

Secondary structure prediction of all-helical proteins in two states

Burkhard Rost and Chris Sander

Protein Design Group, EMBL, D-6900 Heidelberg, Germany

Can secondary structure prediction be improved by prediction rules that focus on a particular structural class of proteins? To help answer this question, we have assessed the accuracy of prediction for all-helical proteins, using two conceptually different methods and two levels of description. An overall two-state single-residue accuracy of ~80% can be obtained by a neural network, no matter whether it is trained on two states (helix and non-helix) or first trained on three states (helix, strand and loop) and then evaluated on two states. For four test proteins, this is similar to the accuracy obtained with inductive logic programming. We conclude that on the level of secondary structure, there is no practical advantage in training on two states, especially given the added margin of error in identifying the structural class of a protein. In the further development of these methods, it is increasingly important to focus on aspects of secondary structure that aid in the construction of a correct 3-D model, such as the correct placement of segments.

Key words: all-helical proteins/logic programming/secondary structure prediction

Introduction

Classically, the problem of secondary structure prediction is formulated in terms of three states: helix, strand and loop (Nagano, 1973; Chou and Fasman, 1974; Lim, 1974; Garnier *et al.*, 1978). It has been argued that more accurate predictions are possible within a particular structural class (Taylor and Thornton, 1984; King and Sternberg, 1990; Kneller *et al.*, 1990; Hayward and Collins, 1992; Presnell *et al.*, 1992), using methods specifically derived for, e.g. the class of all-helical proteins. This idea leads to a two-step procedure: first identify a protein as all-helical, then predict where the helices and loops are located. In this vein, a recent report of 80.5% accuracy for predicting the two states, helix and non-helix, is particularly exciting (Muggleton *et al.*, 1992). The method uses inductive logic programming, a machine learning method that extracts rules from a training set of 12 all-helical proteins and then is tested on four non-homologous all-helical proteins. To learn more about the possible practical advantages of this type of method, we have compared the results of Muggleton, King and Sternberg (Muggleton *et al.*, 1992) referred to here as MKS, with two neural network systems, evaluated on the same test of four proteins.

The first network system, dubbed 'helix net' (a neural network trained explicitly to predict the secondary structure of all-helical proteins in two states), allows a direct comparison of methodology by using the same training set of 12 proteins, in terms of the same two states, helix and non-helix. It is different in that it uses a neural network in training, rather than inductive logic programming and multiple sequence alignments as input, rather

than single sequences. The helix net achieves 82.7% two-state accuracy (Q_2) on the four test proteins.

The second network system, profile neural network prediction from Heidelberg (PHD; a system of neural networks predicting secondary structure in three states available for fully automatic use), provides a comparison of data representation by training on the three helix states, helix, strand and loop (rather than on two states), on a much larger database of 130 proteins (rather than on 12 proteins) and with multifold cross-validation of results, i.e. on multiple pairs of training and test sets (rather than only a single pair). The PHD net achieves 81.2% accuracy (Q_2) on the four test proteins when evaluated as to its ability to predict helix/non-helix.

These test data shed some light on the following issues: (i) does the inductive logic programming method presented by MKS yield the best prediction for all-helical proteins? (ii) is there an advantage in using two-state predictions rather than three-state predictions? (iii) can 80% accuracy of two-state prediction be expected in future practice? (iv) is it reasonable to compare methods scoring as high as 80% on the basis of single-residue scores? After discussion of these issues, we make some practical recommendations.

Table 1. Accuracy of three different prediction methods evaluated in the two states, helical (H or α) and non-helical (L)

	Number of residues			Accuracy	
	Pred H	Pred L	Sum	% obs	% pred
MKS					
Obs H	160	57	217	74	86
Obs L	24	175	199	87	75
Sum	184	232	416	80.5	
Helix net					
Obs H	162	46	208	77	85
Obs L	27	184	211	87	80
Obs H + L	189	230	419	82.7	
PHD					
Obs H	132	71	203	65	94
Obs L	7	204	211	96	74
Obs H + L	139	275	414	81.2	

Accuracy is derived from the number of residues predicted in structure i (pred) and observed in structure j (obs). For example, for MKS, out of 217 residues observed as helical, 160 residues are predicted as H and 57 are predicted as L. Accuracy is expressed as the correctly predicted fraction of residues observed (% obs) or predicted (% pred) in a particular state, given in separate rows for H, L or all residues (H + L). MKS refers to Muggleton, King and Sternberg (Muggleton *et al.*, 1992), helix net to the profile network trained on 12 proteins to distinguish the two states H and L and PHD to a profile network trained on 130 proteins to distinguish three states, H, E and L of secondary structure. All three methods have a two-state accuracy Q_2 just above 80% evaluated on the four test proteins of MKS.

