures. As the final parallel and anti-parallel \( \beta \)-strand propensity for each residue (\( P^\text{Par} \) and \( P^\text{Antipar} \)), the maximum (anti-)parallel window score was taken. This means that the score of the best possible interacting (anti-)parallel region was taken. Pair-wise hydrogen bonding potentials were also determined for \( \alpha \)-helical residues at a sequence separation of four residues, by summing over a 7-residue window to arrive at an extra helix propensity for each residue (\( P^\text{Helix} \)). A last additional propensity concerning \( \beta \)-turns (\( P^\text{Turn} \)) was obtained by summing single-residue propensities in classic \( \beta \)-turn positions 1-4 [41] using a four-residue window. For each of the thus obtained seven independent propensity values, threshold values (\( T \)) were calculated and used in the following five rules. The rules are applied consecutively to get a three-state prediction for each residue:

1. If (\( P^\text{Par} > T^\text{Par} \) or \( P^\text{Antipar} > T^\text{Antipar} \)) and \( P^\text{Helix} < T^\text{Helix} \), then predict \( \beta \)-strand; otherwise, if \( P^\text{Helix} > T^\text{Helix} \), then predict \( \alpha \)-helix, otherwise predict coil.

2. If \( P^\text{C} > T^\text{C} \), then predict coil.

3. If \( P^\text{E} > T^\text{E} \), then predict \( \beta \)-strand.

4. If \( P^H > T^H \), then predict \( \alpha \)-helix.

5. If \( P^\text{Turn} > T^\text{Turn} \), then predict coil.

Apart from the novel scheme to include long-range interactions for strand prediction, the PREDATOR method can also predict using multiple sequence information. However, PREDATOR does not use or construct a multiple alignment, but compares the sequences using pair-wise local alignments [42]. In this sense, the method can be viewed as a predecessor of the successful PSIPRED method (see above), which employs comparable information in the form of local alignment-based position-specific scoring matrices (PSSM) generated by the PSI-BLAST program. Predictions by PREDATOR are carried out for a single base sequence and a number of related sequences. A set of highly scoring local alignments is compiled through matching the base sequence with each of the other sequences. In contrast to PSI-BLAST, more than one local alignment per matched related sequence can be selected (Fig 5). A weight is then compiled for each local fragment based on the score and length of its alignment with the base sequence. For each residue in the base sequence, after gathering exemplars using the base sequence and the stacked fragments, the weighted sum over all exemplars 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 results in a per-residue accuracy of 74.8\% [30], as assessed using one-at-a-time jackknife testing over the RS protein set. PREDATOR is accessible via a Web server (Table 1).

**Linear Discriminant Analysis - the DSC Method**

The DSC method [31] combines the compositional features of multiple alignments with empirical rules that are found important for secondary structure prediction. The information is
processed using linear statistics. The rules and concepts used relating to multiple alignment information are: (i) N-terminal and C-terminal sequence fragments normally adopt a coiled structure; (ii) Alignment positions comprising gaps are indicative for coil regions; (iii) Periodicity in positions of hydrophobic or conserved residues; and (iv) Residue ratios in the alignment. These patterns are detected using autocorrelation, feeding back of predicted secondary structure information and some simple filter rules. Prediction occurs in five consecutive steps. (1) The basic prediction of the secondary structure is carried out using the GOR method (see above), which is used on each of the aligned sequences. The average GOR score for each of the three states is then compiled for each alignment position. (2) For each alignment position a so-called attribute vector is compiled, consisting of 10 attributes: the three averaged GOR scores for H, E and C from step 1; 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. (3) The positional vectors are doubled in number to 20 attributes by adding the same 10 attributes in a smoothed fashion (using running averages). (4) Seven more attributes are added to the 20 attributes of the preceding step: weights for predicted α-helix and β-strand, based on the 20-attributes vectors of step 3, and the fractions of the five most discriminating residue types; His, Glu, Gln, Asp and Arg. To convert these 27-attribute vectors to three-state propensities, a linear discrimination function is used. This effectively is a set of weights for the attributes in the positional vector corresponding to each of the secondary structure states, so three sets of 27 attribute weights are used. The optimal weights used in the DSC method have been gathered using a training set of known 3D structures. After applying the weights to the attribute vectors for each alignment position, the secondary structure associated with the highest scoring vector is taken. (5) A set of 11 simple filter rules are used for a final prediction, such as, for example, \([E/C][E][H/E/C][H/C]\)→C, where [E/C] denotes E or C. These filter rules have been constructed automatically using machine learning techniques.

The accuracy of DSC, as assessed by the authors based on the Rost-Sander protein set, comprises 70.1% [31]. The method shows best performance for moderately sized proteins in the range 90-170 residues. As an additional option, the DSC method can also be used to refine a prediction by the PHD algorithm (see above). The average Q3 of this PHD-DSC combinatorial procedure is 72.4% (King and Sternberg, 1996), which is 0.6% higher than the accuracy of PHD alone. A WWW server for the DSC method is available (Table 1).

