Protein Secondary Structures: Prediction

John-Marc Chandonia, *University of California, San Francisco, California, USA*

Secondary structure prediction methods are computational algorithms that predict the secondary structure of a protein (i.e. α helices, β strands and turns) from the primary structure (the amino acid sequence).

Secondary article

Article Contents

Introduction

For most proteins, the three-dimensional (3D) structure is determined only by the amino acid sequence. Although it has been discovered that molecular chaperones present in cells speed the rate of folding and prevent misfolds, a protein's native 3D structure is currently believed to be determined only by its sequence and the local environment (i.e. the solvent). Recent genome sequencing efforts, such as the Human Genome Project, have caused an explosion in the number of known protein sequences. However, experimental methods for determining 3D structure, such as X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, are comparatively slow and expensive. Therefore, purely computational methods for determining a protein's 3D structure and function from sequence alone are highly desirable, and development of such methods will continue to be one of the major challenges for computational biologists in the coming postgenomic era.

Secondary structure prediction has long been viewed as a useful first step in prediction of 3D structure. Every computational method is parameterized, or trained, on a 'training set' of proteins with known secondary structure. The accuracy of the method is then estimated by testing the method on a different set of proteins, the 'test set'. Although the correct secondary structure of proteins in the test set must be known in order to judge the prediction accuracy, this information is withheld from the algorithm making the prediction. Assuming the proteins in the test set represent a statistically significant sample of the unknown proteins on which the algorithm might actually be used, this procedure gives a good estimate of how accurately an algorithm might perform on unknown proteins. By comparing the accuracy of various methods using several measures, it is easy to demonstrate that significant progress has been made in the field as the sizes of the training sets have increased and the algorithms themselves have become more complex.

Measures of accuracy

The most intuitive measure of secondary structure prediction accuracy, and therefore the most often quoted in describing the accuracy of various methods, is simply the percentage of residues in a protein or set of proteins for which the secondary structure is predicted correctly. Most methods attempt to distinguish the most commonly occurring types of secondary structure, α helices and extended β strands, from the rest of the protein, which is described as 'coil'. If every residue in a protein is predicted to be in one of three states (helix, strand or coil), the three-state prediction accuracy, or Q_3, is measured by simply dividing the number of correct predictions made by the total number of residues in the protein:

$$Q_3 = \frac{R_{\text{helix}} + R_{\text{strand}} + R_{\text{coil}}}{N}$$

Here, R_{strand} is the number of correctly predicted residues of type *strand*, and N is the total number of residues. There are several problems with judging algorithms only on the basis of Q_3. First, the Q_3 score does not account for differing success rates on different types of secondary structure. Also, the Q_3 score overemphasizes prediction of coil, the most common type of secondary structure. Because approximately 50% of most proteins are neither helix nor strand, a hypothetical algorithm that predicted every residue as coil would be 50% accurate (although 100% useless). To quantify an algorithm's performance on different types of secondary structure, researchers often calculate Matthews correlation coefficients for prediction of helix (C_{helix}), strand (C_{strand}) and coil (C_{coil}):

$$C_{\text{helix}} = \frac{(p_{\text{H}}\, n_{\text{H}}) - (u_{\text{H}}\, o_{\text{H}})}{\sqrt{(n_{\text{H}} + u_{\text{H}})(n_{\text{H}} + o_{\text{H}})(p_{\text{H}} + u_{\text{H}})(p_{\text{H}} + o_{\text{H}})}}$$

In this calculation, p_{H} is the number of correctly predicted helical residues, n_{H} is the number of residues that are correctly identified as something other than helix, o_{H} is the number of nonhelical residues that are predicted as helix, and u_{H} is the number of helical residues that are missed by the algorithm. A corresponding calculation is done for C_{strand} and C_{coil}. Matthews coefficients range from -1 for perfectly anticorrelated predictions to $+1$ for perfect predictions, and are close to 0 for random predictions (such as the above example).