351c

.....1.....2.....3.....4.....5.....6.....7.....8..

sequence | EDPEVLFKNKGCVACHAIDTKMVGPAVKDVAAKFACQAGAEELAQRIRKNGSQGVWGPIMPFPNAVSDDEAOTLAKWVLSQK |

DSSP | KHHHHHK GGG HHHHHHHH HHHHHHHHHH HHHHHHHHHHH |

MKS | HHHHH HHHHHHHHHHHHHHHH HHHHHHHHHHHH |

helix net | HHHHHHHH HHHHHHHH HHHHHHHH HHHHHHHHHHHH |

PHD | HHHHH HHHHHHHHHH HHHHHHHHHH HHHHHHHHHHHH |

9pap_A

.....1.....2.....3.....4.....5.....6.....7.....8

sequence | IPEYVDWRQKGAVTPVKNQSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQLLDCDRRSYGCNGGYPWSALQLVAQYGI |

DSSP | HHHHHHHHHHHHHHHHHH HHHHHH HHHHHHHHHH |

MKS | HHHHHHHHHHHHHHHHHHHH HHHHHH HHHH HHHHHHHHHH |

helix net | HHHH HHHHHHHHHH HHHHHHHHHH HHHHHHHH |

PHD | HHHH HHHHHH HHHHHHHHHH |

.....1.....2.....3.....4.....5.....6.....7.....8

sequence | HYRNTYPYEGVQRYCRSREKGPYAAKTD |

DSSP | GGG |

MKS | |

helix net | H |

PHD | |

Fig. 1. Predictions of four test proteins. The predictions come from three methods: the machine learning rules of Muggleton, King and Sternberg (MKS) (Muggleton *et al.*, 1992), the network trained on helical proteins with (only) single cross-validation (helix net) and the network trained to predict three states (helix, strand and loop) on 130 non-similar proteins (PHD) and multiple cross-validation on seven different pairs of training/test sets. The second row gives the observed structure as extracted from the 3-D coordinates by DSSP: H, helix; ., loop; G, 3₁₀ helix (evaluated here as non-helix). For two proteins, helix net and/or PHD use a slightly different homologue than MKS: cytochrome 256b__A instead of cytochrome 156b__A and phospholipase 4bp2 instead of phospholipase 1bp2. The small differences are indicated by the following symbols: * means that the other structure has an H, h that the other one has a loop and . that there is an insertion in the sequence.

Inductive logic programming versus neural network approaches

Comparable performance is achieved with either the rules derived from the inductive logic program Golem (used by Muggleton *et al.* (1992)) using single sequences as input or with the parameters derived from a neural network system using profiles of evolutionary information as input. The overall accuracy on the four test proteins is $Q_2 = 80.5\%$ for Golem and 82.7% for helix net (Table I, Figure 1). It is difficult to assess the relative merits of the two approaches, as they use different inputs. The Prolog-coded rules resulting from inductive logic programming may lead to new physical insight into protein structure formation (King *et al.*, 1992), but the same might be true for networks

(Brunak, 1991). Thus, whether inductive logic programming has a conceptual advantage or disadvantage relative to neural nets in protein structure prediction or vice versa, remains an open question.

Two-state prediction versus three-state prediction

Once the structural type of a protein has been determined as all-helical, e.g. by CD spectroscopy, secondary structure prediction can be reduced to the simpler task of predicting the location of the secondary structure segments, rather than the type (helix or strand) and the location. The rules of such two-state predictions from sequences should be simpler and easier to learn. Does this hold up in practice?

