Protein secondary structure prediction

Geoffrey J Barton

University of Oxford, Oxford, UK

The past year has seen a consolidation of protein secondary structure prediction methods. The advantages of prediction from an aligned family of proteins have been highlighted by several accurate predictions made 'blind', before any X-ray or NMR structure was known for the family. New techniques that apply machine learning and discriminant analysis show promise as alternatives to neural networks.

Current Opinion in Structural Biology 1995, 5:372–376

Introduction

By far the most accurate method of predicting the secondary (and tertiary) structure of a globular protein is by alignment of the sequence to a homologue of known three-dimensional structure. If the sequence similarity is high enough to ensure reliable alignment, then the mean accuracy of prediction of three states (helix, strand or coil) is 88% [1••]. In contrast, until recently, the accuracy of prediction without homology was around 50–60%. With the rapid growth in protein sequence data available and the requirement to infer protein conformation and function from sequence, a pressing need has developed for more accurate methods for protein structure prediction. Fortunately, the expansion in the databases has also meant that it is now common for predictions to be made for proteins that are members of large families of sequences. When accurate multiple alignments of the members of these families can be made, the evolutionary information present in the alignment can be used to identify which residues are of key importance to the fold and function of the protein. Effective exploitation of this information has led to a significant increase in the accuracy of secondary structure prediction methods so that today it is often possible to identify correctly the majority of the secondary structure elements in a protein. Here I will discuss work on protein secondary structure prediction published in this area between late 1993 and February 1995.

The rationale behind prediction methods

Perhaps the most widely used secondary structure prediction method over the past 15 years is that developed by Garnier, Osguthorpe and Robson (known as the GOR method [2,3•]). Although Robson and colleagues recognized the importance of evolutionary information from protein families [2], in the late 1970s studies of aligned families were hampered by lack of data, so the idea could not be exploited fully. In the 1980s Zvelebil *et al.* [4] capitalized on new techniques for multiple sequence alignment [5] and the increased database, to extend the GOR method to use aligned sequence data. This gave an increase in accuracy of 9%, but the available data still limited testing to 11 families. Over the past five years, with an increasing number of suitable protein families to work with, several groups have devised secondary structure prediction methods that exploit multiple sequence alignments [6,7•,8–10]. Of necessity, many methods are developed and tested on proteins of known tertiary structure. If the test is to be objective, the prediction method must be automatic and have parameters that can be optimized against a training set of proteins. Testing is on a group of proteins that are not homologous to the training set. The most exciting test, however, is to predict the structures of proteins for which no experimental structure is known. When making such 'blind' predictions it is possible to bring additional information from spectroscopy or other experimental techniques to bear on the prediction. Many such predictions have now been made and evaluated in the light of the subsequently determined protein structures [11••]. The results have been very encouraging, with predictions that correctly locate the position and type of most of the core secondary structures in the protein [11••,12].

During the review period blind predictions of tyrosine phosphatases [13••], factor XIIIa [13••], isopenicillin N synthase [14••], the pleckstrin homology (PH) domain [15•,16] and matrix metalloproteinases [17•] have been published and can now be compared with experimentally determined structures (see annotations [13••,14••,15•,16,17•]). Predictions for serine/threonine phosphatases [18], isoprenyl diphosphate synthases [19], von Willebrand factor type A domain [20•], integrin α-subunit N-terminal domain [21], prion protein [22•] and the proteasome [23••] await testing by comparison with structures yet to be determined by X-ray or NMR methods.

Improving secondary structure predictions

Accuracy

One problem that has arisen is how to evaluate secondary structure predictions. For prediction of a single protein sequence one might expect the best residue by residue accuracy to be 100%. It is not possible to define the secondary structure of a protein exactly, however. There is always room for alternative interpretations of where a helix or strand begins or ends so failure of a prediction to match exactly the secondary structure definition is not a disaster [24•]. The problem of evaluation is more complicated for prediction from multiple sequences, as the prediction is a consensus for the family and so is not expected to be 100% in agreement with any single family member. The expected range in accuracy for a perfect consensus prediction is a function of the number, diversity and length of the sequences. Russell and I have calculated estimates of this range [11••].