A third quantitative measure of accuracy, segment overlap score (Sov), has recently been proposed as a standard for comparison of secondary structure prediction algorithms (Zemla *et al.*, 1999). The definition is fairly complex and beyond the scope of this article. Like Q_3, Sov can range from 0 to 100%. Compared to Q_3, the Sov score places emphasis on the ability of algorithms to correctly predict the core regions of secondary structure elements, penalizing algorithms that incorrectly break a single helix or strand into two or more separate elements of secondary structure. Therefore, secondary structure prediction methods that produce high Sov scores should be most useful to researchers for whom an accurate prediction of the exact number and sequential order of helices and strands in a protein is more important than a prediction of the precise locations of the boundaries of the elements. Several prediction methods published between 1994 and 1999 report a similar (but somewhat subjectively defined) measure proposed by the same research group, which was also called Sov. Since Sov scores for algorithms published to date have not been recalculated according to the new quantitative definition, further discussion of Sov will be omitted from this article. However, this measure may prove very useful in comparing future prediction methods.

Comparison of accuracy

A comparison of published accuracy statistics for several historical and current methods of secondary structure prediction is shown in **Table 1**. The test sets consist of 62 proteins for the three pre-1983 methods, and at least 124 nonhomologous proteins for the later methods. In most cases, the test sets contained no proteins that were highly homologous to the proteins used to parameterize the methods. There are three cases for which the training and test sets overlap; for these methods, the reported accuracies are approximately 1–5% higher than might be expected for newly discovered proteins. In one case, results are reported for a data set in which short helices and strands were redefined as coil; although this increases the apparent Q_3 of the method, correlation coefficients show the method to be slightly less accurate than similar methods. For methods with a publicly available web server or downloadable program, the name of the program is given. In some cases, statistics could not be computed from the information published.

Helix and Strand Propensities for Individual Amino Acids

In 1961, Anfinsen hypothesized that the conformation of proteins was determined primarily by their amino acid sequence. Shortly thereafter, theoreticians began attempts to predict protein structure from sequence. Early models in the mid-1960s classified amino acids as helix formers or helix breakers, based on the few known X-ray crystal structures (mostly of predominantly α-helical proteins

Table 1 Comparison of accuracy of various secondary structure prediction methods

Type	Method	Year	Q_3 (%)	C_{helix}	C_{strand}	C_{coil}
Statistical	Chou & Fasman	1974, 1978	49.9	0.22	0.22	0.25
	Lim [a]	1974	59.4	0.37	0.29	0.31
	Garnier *et al.* (GOR) [a]	1978	55.9	0.35	0.31	0.30
	King *et al.* (DSC)	1997	70.1	—	—	—
Nearest neighbour	Yi and Lander	1993	68.0	0.52	0.41	0.44
	Salamov and Solovyev (SSPAL)	1997	73.5	0.65	0.53	—
	SSPAL, single sequence input	1997	71.0	0.61	0.49	—
	Frishman and Argos (Predator)[b]	1997	74.8	0.61	0.45	0.44
Neural networks	Qian and Sejnowski	1988	64.3	0.41	0.31	0.41
	Holley and Karplus	1989	63.2	0.41	0.32	0.36
	Rost and Sander (PHD)	1993, 1994	72.2	0.63	0.53	0.52
	Chandonia and Karplus (Pred2ary)	1999	74.8	0.68	0.54	0.55
Combined	Cuff and Barton (JPred) [a]	1999	72.9	—	—	—
Homology modelling	Theoretical limit		88.0	—	—	—

Names of downloadable programs are given in parentheses.
[a]Cases in which the training and test sets overlap.
[b]Data set in which short helices and strands were redefined as coil.
—, Statistics could not be computed from the information published.

such as myoglobin) available at the time. Other methods involved using 'helical wheel' plots to search for sequences that could form amphipathic (charged or polar on one side, and hydrophobic on the other) α helices. Most methods reported at least 60–70% accuracy at helix prediction on test sets consisting of only a few proteins. The first attempts at β-sheet prediction were made in 1970, but these were largely unsuccessful.