**Using Tertiary Structural Models – the PSA Method**

The PSA method [33] predicts secondary structure for single sequences but employs tertiary structural information using a hidden Markov modelling (HMM) approach. The method is based on a threading-like approach (see above), in that for a query sequence the goodness of fit is tested with 15 basic tertiary structural models, named discrete space models (DSM). Included as DSMs are for example the α-helical globin structure and the flavodoxin-type α/β fold. Each is of these models is composed of secondary structure proteins.
elements chosen from 13 types distinguished by the method: N- and C-cap (for helix); average, buried and exposed α-helix; buried β-strand; buried and exposed amphipathic β-strand; 4 β-turn positions; and coil. Using the HMM formalism for modelling the DSMs, the most suitable implementation for each DSM (e.g. the flavodoxin-type DSM can hold 5 to 7 helices) is selected and the secondary structure of the best fitting model is then presented to the user as a probability contour plot. Although the method might not be among the best performers, the approach is interesting and the PSA graphic outputs are useful as a starting point for gaining insight into the probabilities for each of the secondary structures along the sequence.

**Using Secondary Structure Specific Exchanges – the SSPRED Method**

The SSPRED method [43] exploits an alternative aspect of the positional information provided by multiple alignments, in that it uses the amino acid pair-wise exchanges observed for each multiple alignment positions. Using the 3D-ALI database [44], which holds multiple alignments of distantly homologous proteins constructed based on structure superpositioning and sequence alignment, amino acid exchange matrices were compiled for helix, strand and coil, respectively. Each matrix simply contains preference values for amino acid exchanges observed at alignment positions with the corresponding secondary structure in the 3D-ALI database. The matrices are used to predict the secondary structure of a query alignment through listing the unique observed residue exchanges for each alignment position and summing the corresponding preference values over each of the three exchange matrices. Each exchange type (e.g. alanine to/from proline) is counted only once for each query alignment position, which provides implicit weighting of the sequences to avoid predominance of redundant sequences. The secondary structure corresponding to 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:

1. Single position interruptions: (H[E/C]H) becomes (HHH) and (E[H/C]E) is set to (EEE), where [E/C] indicates E or C.
2. Double position interruptions: (HH[EE/CC]H) or (H[EE/CC]HH) becomes (HHHHH) and (EE[HH/CC]E) or (E[HH/CC]EE) is changed to (EEEEEE), where [EE/CC] designates EE or CC.
3. Short fragments: helices ≤ 4 residues and strands ≤ 2 residues are changed into coil.

The accuracy of the method was assessed using a one-at-a-time jackknife test and amounted to 72%, albeit over a relatively small test set of only 38 protein families. Nonetheless, an advantage of the technique is its speed and conceptual clarity.

**Running Multiple Methods for Consensus Prediction – JPRED**

The JPRED server [45] (Table 1) at the EMBL-European Bioinformatics Institute (Hinxton, U.K.) conveniently runs state-of-the-art prediction methods such as PHD (35), PREDATOR (29,30), DSC [31] and NNSSP [28], while the methods ZPRED (24) and MULPRED (Barton, unpublished) are also included. The NNSSP method has to be activated explicitly, as it is the slowest of the ensemble and often will not be finished in the computing time slot allocated to the user. 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 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 then taken as the consensus prediction for the position. In case no sufficient agreement is reached, the PHD prediction is taken. This consensus prediction is somewhat less accurate when the NNSSP method is not included. An example of output by the JPRED server for the signal transduction protein cheY (PDB code 3chy) is given in Fig 6 (vide infra). The JPRED server also accepts a single query sequence, in which case it constructs a set of related sequences by launching a BLAST search [46]. The CLUSTAL X [47] multiple alignment method (see below)
Fig. (6). Secondary structure prediction of protein cheY by the JPRED server. Individual predictions by five methods are given as well as a consensus prediction (for details, see text). Prediction are performed for the alignment of the cheY sequence (PDB code 3chy) with 13 flavodoxin sequences by the PRALINE method. Under the alignment, strands are depicted as arrows, while helices are given as cylinders. Predicted residue solvent accessibility values are also given (Acc): dark boxes indicate buried residues, while transparent boxes represent residues predicted as exposed.

is subsequently invoked to align the query sequence with the found tentative homologues. The resulting multiple alignment is then subjected to the actual JPRED consensus prediction technique.
PREDICTION METHODS FOR TRANS-MEMBRANE REGIONS

Membrane proteins are associated with the cell membrane and comprise one or more transmembrane (TM) segments. The organisation of the TM segments is severely restricted by the lipid bilayer of the cell membrane, in contrast to soluble globular proteins where a great variety in the mutual orientations of individual secondary structural elements can be observed. Because of the restrictions imposed by the hydrophobic environment within the cell membrane, the prediction of TM segments is easier in principle than prediction of their soluble counterparts. This is due to the relatively strong tendency of certain hydrophobic amino acid types with special physico-chemical properties to occur in membrane spanning regions. On the other hand, only few X-ray or NMR tertiary structures of TM proteins have been elucidated to date [48]. Also, experimental data on TM boundaries are not very precise, as they are acquired from site-directed mutagenesis, enzymatic cleavage, immunological methods, and the like. The paucity in accurate structural data limits the possibility of training TM prediction methods, and potentially affects the accuracy of database-oriented methods.