Table II. Evaluation of prediction accuracy in terms of the two states, helix (α) and non-helix (L), for a random method (RAN) and for a neural network method (PHD) applied to different sets of proteins

	N	N _{prot}	% obs		% pred		Corr _α	Corr _L	
			Q ₂	Q _α	Q _L	Q _α			Q _L
RAN	12162	94	54.5	24	68	24	68	-0.09	-0.09
PHD on 22	3221	22	75.9	69	88	90	62	0.55	0.55
PHD on six	761	6	84.2	87	78	89	74	0.64	0.64
PHD on four	414	4	81.2	65	96	94	74	0.65	0.65
PHD on IL-4	130	1	87.7	86	89	88	86	0.75	0.75

RAN: alignments of protein pairs with 5–10% pairwise sequence similarity over an alignment length of more than 80 residues were chosen at random from an all-against-all sequence comparison of a representative set of proteins of known structure (Hobohm *et al.*, 1992). Thus, these pairs are very likely to be dissimilar in their 3-D structure. In each pair, the structure of the second protein was taken to be the prediction for the first, an 'ignorant' or 'random' procedure. For two states, the accuracy of the random prediction is Q₂ = 54.5%, not 50%. For comparison: for three states, random accuracy is Q₃ = 35.2%, not 33.3% (Q_α = 23, Q_L = 47 of observed or of predicted).

PHD: result for a cross-validation test when the network was trained to distinguish three states and applied to classify into helix/non-helix (Rost and Sander, 1993b). The chains were as follows.

On 22: (given in PDB code; number of residues in brackets) 256h_A (106), 4bp2 (118), 1cc5 (83), 2ocy_A (127), 3eln (143), 4cpv (108), 6cta (431), 2cyp (293), 5cya (103), 1eca (136), 1hdb_B (145), 2hrz_A (114), 3icb (75), 1158 (164), 2lb4 (153), 2h1b (149), 1hd_3 (87), 5lyz (129), 1pmb_A (153), 1adh_A (146), 2mv_P (154), 2wrp_R (104); on 6: 1col (197), 1hdd (57), 1pou (72), 1rop (56), 2acp (348), 2zta (31); on 4: 256b_A (106), 4hp2 (118), 351c (82), 9pap_A (108); on IL-4: interleukin 4 as given explicitly in Figure 2 (Redfield *et al.*, 1992; Smith *et al.*, 1992).

N, number of residues in the data set; N_{prot}, number of proteins in the data set.

Q₂, Q_α, Q_L, percentages of correctly predicted residues for two states and for helix and loop.

Corr_α, corr_L, Matthews (1975) correlation coefficients.

Somewhat surprisingly, the two neural networks, one trained on two states (training set, 12 all-helical proteins) and one on three states (training set, 130 proteins of different types), give approximately the same two-state accuracy on the four test proteins: Q₂ = 82.7% for helix net, Q₂ = 81.2% for PHD (Table I, Figure 1). We draw the preliminary conclusion that there is little or no advantage in using a method trained on two structural states, even when a protein is known to be all-helical. The conclusion is preliminary for two reasons: the test set of four proteins used by MKS is very small and two-state prediction for strand/non-strand was not considered here.

The balance shifts in favour of using three-state predictions, when one takes into account the additional problem of determining whether a protein is all-helical or not. This is true since two misclassifications out of 24 proteins have already decreased the accuracy (data not shown). CD spectroscopy or prediction methods may occasionally falsely classify some proteins (Deleage and Roux, 1987; Curtis *et al.*, 1991; Muskal and Kim, 1992; Zhang and Chou, 1992; Rost and Sander, 1993b). So we see the following trend: the improvement in prediction accuracy achieved by training neural networks on all-helical proteins is so marginal that it is offset by any error in classifying a protein as, e.g. all-helical. So, on balance, a one-step prediction of secondary structure is probably preferable to a two-step procedure of first predicting the structural class and subsequently using a class-specific method. For statistical prediction methods the situation appears to be similar (J. Garnier, personal communication).