In 1974, Chou and Fasman presented the first algorithm for predicting α-helix, β-strand and coil regions of globular proteins. The algorithm was parameterized using a set of 15 crystal structures available at the time. In total, the data set contained 2473 residues, of which 36% were α helix, 17% were β sheet and 47% were coil. Compared to current databases, which contain a representative cross-section of known protein structures (containing hundreds of times more data), the data set of Chou and Fasman slightly overrepresents helix and underrepresents strand; however, this data set was significantly larger and more diverse than those used to parameterize prior algorithms.

The Chou and Fasman method uses α-helix, β-strand and coil 'conformational parameters' derived for each amino acid. These are calculated by dividing the observed frequency of each amino acid in each type of secondary structure by the overall frequency of that secondary structure type in the data set. The α-helix conformational parameters are used to classify the 20 amino acids into six categories, ranging from 'strong helix formers' (Glu, Ala and Leu) to 'strong helix breakers' (Pro and Gly). Glu, the strongest helix former, is 1.53 times as likely to be present in a helix as the average amino acid; at the other end of the spectrum, Gly is only 0.53 times as likely to be present in a helix. The 20 amino acids are also classified into six categories according to their relative chances of being found in β sheet; these normalized probabilities range from 1.67 for Met to 0.26 for Glu. Chou and Fasman present 10 rules for helix and strand nucleation, extension and termination. After assigning regions of the sequences as α helices and β strands according to these rules, unassigned regions are predicted as coil. The rules are fairly complex, but easily implemented on a computer. Chou and Fasman claimed 80% (Q_3) accuracy for their algorithm (Chou and Fasman, 1974).

Other prediction algorithms invented in the mid-to-late 1970s included the method of Lim (Lim, 1974), and the Garnier, Osguthorpe and Robson (GOR) method (Garnier *et al*., 1978). The method of Lim is of special interest because it involves no numeric parameters. Instead, amino acids are classified into six categories according to size, hydrophobicity and conformational flexibility. Lim then presents several dozen rules for matching patterns of amino acid types with helices and strand formation, based on the current theory of hydrophobic packing in the core of proteins. Remaining segments of the chain are classified as coil. Lim demonstrated the predictions to be 70% (Q_3) accurate on the sequences he tested. The GOR method takes the opposite approach to that of Lim, attempting to use more information statistically derived from known structures in the prediction process. In the GOR method, secondary structure type at a given sequence position is correlated with the amino acid types at positions up to eight residues away in sequence. These statistics are used to calculate conformation preferences for α helix, β strand, β turn and coil at each position in a new sequence. Predictions are then made using only a few rules, which account for cooperativity in secondary structure formation and the relative frequencies of the four types of secondary structure in the database. GOR claimed 49% (Q_4) accuracy for the four-state prediction problem; although this is low compared with other algorithms that claimed 70–80% Q_3 accuracy, the four-state prediction problem is significantly more difficult.

Because of conflicting accuracy claims for various algorithms, Kabsch and Sander conducted a study in 1983 comparing the Chou and Fasman, GOR and Lim algorithms. The test set contained 62 proteins, the largest such set that had been assembled. Twenty-four of the structures were known prior to 1974, when the Chou and Fasman and Lim methods were published. The 38 post-1974 structures in the data set were therefore thought to be a better test of the algorithms' expected performance on newly discovered proteins. Results of the three methods on the entire test set are shown in **Table 1**. The method of Chou and Fasman performed with about 50% accuracy on both pre-1974 and post-1974 structures, significantly lower than the authors' original claim of 80% accuracy. The other two methods performed significantly better on the earlier test set, indicating possible overparameterization of the algorithms to fit existing data. Based on results on the post-1974 set, both the GOR and Lim methods could be expected to perform at about 56% accuracy on newly discovered sequences. This increase in accuracy in both algorithms relative to Chou and Fasman's technique was largely due to improvements in α-helix prediction, with more modest gains in β-strand prediction accuracy; however, even a 56% rate of successful predictions was significantly lower than the authors' prior claims. Kabsch and Sander concluded that 'an error rate of 44% is unacceptable for many purposes, and newly developing methods must do better' (Kabsch and Sander, 1983).