Simple residue by residue percentage accuracy has long been the standard method of assessment of secondary structure predictions. Although a useful guide, high percentage accuracies can be obtained for predictions of structures that are unlike proteins. For example, predicting myoglobin to be entirely helical (no strand or coil) will give over 80% accuracy but the prediction is of little practical use. Rost *et al.* [25••] and Wang [26] explore these problems and suggest some alternative measures of predictive success based on secondary structure segment overlap. Although such measures help in an objective assessment of the prediction, there is no complete substitute for visual inspection. By eye, serious errors stand out and predictions of structures that are unlike proteins are usually recognizable. By eye, it is also straightforward to weight the importance of individual secondary structures. For example, prediction of what is in fact a core strand to be a helix would seriously hamper attempts to generate the correct tertiary structure of the protein from the predicted secondary structure, whereas prediction of a non-core helix as coil may have little impact on the integrity of the tertiary structure.

'Predictable' regions of secondary structure

When recent predictions are examined in the light of the corresponding experimentally determined structures, the results look good. In general, the regions predicted with the highest confidence measure are also the most accurate. For example, Livingstone and I [13••] assigned 41% of the tyrosine phosphatase structure with high confidence. Within these regions 88% of the residues were correctly predicted. Interestingly, these figures agree with Rost and Sander's observation that 40% of a sequence will be predicted with >88% accuracy by their method [1••]. This agreement suggests that there is a core of 'predictable' regions in a protein. Examination of six blind predictions shows that the most accurately predicted regions are those that have clear periodicity in conservation, where conserved positions either alternate (β-strand) or have a 1, 4, 5, 8 pattern characteristic of one face of an α-helix (CD Livingstone, personal communication). Problems remain with buried α-helices that comprise short runs of conserved hydrophobic amino acids. These often look like potential β-strands and can mislead both automatic and manual predictive methods.

Evolutionarily conserved residues and prediction

The improvements in the accuracy of secondary structure prediction that are seen when multiple alignments are used stem from the observation that positions in an alignment where the identity of the amino acid residue varies slowly during the course of evolution are important to the stability of the fold or the protein function. Patterns of conservation can be discerned by eye, but ideally automatic protocols should be used to improve objectivity. Over the past year Benner and colleagues [27•] have described their heuristics for the prediction of secondary structure and exposure of amino acid residues to solvent. Rost and Sander [1••] have updated their automatic neural network prediction method to include explicit gaps and conservation while Blundell's group [28•–30•] have explained their application of environment-dependent substitution tables to the prediction and orientation of α-helices [28•], and to more general secondary structure prediction [29•,30•]. The study of substitution tables is particularly appealing as it has a direct relationship to the underlying evolutionary processes that lead to change in proteins.

Prediction from single sequences

Although the emphasis in the review period has been on the development and application of multiple sequence data to prediction, some new methods for prediction from single sequences have been described [31•,32••,33,34••,35•]. Of these, the work of Solovyev and Salamov [32••] is particularly promising. Unlike most prediction schemes, which predict at the level of single residues, they apply linear discriminant analysis to assign segments of secondary structure. The overall accuracy when subjected to a full jackknife test, in which each protein is removed from the training set in turn, is a respectable 65%, with especially high accuracy for long α-helices and β-strands.

Structural class prediction

If the structural class of a protein (α, β, α/β, or α+β) is known then the secondary structure prediction problem is simplified. Clearly, if it is known that a protein contains only β-strands, then there is no need to consider α-helix! Spectroscopic data can be used to good effect to infer the structural class. Such data were used for the prion protein [22•] and von Willebrand factor predictions [20•]

but ideally one would like to predict the structural class from sequence data alone. Rost and Sander [1••] showed that their neural network method correctly identifies the structural class of 75% of the proteins of a 250-member set, and Chou and Zhang [36•], using a new approach that calculates Mahalanobis distances in amino acid composition space, claim a remarkable 94.7% accuracy for the structural classification of 131 proteins and 100% accuracy for classification of α- and β-class proteins.