The most frequently observed secondary structure in trans-membrane segments is the $\alpha$-helix, albeit also a few protein families involving TM $\beta$-strands, typically constituting $\beta$-barrel topologies, have recently been identified. TM prediction efforts have generally been focussed on the determination of the membrane sequence boundaries and their tentative orientation with respect to the membrane. As to the latter, it was observed in prokaryotes, and to a lesser degree in eukaryotes, that positively charged residues within TM-connecting loop regions tend to reside in the cytosol, a trend coined the “positive inside rule” [49,50]. The available TM prediction methods generally assume the $\alpha$-helical structure, as this structure is most appropriate due to its ability to mutually satisfy main-chain hydrogen donors and acceptors through formation of hydrogen bonds (see above). This is particularly important given the severe hydrophobicity of the membrane environment.

The early approaches to TM prediction relied on the recognition of stretches of hydrophobic amino acids, and typically delineated those using sliding window approaches, so that TM regions would appear as peaks in hydrophobicity plots, also called hydropathic profiles. The window length should correspond to the expected length of a transmembrane segment, in the range 16-25 residues. Given that the average membrane thickness is about 30 Å, about 20 residues form a helix stretching from one lipid bilayer surface to another. A classical hydrophobicity scale in TM prediction was developed by Kyte and Doolittle [51] and derived from globular protein interior data. Following Kyte and Doolittle, many different prediction techniques have been devised over the years. Although a detailed description of TM prediction methods is outside the scope of this review (for an overview of methods for TM prediction, see Ref. [52], references for current state-of-the-art prediction methods and related WWW services are given in Table 2. These can

<table>
<thead>
<tr>
<th>Method</th>
<th>Reference</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>TMHMM</td>
<td>[107]</td>
<td><a href="http://www.cbs.dtu.dk/services/TMHMM-1.0/">http://www.cbs.dtu.dk/services/TMHMM-1.0/</a></td>
</tr>
</tbody>
</table>
also be conveniently activated from the ExPaSy webserver (http://www.expasy.ch/tools/).

No general method appears available aimed at predicting TM β-strands. Delineating those is indeed a difficult task: While every second β-strand residue protruding into the lipid bilayer tends to be hydrophobic, the side chains facing the pore of the β-barrel display no definitive tendency and can therefore dramatically lower the amphipathic signal. Moreover, the number of amino acid residues in β-strands needed to span the membrane is much smaller than that for the helical conformation, typically about 10 only, so that they can go undetected easily by smoothed hydropathic profile approaches.

**MULTIPLE ALIGNMENT QUALITY AND SECONDARY STRUCTURE PREDICTION**

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. Over the years, many different multiple alignment methods have been developed, featuring different fundamental strategies and heuristics, which can result in significant differences between the alignments produced by these methods, particularly when divergent sequences are included.

**Multiple Alignment Methods**

Multiple sequence comparison involves the search for similarity in more than two sequences. The alignment of a set of sequences can provide important information about structure-function relationships within the proteins, such as the evolutionary conservation of functional amino acids at certain sequence positions or conserved hydrophobicity patterns in particular regions. A multiple alignment often serves as a starting point for site-directed mutagenesis experiments. In addition to secondary structure prediction, there are also other techniques for sequence analysis which rely on a multiple alignment, including phylogenetic analysis, and threading techniques (see above).

Two basic classes of multiple alignment programs have been developed: global and local methods. Global alignment programs attempt to align the sequences over their whole length, whereas local programs search only for the most conserved regions and leave the other parts of the sequences unaligned. The most effective alignment algorithm depends on the nature of the sequences to be aligned. Global algorithms produce the most accurate and reliable alignments when all the sequences in the data set are of similar length. However, when the sequences differ greatly in length, local alignment programs are often more successful in identifying the conserved regions.