Use of Sequence Alignment to Improve Predictions

The use of larger training data sets and more complex statistical methods gradually improved prediction accuracy to over 60% by the late 1980s. A major increase in accuracy came in 1993, when Rost and Sander began using multiple sequence profiles to predict secondary structure. Rather than training and testing their method on single

sequences, Rost and Sander first constructed profiles of sequences by using local alignment algorithms to collect evolutionarily related sequences from current sequence databases. Although the majority of the sequences in each profile did not have known 3D structures, the authors hypothesized that the sequences that were highly similar to the sequence being predicted would also assume a similar fold. This additional information, such as the sequence conservation rate at each position and the range of possible amino acid types that might be accommodated at each position in the 3D structure, enabled a 7% increase in the accuracy of their algorithm, from 62% to 69% (Rost and Sander, 1994).

Prediction algorithms which rely only on simple statistics of amino acid frequencies have been updated to take advantage of multiple sequence profiles. The recently published DSC algorithm (King *et al.*, 1997) achieves 70.1% overall accuracy on a test set of 126 proteins which were not homologous to those used to parameterize the algorithm. Although this figure is lower than that for other contemporary algorithms, the authors claim an advantage in the simplicity of the method. The most accurate prediction methods to date are based on nearest neighbour methods and neural networks (discussed in the following two sections); because these methods employ complex nonlinear statistics, they have been criticized by some as 'black boxes', as the exact derivation of any individual prediction is often unclear.

Nearest Neighbour Methods

In the late 1980s, a new class of secondary structure prediction algorithm began to take advantage of the large number of 3D structures that were becoming available in public databases. These 'nearest neighbour' methods use sequence alignment techniques to search the database of available structures for segments of known proteins that are homologous to the query sequence. At each residue in the query sequence, a secondary structure prediction is made by choosing the secondary structure state (α helix, β strand or coil) held by the majority of the residues in homologous sequences of known structure.

An important element of nearest neighbour algorithms is the choice of sequence alignment algorithms used to find the neighbours. Early algorithms used standard 20×20 amino acid substitution matrices, such as the BLOSUM or PAM matrices (Dayhoff, 1978). In 1991, the environmental scoring method of Bowie and colleagues introduced substitution matrices based on amino acid type, secondary structure and solvent exposure of residues, significantly improving the accuracy of sequence alignment. These improved substitution matrices enabled the nearest neighbour method of Yi and Lander to achieve prediction accuracy of 68.0% on single sequences (Yi and Lander, 1993). Nearest neighbour methods are particularly useful in cases where structures of proteins highly homologous to the query sequence are present in the database, but unknown to the investigator using the method. In these cases, the homologous structure may be automatically identified through sequence alignment and heavily weighted in making predictions, resulting in accuracy levels comparable with homology modelling (85–90%).

In 1995, Salamov and Solovyev used multiple sequence profiles and a larger number of environmental classes (72, as opposed to Yi and Lander's 15) to improve the accuracy of nearest neighbour methods to 72.7%. An improved version of this method that uses a more advanced local sequence alignment algorithm produces the most accurate results of any nearest neighbour algorithm published to date, with a Q_3 of 73.5%. Nearest neighbour algorithms such as this one also perform well compared with statistical or neural network-based methods when single sequences rather than sequence profiles are used as input. The updated Salamov and Solovyev method achieves 71.0% accuracy in this case, the highest reported to date for single sequences (Salamov and Solovyev, 1997).

Another notable prediction algorithm that uses nearest neighbour methods is that of Frishman and Argos (1997). As in other, similar algorithms, homologous neighbours are identified using a local sequence alignment algorithm on a large database of sequences. Seven propensities for various types of secondary structure are then calculated for each position in the alignment; in addition to the standard three secondary structure categories, propensities for all helices (including α, 3_{10} and π helices), parallel and antiparallel β strands, and turns are also calculated. The propensities at each residue are then translated into predictions using a set of heuristic rules, similar to the original method of Chou and Fasman (Frishman and Argos, 1997). Although the authors claim 74.8% accuracy for the algorithm, this was accomplished by redefining the shortest helices and strands (which are the most difficult for all methods to predict reliably) in the data set as coil. When evaluated on this basis of Matthews correlation coefficients, the method of Frishman and Argos is less accurate than other recently developed nearest neighbour algorithms (**Table 1**).