Conclusions

The review period has seen consolidation of techniques to predict protein secondary structure from multiply aligned sequences. Predictions by both automatic and semi-automatic methods now reach accuracies of around 70% for three-state assignments (helix, β-strand and coil) and over 80% for regions assigned with high confidence. Assessment of predictions made blind suggests that the regions most accurately predicted are those that exhibit characteristic patterns of conservation for α-helix and β-strand. Reliable predictions are also made for variable loops as these regions tolerate insertions and deletions that can be identified easily from a multiple alignment.

Current secondary structure predictions are good enough to provide a starting point for tertiary structure prediction or for searching libraries of known structures to find topologies consistent with the secondary structure and other restraints [37••]. Such techniques complement methods of fold recognition based either on threading (for a review, see [38]) or on the recognition of distant sequence similarities [39]. Secondary structure prediction is not yet perfect, but is now accurate enough to be taken seriously as a tool to assist in the design of experiments to probe protein structure and function.

References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- • of special interest
- •• of outstanding interest

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

The latest in a long line of papers describing Rost and Sander's neural network method for secondary structure prediction from aligned sequences. The method is reviewed and recent developments that include position dependent conservation weights and explicit consideration of gaps are described. This paper also shows that the accuracy of prediction of protein structural class by the algorithm is similar to that obtained from circular dichroism spectra.

2. Garnier J, Osguthorpe DJ, Robson B: **Analysis of the accuracy and implication of simple methods for predicting the secondary structure of globular proteins.** *J Mol Biol* 1978, **120**:97–120.

3. • Ellis LBM, Milius RP: **Valid and invalid implementations of GOR secondary structure predictions.** *Comput Appl Biosci* 1994, **10**:341–348.

The Garnier, Osguthorpe and Robson (GOR) algorithm for protein secondary structure prediction has been one of the most widely used methods since its description in 1978. This paper points out that many of the available implementations of the original GOR algorithm are incorrect and provides some simple test data to check a GOR program. Correct implementations give accuracies up to 13 percentage points higher than incorrect ones.

4. Zvelebil MJJM, Barton GJ, Taylor WR, Sternberg MJE: **Prediction of protein secondary structure and active sites using the alignment of homologous sequences.** *J Mol Biol* 1987, **195**:957–961.

5. Barton GJ, Sternberg MJE: **A strategy for the rapid multiple alignment of protein sequences: confidence levels from tertiary structure comparisons.** *J Mol Biol* 1987, **198**:327–337.

6. Rost B, Sander C: **Prediction of protein secondary structure at better than 70 percent accuracy.** *J Mol Biol* 1993, **232**:584–599.

7. • Rost B, Sander C, Schneider R: **PHD — an automatic mail server for protein secondary structure prediction.** *Comput Appl Biosci* 1994, **10**:53–60.

One of the strengths of the Rost and Sander prediction method is that anyone with an email account can use it to predict secondary structure. This paper describes how the email server is implemented and provides some useful guidelines on how to make best use of the server.

8. Benner SA: **Patterns of divergence in homologous proteins as indicators of tertiary and quaternary structure.** *Adv Enzyme Regul* 1989, **28**:219–236.

9. Barton GJ, Freemont PF, Newman R, Crumpton M: **Sequence analysis of the annexin super gene family of proteins.** *Eur J Biochem* 1991, **198**:749–760.

10. Levin JM, Pascarella S, Argos P, Garnier J: **Quantification of secondary structure prediction improvement using multiple alignments.** *Protein Eng* 1993, **6**:849–854.

11. •• Russell RB, Barton GJ: **The limits of protein secondary structure prediction accuracy from multiple sequence alignment.** *J Mol Biol* 1993, **234**:951–957.

A prediction of the consensus secondary structure for a family of sequences is unlikely to agree 100% with every member of the family. This paper determines the scatter in accuracy expected for secondary structure predictions of families composed of sequences of varying lengths and similarity. The paper summarizes the results of blind predictions performed up to 1993.

12. Benner SA, Gerloff DL, Jenny TF: **Predicting protein crystal-structures.** *Science* 1994, **265**:1642–1644.

13. •• Livingstone CD, Barton GJ: **Secondary structure prediction from multiple sequence data — blood-clotting factor-XIII and *Yersinia* protein-tyrosine-phosphatase.** *Int J Pept Protein Res* 1994, **44**:239–244.