The two most explored computational techniques for multiple sequence alignment are the dynamic programming (DP) technique [53,54] and, more recently, Hidden Markov Modelling (HMM) [55,56]. While the DP technique is deterministic, HMM is a stochastic approach, which has proven powerful if applied to sequence database searches. Krogh et al. [56] described a HMM procedure for multiple alignment in which the alignment process is modelled in a finite automata fashion with three basic alignment steps considered as states: match, insert and delete. Probabilities are attached to the state transitions by an expectation maximisation algorithm trained during the alignment such that position-dependent amino acid substitution, insert and delete probabilities are generated. Although HMM methods incorporate more detail than the classical DP methods, such as the mentioned position-specific scoring schemes, HMM approaches to multiple sequence alignment generally perform poorly when compared with other methods [57,58]. This is mainly due to the inherently complex parameterisation of the technique, often leading to local traps in the sequence space searched. As a consequence, the best performing methods for multiple alignment are based on the DP technique.

**Global Multiple Alignment Methods**

The problem of finding an optimal or highest scoring global alignment of two sequences was solved three decades ago with the DP technique [53], which guarantees the finding of the highest scoring or optimal alignment based on an amino acid substitution scoring scheme and insertion/deletion penalties. Unfortunately, the calculation of such an optimal simultaneous alignment becomes computationally unfeasible for four or more sequences. In general, an algorithm for
optimal alignment needs a number of computational steps and an amount of memory of at least the order of the product of the sequence lengths. Rigorous methods for the simultaneous alignment of four or more sequences thus cannot evaluate all possible matches, but attempt to find the optimal alignment by considering only a small fraction of all possible comparisons. Over the years, various heuristic approaches have been developed leading to a large number of multiple alignment programs adopting different strategies. Carillo and Lipman [59] showed that the optimal alignment path of \( N \) sequences can be limited to a small region in the \( N \)-dimensional search matrix, of which the upper bounds can be inferred from pairwise comparisons of the sequences. The algorithm MSA [60] is based on this approach and generalises the pair-wise diagonal strip method of Ficket [61] to \( N \) dimensions, where \( N \) is the number of sequences to be aligned. Up to 9 sequences of 200-300 residues in length can be aligned with the MSA method. Recently, the basic MSA technique was extended in order to cope with larger data sets using a divide-and-conquer strategy [62] implemented in the method DCA [63]. However, the approach remains extremely CPU and memory intensive and thus applicable only to small data sets.

In general, the most popular and successful approach has been the progressive alignment strategy [64,65] where a multiple alignment is built up gradually by aligning the closest sequences first and successively adding in the more distant ones. Most widely used methods thus work in an agglomerative way by aligning sequences, following an heuristically determined order, until all sequences are joined in a final multiple alignment. The sequences are aligned either according to a previously determined order which, for instance, can be derived from a known

![Diagram](a)
Fig. (7). Guide tree and alignment by the program CLUSTALX [47] of the signal transduction protein cheY (PDB code 3chy) and 13 flavodoxin sequences. (a) Guide tree based on pair-wise sequence identity values. (b) Multiple alignment of the 14 sequences.

phylogenetic tree, or is reassessed at each step during the progressive alignment. Most present day methods use a dendrogram constructed from the pair-wise sequence similarities as a guide tree and then invoke a DP algorithm to compare the sequence pairs, a block of aligned sequences with a single sequence, or two blocks of aligned sequences (Fig 7). The initial step thus involves performing all pair-wise comparisons between the sequences, which has a time complexity of $N^2L^2$, where $N$ is the number of sequences and $L$ the average sequence length. This complexity is often the bottleneck for computing large alignments. However, using a guide tree is a good heuristic as the sequences are progressively aligned from similar to divergent, which results in less error because alignment of similar sequences is more accurate than that of distant sequences, such that error propagation is minimised. A consequence of this scenario is that whenever a gap is introduced in any sequence (block) during an alignment step, the gap will remain in further steps.

Hogeweg and Hesper [64] were the first to devise an integrated agglomerative algorithm. In their method, a dendrogram is constructed based on all pair-wise similarities of sequences matched by dynamic programming. The sequences are aligned progressively following the branch order of the dendrogram. During the alignment, internode or ancestral sequences are constructed to represent already aligned groups of sequences. Hogeweg and Hesper [64] argued that a multiple alignment and the associated phylogenetic tree cannot be separated. They showed this by using an iterative procedure for the first time: From the initial tree based on pair-wise alignments, carrying no
information yet of related groups of sequences, a multiple alignment is generated and the pair-wise similarities associated with the latter alignment are inferred. Then, iteratively a new tree is constructed from which a succeeding alignment is created, based on the increased information.