Unfortunately, the comparison of two- and three-state

prediction methods is sometimes confused in the literature. The confusion can be avoided by noting that two-state prediction accuracy (Q₂) by definition always has more impressive values than three-state accuracy (Q₃), for the same method (Maxfield and Scheraga, 1976). This is simply a consequence of a trade-off between more refined distinction of states and higher prediction accuracy. For example, evaluation of the PHD prediction on the four test proteins yields a lower mean accuracy on three states, Q₃ = 70%, than on two states, Q₂ = 81%. The difference comes, e.g. from counting loop residues predicted as β-strands as incorrect in Q₃ and as correct in Q₂. The difference between the two measures is also obvious from the fact that the 'ignorance level' of Q₂ is ~55% while that for Q₃ is ~36% (these numbers were derived from alignments of sequences with dissimilar tertiary structure, Table II).

Expected prediction accuracy

Previous two-state predictions have achieved overall accuracies of Q₂ = 76% (Presnell *et al.*, 1992), Q₂ = 76% (Maxfield and Scheraga, 1976) and Q₂ = 79% (Kneller *et al.*, 1990). Although the level reached by all three methods described here is several percentage points higher, at Q₂ ≈ 80% it is premature to expect generally this level of accuracy in future applications to newly determined protein sequences. The reason is in the very limited size of the test set used. Experience with other prediction methods (Garnier and Levin, 1991; Robson and Garnier, 1993) clearly indicates that accuracy can vary considerably between different test sets selected from a large database. For example, a simple multilayered network was reported to yield Q₃ = 64.3% when tested on a single set of 15 proteins (Qian and Sejnowski, 1988). A comparable network reached only a mean of Q₃ = 61.7% when seven different test sets were selected from the database in independent trials ('7-fold cross-validation') (Rost and Sander, 1993b). The difference between best and worst test set was more than five percentage points. The fluctuation in accuracy is even larger when single proteins are considered: the accuracy per chain typically ranges somewhere between 40 and 90%. Just as Qian and Sejnowski (1988) had picked, by chance, a particularly favourable set of proteins, it is conceivable that the set chosen by MKS, two cytochromes, papain and phospholipase, yields a higher than mean prediction score. Indeed this appears to be the case: while PHD scored Q₂ = 81.2% on these four test proteins, it scored only Q₂ = 77.8% on another set of 10 helical proteins (chosen from the training set of MKS, but properly cross-validated, i.e. without using homologues in training).

An analysis of the network system when trained in three states and applied to a two-state prediction based on 22 all-helical proteins yields Q₂ = 75.9% with an SD of 8.2% (using multiple cross-validation, Table II). For six further chains and interleukin-4 (Figure 2), a recently predicted (Klein and DeLisi, 1986) and experimentally solved (Redfield *et al.*, 1992; Smith *et al.*, 1992) structure, the predictions were more accurate (Table II). In summary, the 80% level of accuracy is only an approximate estimate of future performance.

Fundamental limitation of single-residue accuracy scores

A general limitation in the evaluation of an accuracy as high as 80% is due to the fact that secondary structure is not uniquely defined. Although the DSSP method of extracting strings of secondary structure symbols from 3-D coordinates has become a *de facto* standard, it is interesting to compare other automatic

1.....2.....3.....4.....5.....6.....7.....8
sequence	MHKCDITLQEI IKTLSLSEKTLCTELTVTDIFPAASKNITTEKETPCRAATVLRQFYSHHEKDTRCLGATAQQPFRHKQ
observed	HHHHHHHHHHHH EEE HHHHHHHHHHHHHHHHHHH HHHHHHH
PHD	HHHHHHHHHHHH EEEEE HHH HHHHHHHHHH HHHHHHHHH
1.....2.....3.....4.....5.....6
sequence	LIRFLKRLDRNLWGLAGLNSCPVKEANQSTLENFLERLKTIMREKYSKCSS
observed	HHHHHHHHHHHH EEE HHHHHHHHHHHHHHH
PHD	HHHHHHHHHHHHH HHHHHHHHHHHHHH

Fig. 2. Predictions of an interesting recently solved structure, interleukin-4. Prediction of the four-helix bundle interleukin-4 by PHD in three states: H (α -helix), E (extended β -strand) and loop (space). The assignment of secondary structure was taken from Rodfield *et al.* (1992) and Smith *et al.* (1992). This prediction illustrates the limitations of two-state methods. The protein had been classified as all-helical by CD and the tertiary structure predicted on this basis (Curtis *et al.*, 1991). Although the overall bundle architecture can be predicted this way, the small but structurally important β -sheet (predicted here in part) was missed.