Neural Networks

The neural network model is based on attempts to simulate computationally processes that take place in biological neural networks. The model consists of a set of units that represent neurons, and weighted connections between the units. Each unit sums the input from various other neurons and/or outside sources, processes the input with a sigmoidal 'activation function' that represents the activation threshold of biological neurons, and passes the

resulting output to other neurons to which it is connected. Computationally, the neural network can be thought of as a complex nonlinear function that maps a set of inputs (units with no input from other neurons, only from external sources) to a set of outputs. Well-established algorithms have been developed to 'train' a neural network to map a series of given input patterns to their desired output. The trained network can then be used to evaluate new data that become available. This pattern recognition ability of neural networks has been a topic of research in the field of artificial intelligence for many years.

In the late 1980s, several research groups trained neural networks to map patterns present in amino acid sequences to the correct secondary structure. Qian and Sejnowski developed a series of cascading networks that map single sequences to structure. The first network in the cascade is shown a window of 15 sequential residues as input, and trained to predict propensities for helix, strand and coil of the central residue in the window. The input window is 'slid' along the sequence, and centred on each residue until all are predicted. Output from the first network is fed into a second network (this time using a window of 17 residues), which refines the results of the first network and produces a final prediction. The overall accuracy (Q_3) of the algorithm is 64%, significantly better than statistical methods available at the time (Qian and Sejnowski, 1988). Holley and Karplus experimented with different network topologies and a variety of methods for encoding the amino acids in the input window; with only two outputs representing helix and strand propensities, their network produces predictions which are 63% accurate, and slightly better at predicting β strand than the networks developed by Qian and Sejnowski (Holley and Karplus, 1989).

In 1993, Rost and Sander developed a neural network algorithm which was trained and tested on profiles of multiple sequences. While this innovation was responsible for a 7% improvement in accuracy, other improvements included a two-network cascade similar to the method of Qian and Sejnowski, and the use of a 'jury' of seven different encoding/topological schemes similar to those tested by Holley and Karplus. The combination of improvements resulted in predictions that were 71.2% accurate, the highest accuracy available at the time. Minor improvements in the algorithm have since brought the accuracy to 72.2%, as shown in **Table 1** (Rost and Sander, 1994).

Recently, Chandonia and Karplus (1999) published a method that is similar to that of Rost and Sander but uses networks trained on much larger protein databases. Network topology is tuned to the increased information content of the larger training sets. The method also uses an iterative algorithm of predicting the structural class of a protein, then using that information to invoke neural networks specialized for predictions of each class of proteins. Accuracy of this method is the highest published to date, at 74.8% Q_3. Matthews correlation coefficients for helix prediction are over three times higher than for the original Chou and Fasman method, and correlation coefficients for strand and coil are improved more than twofold.

Another recently published method (Cuff and Barton, 1999) uses a jury of various prediction methods described here (Rost and Sander, 1994; Frishman and Argos, 1997; King *et al*., 1997; Salamov and Solovyev, 1997) to produce a combined prediction. Although the jury prediction is more accurate (72.9% Q_3) than any of the individual methods that were used, results must be interpreted with care, as several of the algorithms in the jury were trained on proteins homologous to those in the test set. However, the authors make the convincing point that the different methods of assigning secondary structure as helix, strand and coil used by the authors of each of the algorithms can affect the apparent accuracy. For example, some methods attempt to predict 3_{10} and π helices as well as α helices, while others predict β bridges as well as β strand. Different definitions of helix and strand may produce an apparent difference in accuracy between various algorithms of up to 3%. This means that all users of prediction methods should be aware of exactly what types of secondary structure their algorithms were designed to predict.

Neural networks have also been used to make secondary structure predictions other than the standard predictions of helix, strand and coil. Recently, a neural network method developed by Shepherd and colleagues was able to distinguish β turns from non-β-turn regions of proteins at 75% accuracy; the Matthews correlation coefficient for this prediction increased to 0.35, relative to around 0.20 for older β-turn prediction methods. The algorithm can also distinguish between common types of β turn with more limited accuracy (Shepherd *et al*., 1999).