Blind predictions made for two protein families for which X-ray structures are now known are reported. Conventional single-sequence secondary-structure prediction methods were combined with an analysis of residue conservation patterns. Factor XIIIa is a difficult challenge for prediction because it is a large (674 residue) multidomain protein, and comparatively few homologous sequences are available. The overall accuracy for this prediction was 61%, but the 32% of the structure that we predicted with confidence showed 83% accuracy. The tyrosine phosphatase prediction fared better, with 14 of the 17 regular secondary structures being correctly located and an overall accuracy of 76% when compared to the refined 2.3 Å structure for the human enzyme. Note: the appendix to this paper includes a comparison of the prediction to an early 2.8 Å structure of the human tyrosine phosphatase. The secondary structures were not as well resolved in this structure and the agreement with the prediction was 68%.

14. •• Benner SA, Jenny TF, Cohen MA, Gonnet GH: **Predicting the conformation of proteins from sequences. Progress and future progress.** *Adv Enzyme Regul* 1994, **34**:269–353.

An appendix to this paper describes (in approximately 44 pages) predictions for isopenicillin N synthase and related proteins. The X-ray structure of a member of this family has recently been determined. The preferred consensus secondary structure prediction for the family showed 10 secondary structures with confidence, of which eight were correct

in type and position and two α-helices were incorrectly predicted as β-strand. Of the 12 more problematic regions predicted, nine were correct in type, two were incorrectly predicted to be β-strand where there is in fact coil and one region was predicted to be α-helical when the actual secondary structure was a β-strand followed by an α-helix. Three β-strands (one of which is a core secondary structure) and two short α-helices were incorrectly predicted as coil.

15. • Musacchio A, Gibson T, Rice P, Thompson J, Saraste M: **The PH domain: a common piece in the structural patchwork of signalling proteins.** *Trends Biochem Sci* 1993, **18**:343–348.

This article includes a prediction of the consensus secondary structure of 45 pleckstrin homology (PH) domains. The prediction is very similar to that obtained by Jenny and Benner [16] and agrees well with the experimentally determined structure, which contains seven β-strands and a C-terminal α-helix. Both predictions miss a second α-helix between β-strands 3 and 4, but this α-helix is probably not present in all members of the family.

16. Jenny TF, Benner SA: **A prediction of the secondary structure of the pleckstrin homology domain.** *Proteins* 1994, **20**:1–3.

17. • Hodgkin EE, Gilman IC, Gilbert RJ: **Retrospective analysis of a secondary structure prediction — the catalytic domain of matrix metalloproteinases.** *Protein Sci* 1994, **3**:984–986.

This secondary structure prediction was performed before the first tertiary structure of a member of the family was known. Unfortunately the paper did not reach print before the first structure was published. This short note compares the blind prediction with the crystal structure and shows that all five β-strands and three α-helices were correctly predicted, but one additional β-strand was predicted that is not seen in the structure.

18. Barton GJ, Cohen PTW, Barford D: **Conservation analysis and structure prediction of the protein serine/threonine phosphatases — sequence similarity with diadenosine tetraphosphatase from *Escherichia coli* suggests homology to the protein phosphatases.** *Eur J Biochem* 1994, **220**:225–237.

19. Chen A, Kroon PA, Poulter CD: **Isoprenyl diphosphate synthases: protein sequence comparisons, a phylogenetic tree, and predictions of secondary structure.** *Protein Sci* 1994, **3**:600–607.

20. • Perkins SJ, Smith KF, Williams SC, Haris PI, Chapman D, Sim RB: **The secondary structure of the von-Willebrand-factor type-A domain in factor-b of human-complement by Fourier-transform infrared-spectroscopy — its occurrence in collagen type-VI, type-VII, type-XII and type-XIV, the integrins and other proteins by averaged structure predictions.** *J Mol Biol* 1994, **238**:104–119.

In this paper analysis of the secondary structure content of a von Willebrand factor type A domain obtained by Fourier transform infrared spectroscopy is used to predict occurrence of the domain in other proteins, by using the averaged GOR (Garnier, Osguthorpe and Robson) algorithm for predictions on 75 sequences. The X-ray structure of a von Willebrand domain is expected to be determined this year, so this prediction should be evaluated shortly.