methods or definitions based on visual definition, as these generally give different assignments in detail. For example, Woodcock *et al.* (1992) report only 79% agreement between DSSP and PCURVES assignments. In addition, variation in secondary structure segments in different crystal forms is non-negligible and proteins homologous in 3-D structure rarely have identical secondary structure strings. So it appears unwise to measure the success of structure prediction methods exclusively on the basis of single-residue accuracy, with 100% as the (misleading) goal. What is needed are measures that evaluate whether or not the presence of a helical or strand segment has been predicted, irrespective of slight shifts in segment ends (Taylor, 1984; Taylor and Thornton, 1984; Presnell *et al.*, 1992; Rost *et al.*, 1993; B.Rost, R.Scheider and C.Sander, submitted for publication).

Recommendations for practical use of secondary structure predictions

With many new protein sequences coming out of molecular biology laboratories these days, a practical guide to structure prediction for non-specialists would be useful. The main points are as follows:

- (i) The best way to predict protein tertiary structure (and, by implication, secondary structure) is to detect a protein in the database of known structure with sufficient sequence similarity to imply structural homology, e.g. >25% identical residues over a length of 80 or more residues, with not more than three or four gaps (Schneider and Sander, 1991). Only when similarities cannot be found does secondary structure prediction become meaningful.
- (ii) Whenever there are several structurally homologous sequences, as detected by alignment methods, in a family, it is advantageous to use methods that exploit the information contained in the multiple sequence alignment, as these generally give better results (Zvelebil *et al.*, 1987; Frampton *et al.*, 1989; Barton *et al.*, 1991; Niemann and Kirschner, 1991; Benner *et al.*, 1992; Musacchio *et al.*, 1992; Rost and Sander, 1992; Russell *et al.*, 1992; Barton and Russell, 1993; Benner *et al.*, 1993; Gerloff *et al.*, 1993; Gibson *et al.*, 1993; Rost and Sander, 1993a,b).
- (iii) Whenever a sequence family has only one member, any of

a much larger number of methods are useful (we do not discuss these in detail). The expected accuracy Q_3 is approximately five percentage points lower when no family information is available.

- (iv) Given knowledge of the folding type of a protein, e.g. as a result of spectroscopic experiment, one can either use a method trained on just that structural type, e.g. on all-helical proteins or use a general method and interpret the result within the constraints for this type of fold, e.g. by interpreting all predicted β -strands as non-helix. Surprisingly, the tests reported here on a limited data set indicate that either way has approximately the same level of accuracy: the general three-state network using multi-sequence alignments as input performs approximately as well as a method exclusively developed for helical proteins.

In spite of the current improvements in secondary structure prediction methods, the principal goal of predicting the overall 3-D structure has not yet been reached (Rost and Sander, 1993b). It therefore becomes crucial to evaluate 1-D secondary structure predictions according to their ability to deliver a correct image of the 3-D structure, e.g. by segment-based criteria. Introduction of more aspects of tertiary structure into the recent generation of new prediction methods is eagerly awaited.

References

- Barton,G.J. and Russell,R.B. (1993) *Nature*, **361**, 505–506.
 Barton,G.J., Newman,R.H., Freemont,P.S. and Crumpton,M.J. (1991) *Eur. J. Biochem.*, **196**, 749–760.
 Benner,S.A., Cohen,M.A. and Gerloff,D. (1992) *Nature*, **359**, 781.
 Benner,S.A., Cohen,M.A. and Gerloff,D. (1993) *J. Mol. Biol.*, **229**, 295–305.
 Bronak,S. (1991) In Benhar,O., Bosio,C., Del Giudice,P. and Tabet,E. (eds), *Neural Networks From Biology to High Energy Physics*. Elba, Italy, pp. 277–288.
 Chou,P.Y. and Fasman,U.D. (1974) *Biochemistry*, **13**, 211–215.
 Curtis,B.M., Presnell,S.R., Srinivasan,S., Sassenfield,H., Klinke,R., Jeffery,E., Cosman,D., March,C.J. and Cohen,J.E. (1991) *Proteins*, **11**, 111–119.
 Deleage,G. and Roux,B. (1987) *Protein Engng.*, **1**, 289–294.
 Frampton,J., Lutz,A., Gibson,T.J. and Grif,T. (1989) *Nature*, **342**, 134.
 Garnier,J. and Levin,J.M. (1991) *CABIOS*, **7**, 133–142.
 Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **130**, 97–120.
 Gerloff,D.L., Jenny,T.F., Knecht,L.J., Gonnert,G.H. and Benner,S.A. (1993) *FEBS Lett.*, **318**, 118–124.
 Gibson,T.J., Thompson,J.D. and Abagyan,R.A. (1993) *Protein Engng.*, **6**, 41–50.
 Hayward,S. and Collins,J.F. (1992) *Proteins*, **14**, 372–381.