Higgins and Sharp [66] early on constructed a fast and widely used method CLUSTAL which was especially designed for use on small workstations. Speed was obtained during the pair-wise alignments of the sequences, which is the most time-consuming step of progressive alignment, by applying the fast pair-wise diagonal algorithm of Wilbur and Lipman [67,68]. From the pair-wise similarities, a tree is constructed using the UPGMA clustering criterion. The sequences are then aligned following the branching order of the tree. For the comparison of groups of sequences, Higgins and Sharp [66] used consensus sequences to represent aligned subgroups of sequences and also employed the Wilbur-Lipman technique to match these. Since its earliest version, the CLUSTAL package has been subjected to a number of revision cycles. Higgins et al., [69] implemented an updated version CLUSTAL V in which the memory-efficient dynamic programming routine of Myers and Miller [70] is used, enabling the alignment of large sets of sequences using little memory. The largely extended version CLUSTAL W [71] uses the Neighbour-Joining (NJ) algorithm [72], which is widely used in phylogenetic analysis, to construct a guide tree. Sequence blocks are represented by a profile, in which the individual sequences are additionally weighted according to the branch lengths in the NJ tree. An integrated user interface has been implemented in CLUSTAL X [47] comes with accessory programs for tree depiction. The CLUSTAL W method together with its interface suite CLUSTAL X have generally become the most popular methods for multiple sequence alignment.

The PILEUP routine from the GCG package [73] follows the earlier version V of CLUSTAL closely. It generates an UPGMA-based tree and, for the alignment of two sets of matched sequences, uses the average alignment similarity score of Corpet [74].

The method MULTAL of Taylor [75] is extremely rapid and constructs a tree during the progressive alignment as in the method of Feng and Doolittle [65]. It uses a fast sequential branching method to align the closest pairs of sequences first and then subsequently align the next closest sequences to those already aligned. The order in which the sequences are aligned is largely based on the global amino acid composition of the sequences. However, the blocks of sequences are aligned using dynamic programming during the alignment progression.

The recent method PRALINE [5] does not use a pre-calculated search tree but performs at each alignment step a full profile search with the most recently aligned sequence block. This is to re-evaluate at each alignment step which sequences or blocks of sequences should be aligned, so that the alignment order is determined during the progressive alignment. The technique offers a number of strategies to optimise the quality of multiple alignment, including profile pre-processing and secondary structure prediction-based alignment. The profile pre-processing strategy is aimed at incorporating reliable information from other sequences into each sequence, before progressive alignment takes place. For each sequence, an alignment is created by stacking other sequences (N-to-1 alignment) using their pair-wise alignment with the sequence considered if the alignment scores beyond a user-specified threshold when. The effect of a low threshold is that many distant sequences will be included each pre-processed alignment, while higher thresholds would allow less sequences into the alignments, but matched with greater reliability. For each of the thus formed pre-processed alignments, a profile is constructed, so that effectively for each sequence, position-specific compositional information as well as position-specific gap penalties is created. PRALINE then performs progressive multiple alignment, where each sequence is now represented by its pre-processed profile. The pre-processed profiles for each of the sequences thus incorporate knowledge about other sequences, in particular similar sequences. This enables increased matching of distant sequences and likely placement of gaps outside the ungapped core regions in the pre-processed profiles during progressive alignment. The multiple alignment of the preprocessed profiles can also be used to derive consistency scores for each amino acid in the final alignment, which reflect the consistency of the position of the residue within its associated pair-wise alignments.
The second strategy of exploiting secondary structure prediction to optimise alignments (and _vise versa_) will be described below.

Several new alignment algorithms have recently been developed, where a common point of interest has been the application of iterative strategies to refine and improve the initial multiple alignment, a concept which has been conceived and applied first nearly two decades ago [64]. Iterative strategies provide an interesting alternative to simultaneous alignment (see above), as they are applicable to relatively large data sets. Although they do not provide any guarantees about finding an optimal solution, they are reasonably robust and certainly much less sensitive to the number of sequences than, for example, the aforementioned deterministic method MSA [60]. The PRRP program [76] optimises a progressive, global alignment by iteratively dividing the sequences into two groups which are subsequently realigned using a global group-to-group alignment algorithm. Pair-wise sequence weights are derived from a tree constructed with the UPGMA cluster criterion and used to calculate the alignment scores when sequence blocks are matched. The PRRP method [76] is especially sensitive for sequence sets containing distantly related sequences. The aforementioned PRALINE method [5] also features iteration schemes: one is based on increasing the weights of consistently aligned residues in subsequent iterations, thereby driving the alignment towards greater overall consistency. Another scheme involves updating the stacked sequences within each pre-processed profile according to the constructed multiple alignment before a subsequent alignment of the preprocessed profiles takes place. A third scheme is based on iteratively performing multiple alignment and secondary structure prediction (see below).

**Local Multiple Alignment Methods**

Local multiple alignment methods focus on the comparison of conserved motifs in a set of protein sequences. Motif-based methods generally align shorter and ungapped sequence fragments and succeed when these fragments are recognisable as motifs in all or most of the sequences. Most motif-based methods therefore are not particularly suitable for the alignment of distant sequence groups. Generally, as these methods are not designed to yield full alignments, they should be attempted on sets of sequences with varying lengths if there is suspicion of shared motifs or domain structures.