21. Tuckwell DS, Humphries MJ, Brass A: **A secondary structure model of the integrin α-subunit N-terminal domain based on analysis of multiple alignments.** *Cell Adhes Commun* 1994, **2**:385–402.

22. • Huang ZW, Gabriel JM, Baldwin MA, Fletterick RJ, Prusiner SB, Cohen FE: **Proposed 3-dimensional structure for the cellular prion protein.** *Proc Natl Acad Sci USA* 1994, **91**:7139–7143.

The authors apply a variety of secondary structure prediction methods to a family of prion sequences. In combination with spectroscopic data the prediction methods suggest that the protein has a four-helix core. Application of combinatorial packing algorithms suggest alternative four-helix bundle models for the prion protein.

23. •• Lupas A, Koster AJ, Walz J, Baumeister W: **Predicted secondary structure of the 20-S proteasome and model structure of the putative peptide channel.** *FEBS Lett* 1994, **354**:45–49.

In this paper most current secondary structure prediction algorithms are applied to a multiple alignment of proteasome sequences and the results combined into a consensus prediction. A three-dimensional model of the N-terminal region based on the prediction is proposed and shown to fit data from electron microscopy studies. An experimentally determined tertiary structure for the proteasome complex should be available this year from the same group.

24. • Jenny TF, Benner SA: **Evaluating predictions of secondary structure in proteins.** *Biochem Biophys Res Commun* 1994, **200**:149–155.

The authors discuss a variety of issues and highlight the point that definitions of secondary structures vary. Thus, no consensus prediction will match an individual protein, so accuracy will never be 100%.

25. •• Rost B, Sander C, Schneider R: **Redefining the goals of protein secondary structure prediction.** *J Mol Biol* 1994, **235**:13–26.

Residue by residue accuracies do not give a strong indication of whether the pattern of secondary structures predicted agrees well with reality. This paper describes a variety of 'segment' scoring schemes that attempt to get around this problem.

26. Wang ZX: **Assessing the accuracy of protein secondary structure.** *Nature Struct Biol* 1994, **3**:145–146.

27. • Benner SA, Badcoe I, Cohen MA, Gerloff DL: **Bona-fide prediction of aspects of protein conformation — assigning interior and surface residues from patterns of variation and conservation in homologous protein sequences.** *J Mol Biol* 1994, **235**:926–958.

A lengthy paper that details the approach taken by Benner and colleagues for analyzing families of proteins to predict secondary structure and accessibility of amino acid residues to solvent. The proposed heuristics are evaluated by application to seven protein families.

28. • Donnelly D, Overington JP, Blundell TL: **The prediction and orientation of α-helices from sequence alignments — the combined use of environment-dependent substitution tables, Fourier-transform methods and helix capping rules.** *Protein Eng* 1994, **7**:645–653.

The authors describe the specific application of amino acid substitution tables to α-helix prediction (see also [29•,30•]). Predictions for several all α-helical proteins are made. The method also suggests which face of the helix should pack in the core of the protein.

29. • Wako H, Blundell TL: **Use of amino-acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. 1. Solvent accessibility classes.** *J Mol Biol* 1994, **238**:682–692.

The authors describe a procedure to predict classes of solvent accessibility of amino acid residues from aligned protein sequences. The assumption is that a given residue will have a different pattern of substitution during the course of evolution if buried when compared to the same residue in an exposed enviroment. This paper builds on the extensive work by Blundell and colleagues on deriving substitution tables for amino acids in many different structural environments. The method gives 77% accuracy, but unfortunately this figure is based on predictions for only 13 protein families. See also [28•,30•].

30. • Wako H, Blundell TL: **Use of amino-acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. 2. Secondary structures.** *J Mol Biol* 1994, **238**:693–708.

An extension of the accessibility class predictions described in [29•] to a full secondary structure prediction method. The method is automated but the reasoning that leads to a prediction can be extracted. The method gives a mean three state (helix, sheet and coil) accuracy of 69%, but again this figure comes from predictions on only 13 families.