- Hobahn, L., Scharf, M., Schneider, R. and Sander, C. (1992) *Protein Sci.*, **1**, 409–417.
- King, R. D. and Sternberg, M. J. (1990) *J. Mol. Biol.*, **216**, 441–457.
- King, R. D., Muggleton, S., Lewis, R. A. and Sternberg, M. J. E. (1992) *Proc. Natl Acad. Sci. USA*, **89**, 11322–11326.
- Klein, P. and DeLisi, C. (1986) *Biopolymers*, **25**, 1659–1672.
- Kueler, D. G., Cohen, F. E. and Langridge, R. (1990) *J. Mol. Biol.*, **214**, 171–182.
- Lim, V. I. (1974) *J. Mol. Biol.*, **88**, 857–872.
- Matthews, B. W. (1975) *Biochim. Biophys. Acta.*, **405**, 442–451.
- Maxfield, F. R. and Scheraga, H. A. (1976) *Biochemistry*, **15**, 5138–5153.
- Muggleton, S., King, R. D. and Sternberg, M. J. E. (1992) *Protein Engng.*, **5**, 647–657.
- Musacchio, A., Gibson, T., Lelito, V. P. and Saraste, M. (1992) *FEBS Lett.*, **307**, 55–61.
- Muskal, S. M. and Kim, S.-H. (1992) *J. Mol. Biol.*, **225**, 713–727.
- Nagano, K. (1973) *J. Mol. Biol.*, **75**, 401–420.
- Niermann, T. and Kirschner, K. (1991) *Protein Engng.*, **4**, 359–370.
- Presnell, S. R., Cohen, B. J. and Cohen, F. E. (1992) *Biochemistry*, **31**, 983–993.
- Qian, N. and Sejnowski, T. J. (1988) *J. Mol. Biol.*, **202**, 865–884.
- Redfield, C., Boyd, J., Smith, L. J., Smith, R. A. G. and Dobson, C. M. (1992) *Biochemistry*, **31**, 10431–10437.
- Robson, B. and Garnier, J. (1993) *Nature*, **361**, 506.
- Rost, B. and Sander, C. (1992) *Nature*, **360**, 540.
- Rost, B. and Sander, C. (1993a) *Proc. Natl Acad. Science, USA*, in press.
- Rost, B. and Sander, C. (1993b) *J. Mol. Biol.*, **232**, in press.
- Rost, B., Sander, C. and Schneider, R. (1993) *TIBS*, **18**, 120–123.
- Russell, R. B., Breed, J. and Barton, G. J. (1992) *FEBS Lett.*, **304**, 15–20.
- Schneider, R. and Sander, C. (1991) *Proteins*, **9**, 56–68.
- Smith, L. J., Redfield, C., Boyd, J., Lawrence, G. M. P., Edwards, R. G., Smith, R. A. G. and Dobson, C. M. (1992) *J. Mol. Biol.*, **224**, 899–904.
- Taylor, W. R. (1984) *J. Mol. Biol.*, **173**, 512–521.
- Taylor, W. R. and Thornton, J. M. (1984) *J. Mol. Biol.*, **173**, 487–514.
- Wozniak, S., Morion, J.-P. and Henriusat, B. (1992) *Protein Engng.*, **5**, 629–635.
- Zhang, C.-T. and Chou, K.-C. (1992) *Protein Sci.*, **1**, 401–408.
- Zvelebil, M. J., Barton, G. J., Taylor, W. R. and Sternberg, M. J. E. (1987) *J. Mol. Biol.*, **195**, 957–961.

Received December 9, 1992; revised March 23, 1993; accepted July 9, 1993