The Boguski _et al._ [77] semi-manual program suite holds the space-efficient local alignment routine SIM of Huang _et al._ [78] as well as the MSA method for multiple sequence alignment [60]. For each pair of sequences, the highest scoring local alignments, containing gaps, are determined from which non-gapped regions occurring in each of the sequences are extracted. Neighbouring blocks of such motifs with the intervening sequence fragments are then aligned using the MSA method, thus allowing gaps. The MACAW method [79] is semi-automatic and produces blocks of alignments shared by all or a subset of the sequences. Lawrence _et al._ [80] developed an automatic multiple alignment strategy based on the Gibbs statistical method of iterative sampling. The GIBBS algorithm searches for gap-free motifs of a certain pre-set length _W_, which are found by a random optimisation procedure. The recent method DIALIGN [81] is a segment-based local procedure, which constructs a multiple alignment by assembling a collection of high scoring segments in a sequence independent progressive manner. The segments are incorporated into a multiple alignment using an iterative procedure. Other local programs include MEME [82] or ITERALIGN [83]. In the latter method, individual sequences are edited by replacing amino acids by those that are preponderant at a corresponding position in a multiple alignment to achieve better recognition of crucial alignment regions.

**Alternative Techniques**

In addition to the DP technique, also other computational strategies have been attempted. The program SAGA [84] uses a genetic algorithm (GA) to select from an evolving alignment population the alignment which optimises, as an Objective Function (OF), the weighted sum of pairs as used in the MSA program. More recently, a measure of consistency between the considered multiple alignment and a corresponding library of CLUSTAL pair-wise alignments was taken [57]. Although the method is sensitive, the inherent time consumption makes the SAGA method unfeasible for larger alignments. As mentioned above, hidden
Markov models (HMM) have also been attempted as statistical models of the primary structure consensus for a sequence family [55,56]. The program HMMR [85] uses a simulated-annealing method to maximise the probability that a HMM represents the sequences to be aligned.

**Evaluating Multiple Alignment Methods**

Some recent evaluations of available multiple alignment techniques have been carried out [58,86] using the versatile database of benchmark alignments BAli BASE [87]. The database holds 141 expert alignments of sequences based partly on structural superpositioning of sequences with known tertiary structure. The alignments have been annotated to identify core alignment regions for which structural evidence is deemed conclusive. The evaluation studies based on BAliBASE revealed that the method PRRP [76] was marginally the most accurate, closely followed by CLUSTALX [47], which is a much faster program. Virtually the same accuracy as CLUSTALX was attained by the PRALINE method [5] when run on its default parameters [Notredame, unpublished results]; i.e., optimisation strategies such as profile-preprocessing or predicted secondary structure-induced alignment were not used. Other methods included in the assessment tests generally were less successful, such as the local alignment method DIALIGN [81], the HMM-based method HMMT [85] or the Gibbs-sampling method GIBBS [80]. The DIALIGN method however was relatively successful in aligning sequences with very large insertions or deletions.

**USING ALTERNATIVE MULTIPLE ALIGNMENTS AND SECONDARY STRUCTURE PREDICTION**

Using alternative alignments by running different multiple alignment methods can lead to dramatically different secondary structure predictions for each of the alignments. Heringa [88] constructed automatically two different