31. • Geourjon C, Deleage G: **SOPM — a self-optimized method for protein secondary structure prediction.** *Protein Eng* 1994, **7**:157–164.

This novel prediction method is based on sequence similarity. A protein is compared with a database of proteins of known structure and the subset of most similar proteins selected. The parameters are then optimizsed to give the best prediction of the secondary structure of the proteins in this subset and these parameters are applied to predicting the secondary structure of the protein of interest. The method gives 69% accuracy on a database of 239 protein chains; however, the database contains many similar protein sequences, some with pairwise identities of >30%. It will be interesting to see how the method performs on a non-homologous dataset.

32. •• Solovyev V, Salamov AA: **Predicting α-helix and β-strand segments of globular proteins.** *Comput Appl Biosci* 1994, **10**:661–669.

This unique method applies linear discriminant analysis to the secondary structure prediction problem. Rather than initially predicting the conformation of single residues, the method aims to predict segments

as either α-helix or β-strand. The overall three state (helix, sheet and coil) accuracy of this method (without multiple alignments) on the same dataset used by Rost and Sander [6] is 65.1%. A simple extension to multiple alignments raises the accuracy to 68.2%. This paper also gives a nice summary of alternative methods to evaluate predictions and includes a table that compares the results of many current methods.

33. Wintjens RT, Rooman MJ, Wodak SJ: **Identification of short turn motifs in proteins using sequence and structure fingerprints.** *Isr J Chem* 1994, **34**:257–269.

34. •• Sternberg MJE, King RD, Lewis RA, Muggleton S: **Application of machine learning to structural molecular-biology.** *Philos Trans R Soc Lond [Biol]* 1994, **344**:365–371.

Although neural network techniques have been applied very successfully to secondary structure prediction, it is difficult to extract explanations from a network of why a particular residue has been assigned to the given class of secondary structure. In contrast, the machine learning techniques discussed in this article seek to derive rules automatically from a set of specific observations (e.g. the location of hydrophobic amino acids). Such rules provide understandable explanations of a particular prediction. Applications to drug design are also discussed.

35. • Zimmermann K: **When awaiting bio-Champollion — dynamic-programming regularization of the protein secondary structure predictions.** *Protein Eng* 1994, **7**:1197–1202.

A systematic procedure to make secondary structure predictions reflect more accurately actual proteins is described. This systematically cleans up predictions by any probabilistic method to remove predictions of unrealistically short β-strands and α-helices.

36. • Chou KC, Zhang CT: **Predicting protein-folding types by distance functions that make allowances for amino-acid interactions.** *J Biol Chem* 1994, **269**:22014–22020.

The amino acid composition of a protein can be represented in 20 dimensional space where each dimension represents the frequency of occurrence of an amino acid. The distance between proteins in this compositional space gives a measure of their similarity. There are many different ways of calculating distances, the simplest being the linear distance that is familiar from everyday life; however, this does not take into account scatter in the data, or correlations (e.g. a protein containing many alanines may contain few aspartates). Here, distance measurements that take correlations into account are used. High accuracy for class prediction is claimed, but the testing set contains homologous proteins and it is not clear if a full jackknife test has been performed.

37. •• Russell RB, Copley RR, Barton GJ: **Protein fold recognition from secondary structure assignments.** *In IEEE Proceedings of the 28th Annual Hawaii International Conference on System Sciences*, vol 5. Edited by Hunter L, Shriver BD. Los Alamitos: Institute for Electrical and Electronic Engineers' Press; 1995:302–311.

A preliminary account is given of a new technique to identify potential folds for a protein given the secondary structure and restraints. The method searches a database of the three-dimensional structure of domains to find domains that are consistent with the secondary structure prediction and any distance restraints. The technique allows for deletions of complete secondary structures from either the query protein or the database and so may find similarities between proteins that share little sequence similarity.

38. Rooman MJ, Wodak SJ: **Generating and testing protein folds.** *Curr Opin Struct Biol* 1993, **3**:247–259.

39. Taylor WR: **Protein-structure modeling from remote sequence similarity.** *J Biotechnol* 1994, **35**:281–291.

GJ Barton, Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK.

E-mail: geoff@biop.ox.ac.uk