Fold Recognition Algorithms

As more protein structures are discovered, the problem of protein structure prediction is increasingly being reduced to the problem of identifying the correct fold for a newly discovered sequence among the folds in the current structural database. If the correct fold for a protein can be identified, the problem of secondary structure prediction is largely solved; a 1994 study by Rost and Sander showed that, on average, the secondary structure in pairs of structurally homologous proteins are 88% identical. In practice, this means that highly accurate secondary structure predictions can currently be made for proteins which are at least 30% identical in sequence to a protein with known structure. Most newly discovered proteins still fall below this threshold, so conventional secondary structure prediction algorithms must be used.

Secondary structure predictions are also being used as a starting point for fold recognition algorithms. Several

studies in recent years have shown that alignment using a scoring scheme based on predicted secondary structure in combination with traditional scoring methods is more accurate than the traditional methods alone. For more information on such methods, see Fischer and Eisenberg (1996).

References

Chandonia JM and Karplus M (1999) New methods for accurate prediction of protein secondary structure. *Proteins* **35**: 293–306.

Chou PY and Fasman GD (1974) Prediction of protein conformation. *Biochemistry* **13**: 222–245.

Cuff JA and Barton GJ (1999) Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins* **34**: 508–519.

Dayhoff MO (1978) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3, pp. 353–358. Washington, DC: National Biomedical Research Foundation.

Fischer D and Eisenberg D (1996) Protein fold recognition using sequence-derived predictions. *Protein Science* **5**: 947–955.

Frishman D and Argos P (1997) Seventy-five percent accuracy in protein secondary structure prediction. *Proteins* **27**: 329–335.

Garnier J, Osguthorpe DJ and Robson B (1978) Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. *Journal of Molecular Biology* **120**: 97–120.

Holley H and Karplus M (1989) Protein secondary structure prediction with a neural network. *Methods in Enzymology* **202**: 204–224.

Kabsch W and Sander C (1983) How good are predictions of protein secondary structure? *FEBS Letters* **155**: 179–182.

King RD, Saqi M, Sayle R and Sternberg MJE (1997) DSC: public domain secondary structure prediction. *Computer Applications in the Biosciences* **134**: 473–474.

Lim VI (1974) Algorithms for prediction of Ã-helical and ã-structural regions in globular proteins. *Journal of Molecular Biology* **88**: 873–894.

Qian N and Sejnowski TJ (1988) Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology* **202**: 865–884.

Rost B and Sander C (1994) Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins* **19**: 55–72.

Salamov AA and Solovyev VV (1997) Protein secondary structure prediction using local alignments. *Journal of Molecular Biology* **268**: 31–36.

Shepherd AJ, Gorse D and Thornton JM (1999) Prediction of the location and type of β-turns in proteins using neural networks. *Protein Science* **8**: 1045–1055.

Yi TM and Lander ES (1993) Protein secondary structure prediction using nearest neighbor methods. *Journal of Molecular Biology* **232**: 1117–1129.

Zemla A, Venclovas C, Fidelis K and Rost B (1999) A modified definition of Sov, a segment-based measure for protein secondary structure prediction assessment. *Proteins* **34**: 220–223.

Further Reading

Barton GJ (1995) Protein secondary structure prediction. *Current Opinion in Structural Biology* **5**: 372–376.

Bohm G (1996) New approaches in molecular structure prediction. *Biophysical Chemistry* **59**: 1–32.

Bowie JU, Luthy R and Eisenberg D (1991) A method to identify protein sequences that fold into a known three-dimensional structure. *Science* **253**: 164–170.

Fischer D and Eisenberg D (1996) Protein fold recognition using sequence-derived predictions. *Protein Science* **5**: 947–955.

Kabsch W and Sander C (1983) How good are predictions of protein secondary structure? *FEBS Letters* **155**: 179–182.

Rost B and O'Donoghue S (1997) Sisyphus and prediction of protein structure. *Computer Applications in the Biosciences* **13**: 345–356.