<table>
<thead>
<tr>
<th></th>
<th>2fcr</th>
<th>4fxn</th>
<th>FLAV_ANASP</th>
<th>FLAV_AZOV1</th>
<th>FLAV_CLOAB</th>
<th>FLAV_DESDE</th>
<th>FLAV_DESG1</th>
<th>FLAV_DESSA</th>
<th>FLAV_DESVH</th>
<th>FLAV_ECOLI</th>
<th>FLAV_ENTAG</th>
<th>FLAV_MEGEL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1fx1</td>
<td>29</td>
<td>30</td>
<td>29</td>
<td>18</td>
<td>17</td>
<td>44</td>
<td>54</td>
<td>53</td>
<td>95</td>
<td>35</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>2fcr</td>
<td>23</td>
<td>27</td>
<td>35</td>
<td>16</td>
<td>24</td>
<td>28</td>
<td>26</td>
<td>24</td>
<td>33</td>
<td>34</td>
<td>18</td>
<td>16</td>
</tr>
<tr>
<td>4fxn</td>
<td>20</td>
<td>21</td>
<td>25</td>
<td>28</td>
<td>33</td>
<td>46</td>
<td>39</td>
<td>21</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_ANASP</td>
<td>49</td>
<td>23</td>
<td>28</td>
<td>33</td>
<td>31</td>
<td>46</td>
<td>39</td>
<td>21</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_AZOV1</td>
<td>22</td>
<td>22</td>
<td>28</td>
<td>26</td>
<td>26</td>
<td>45</td>
<td>47</td>
<td>23</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_CLOAB</td>
<td>18</td>
<td>18</td>
<td>18</td>
<td>21</td>
<td>26</td>
<td>20</td>
<td>20</td>
<td>14</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_DESDE</td>
<td>47</td>
<td>44</td>
<td>47</td>
<td>35</td>
<td>25</td>
<td>30</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_DESG1</td>
<td>55</td>
<td>58</td>
<td>32</td>
<td>28</td>
<td>34</td>
<td>18</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_DESSA</td>
<td>56</td>
<td>33</td>
<td>27</td>
<td>35</td>
<td>12</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_DESVH</td>
<td>37</td>
<td>26</td>
<td>28</td>
<td>16</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_ECOLI</td>
<td>41</td>
<td>26</td>
<td>19</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_ENTAG</td>
<td>24</td>
<td>20</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FLAV_MEGEL</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
alignments of the signal transduction protein cheY (PDB code 3chy) with 13 distant flavodoxin sequences, using the popular multiple alignment program CLUSTALX [47] and the PRALINE method [5], respectively. The 3chy structure adopts the basic flavodoxin fold (see Fig. 2) despite very low sequence similarities with genuine flavodoxins. As the cheY sequence is evolutionarily extremely divergent from the flavodoxin sequences, the sequence set is an example of a set of related sequences (although fairly divergent as well) with one distant orphan sequence (Fig 8). With the current genome projects, a plethora of such orphan sequence data sets are likely to become aligned and analysed. The CLUSTALX and PRALINE alignments were given to the JPRED server [45] (see above), which performed a consensus prediction using 5 different secondary structure methods. Fig 6 shows the PRALINE alignment with the corresponding secondary structure predictions by the techniques invoked by the JPRED server as well as the JPRED consensus prediction. The accuracy of this consensus prediction was 85.1%, in contrast to the accuracy of the JPRED consensus prediction based on the CLUSTALX alignment, which was 53.9% (Fig 9). The difference in accuracy of the two consensus predictions amounts to 31%, an order of magnitude more than the increase in prediction accuracy obtained over the last five years [88]. It might therefore be argued that there is a need for a consensus secondary structure prediction based on a variety of multiple sequence alignments by alternative high quality alignment methods. A homologous alignment for the 3chy sequence was also created by the JPRED server. The server constructed an evolutionary related set of homologues through a BLAST search [46] and aligned the 32 resulting sequences using CLUSTALX. The accuracy of the consensus secondary structure prediction for the latter multiple alignment was 88.3% (Fig 9) and hence 3% higher than that obtained for the PRALINE alignment of the cheY-flavodoxin set [88]. Moreover, the second β-strand of the 3chy structure was successfully predicted, whereas it was missed by the predictions based on the CLUSTALX or PRALINE alignments. The prediction results over these three example alignments stress the importance of having a balanced and non-redundant set of homologous sequences for accurate prediction. Also, including divergent sequences can greatly enhance the prediction accuracy. However, these sequences must be aligned properly, which often is a difficult task.

ITERATED MULTIPLE ALIGNMENT AND SECONDARY STRUCTURE PREDICTION

Most reliable secondary structure prediction methods utilise sequence information in multiple alignments and their prediction accuracy relies on the quality of the multiple alignment used. If in turn the multiple alignment would be guided by the predicted secondary structure, an iterative scheme arises that optimises both the quality of the multiple alignment and that of the secondary structure prediction. A procedure to use secondary structure information in multiple alignment is implemented in the PRALINE method [5]. Initially, a multiple alignment is constructed without information about the corresponding secondary structure (Fig 10). Using this alignment,
Fig. (10). Schematic representation of iterative multiple alignment/secondary structure prediction scheme. In the figure is indicated that information from secondary structure prediction programs based on single sequences as well as multiple sequences can be used.

the secondary structure is predicted (for which in principle any of the discussed methods can be taken). Then, a new alignment is constructed now using the predicted secondary structure. This process can be iterated until convergence is reached or a user-defined number of iterations have been carried out. As the PRALINE method employs dynamic programming to progressively construct a multiple alignment for a query set of sequences, it therefore relies on an amino acid exchange weights matrix and a pair of gap penalties [for a review, see Ref. 52]. The initial alignment is constructed using
a default residue exchange matrix (e.g. the BLOSUM 62 matrix) and gap penalties. After secondary structure prediction, resulting in a tentative secondary structure for each sequence (Fig 10), PRALINE utilises the thus obtained secondary structure information as illustrated in Fig 11. 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 [89] and associated gap penalties. The gap penalties have different values for the three secondary structures, in order to make it relatively difficult to insert gaps in helices and strands as compared to coil. As shown in fig 11, the residue exchange weights for matched sequence positions with identical secondary structure states are taken from the corresponding residue exchange matrix, whereas sequence positions with different secondary structure states are treated with the default exchange matrix. When blocks of sequences are aligned, the secondary structures are only considered for matched alignment positions at which all sequences show an identical state; otherwise, the default exchange matrix is taken. 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.

**ASSESSING SECONDARY STRUCTURE PREDICTION METHODS–THE STANDARD OF TRUTH**

How feasible is it to demand that prediction methods should ultimately predict each individual

![Diagram](image_url)
structure 100% accurately? Is our standard of truth up to such a demand? In the vast majority of evaluation studies for secondary structure prediction methods, accuracy is assessed by using known tertiary structures from the protein data bank (PDB) [13] with corresponding secondary structures assigned using the DSSP method of Kabsch and Sander [90]. Colloc'h et al. [91] compared the DSSP method with two other secondary structure determination algorithms. All three of these methods were designed to delineate objectively and automatically the observed secondary structures. Nonetheless, Colloc'h et al. found significant differences in their secondary structural assignments. The ambiguity in secondary structural assignments can be dramatic for particular proteins with agreement of the methods as low as 65% [92,93]. Even in sets of clearly homologous proteins with known tertiary structure, the corresponding secondary structural elements can vary in length or show shifts of one to a few residues. Moreover, tertiary structures elicited by Nuclear Magnetic Resonance (NMR) are often represented in the PDB by an ensemble of closely related structures, and also here the corresponding helix and strand assignments normally do not line up well. A realistic maximum prediction accuracy per residue would therefore be in the range 80-100% [94]. Many researchers early on have suggested that prediction evaluation should be based on the overlap of predicted and observed segments rather than on individual positions [95-100]. This is consistent with efforts to use predicted secondary structure in fold recognition or topology modelling, where recognition of the core secondary structures is more important than their exact delineation (vide supra). A recent secondary structure assignment program that integrates many of the features of earlier methods, such as checking hydrogen bonding patterns and stereochemical characteristics, is the knowledge-based method STRIDE [101]. It is claimed to yield assignments in close agreement to those made by crystallographic experts. The STRIDE method tends to be less strict in assigning helices and strands than the popular DSSP method. However, prediction methods do not necessarily come out better when using an assignment method in agreement with human experts, as tuning a prediction method according to strict assignments is often more easily accomplished. For example, Frishman and Argos [30] found that their PREDATOR method attained a (seemingly?) better accuracy when DSSP assignments were used as opposed to using those of STRIDE as the standard of truth. Furthermore, if only the DSSP method is used, the way in which the eight DSSP secondary structure characters (H: α-helix, G: 3/10-helix, I: π-helix, E: extended (β-strand), B: β-bulge, T: hydrogen-bonded turn, S: bend, ‘ ‘ : coil) are grouped into helix, strand and coil is significant for evaluating 3-state predictions. Jones [37] recently found a clear difference in prediction accuracy of the PSIPRED method depending on the mapping scheme used. When H, G and I were taken as helix while E and B were taken as strand, which is the scheme used in most assessments, the PSIPRED method showed an accuracy of 76.5%. However, when a simpler mapping scheme was used, where H is mapped to helix and E to strand while the rest is taken to be coil, the prediction accuracy increased to a massive 78.3%.

CONCLUSIONS

For secondary structure prediction methods relying on multiple alignments, it is generally important to assess the consistency and quality of a target alignment, since this can have dramatic consequences for the prediction accuracy. It is advisable to try more than one multiple alignment routine, in order to trace misaligned regions in the multiple alignment. As there is no single best prediction method for all cases, it is also recommended to use more than one secondary structure prediction method and construct a consensus prediction. A convenient aid for this is the JPRED server [45]. Alternative alignments can also be included in secondary structure prediction attempts, as this can provide insight in the consistency of the predictions. Some of the prediction methods offer positional reliability indices, which should be included in interpreting the results. The recognition of super-secondary or higher-order structural features from the predicted secondary structure elements can be used to interpret and correct the basic prediction results. Easily recognisable errors might be disruptions in alternating α-helix/β-strand predictions in a tentative α/β protein fold (e.g. the missed second β-strand in the flavodoxin example above) or the occurrence of a single β-strand within a probable α-helical protein. Any conserved features within a multiple alignment, such as likely active site
residues or disulphide bridges, should also be included in interpreting the predictions.

Important improvements in secondary structure prediction have come from computational concepts such as neural network and k-nearest neighbour techniques. Recent progress has been made as well based on new developments in multiple alignment methodology. An example is the new successful PSIPRED method [37], which relies on position-specific scoring matrices as generated by the recent PSI-BLAST program [38]. Further developments in multiple alignment methodology, such as the secondary structure-induced alignment strategy implemented in the PRALINE method [5], are likely to enhance the prediction technology. The rapid growth in the number of structures in the PDB and the current plethora of sequence data from genome sequencing projects will allow improved training and tuning of the prediction methods, which is expected to lead to a further increase in prediction accuracy.

REFERENCES


