Protein Engineering vol.6 no.3 pp.261–266, 1993

Secondary structure prediction for modelling by homology

P.E.Boscott[1], G.J.Barton[2] and W.G.Richards[1,3]

[1]Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ and [2]Laboratory of Molecular Biophysics, Rex Richards Building, South Parks Road, Oxford OX1 3QU, UK

[3]To whom correspondence should be addressed

An improved method of secondary structure prediction has been developed to aid the modelling of proteins by homology. Selected data from four published algorithms are scaled and combined as a weighted mean to produce consensus algorithms. Each consensus algorithm is used to predict the secondary structure of a protein homologous to the target protein and of known structure. By comparison of the predictions to the known structure, accuracy values are calculated and a consensus algorithm chosen as the optimum combination of the composite data for prediction of the homologous protein. This customized algorithm is then used to predict the secondary structure of the unknown protein. In this manner the secondary structure prediction is initially tuned to the required protein family before prediction of the target protein. The method improves statistical secondary structure prediction and can be incorporated into more comprehensive systems such as those involving consensus prediction from multiple sequence alignments. Thirty one proteins from five families were used to compare the new method to that of Garnier, Osguthorpe and Robson (GOR) and sequence alignment. The improvement over GOR is naturally dependent on the similarity of the homologous protein, varying from a mean of 3% to 7% with increasing alignment significance score.

Key words: homology/prediction/secondary structure/sequence alignment

Introduction

Knowledge of a protein's tertiary structure is a fundamental guide to the understanding of biological function. Such knowledge can aid the design of inhibitors and transition state mimics (Richards, 1989; Sander and Smith, 1989). However, of the 40 000 proteins currently sequenced only 1000 3-D structures have been determined by X-ray crystallography or NMR. Where a protein of unknown 3-D structure shows clear homology to a protein of known structure, this information gap can be bridged by application of molecular modelling techniques (Blundell *et al.*, 1987; Swindells and Thornton, 1991).

The modelling procedure follows four basic steps: (i) sequence alignment of the proteins of known and unknown 3-D structure, (ii) substitution of side chains in the conserved core, (iii) modelling of insertions and deletions and (iv) refinement of the model. The critical first step in any modelling study is to obtain an accurate alignment of the two protein sequences. When sequence similarity is high, alignment is usually unambiguous (Barton and Sternberg, 1987). However, when sequence similarity is weak or bounded by large insertions and deletions, an accurate alignment may be difficult to obtain. When performing a sequence alignment to a protein of known 3-D structure, each aligned residue is being predicted to adopt a specific conformation. At its simplest this may be viewed as the prediction of the protein secondary structure as α-helix, β-strand and aperiodic (coil).

In this paper we are considering the most effective strategy for predicting the secondary structure of a protein, given the tertiary structure of at least one other member of the family. To provide a benchmark for improvement we first assess the accuracy of prediction by a conventional sequence alignment method (Barton and Sternberg, 1987; Barton, 1990) and a *de novo* secondary structure prediction method (Garnier *et al.*, 1978). We then show that a combined secondary structure prediction method that is trained on a member of the protein family provides a useful improvement in prediction accuracy for proteins that show weak sequence similarity. When predicting something as complex as protein secondary structure it is important to correlate as much information as possible. This method optimizes the prediction for one sequence from one, homologous, known structure. Should both a known structure and multiple sequences be available, e.g. the P450 superfamily, the weighted average structure prediction (WASP) algorithm can replace any standard algorithm in methods such as Zvelebil *et al.* (1987), to take full advantage of the available data. The WASP method is also complementary to sequence alignment and even when the latter is more accurate it can still play an important role in detecting incorrect assignments.

Methods

Secondary structure definitions

The secondary structure used in this study was obtained from the database program IDITIS (Oxford Molecular Ltd) using the DSSP algorithm (Kabsch and Sander, 1983). To obtain a three state assignment, α-helix (H), π-helix (P) and 3/10 helix (G) were classed as helix, extended (E) remained a class of its own and turn (T), bridge (B), bend (S) and coil were combined to form the coil class.

In the work presented here accuracy values are calculated using equation (1). This states the percentage of residues correctly predicted:

$$\text{accuracy} = \frac{\textit{correct} \times 100}{\textit{seqlen}} \qquad (1)$$

where *correct* is the number of residues correctly predicted and *seqlen* is the number of residues in the sequence.

All alignments were performed using the AMPS package (Barton and Sternberg, 1987; Barton, 1990) and the alignment scores are given as significance scores (SD) above the mean obtained for random sequences of the same length and composition (see Barton and Sternberg, 1987 for details). SD score values can be converted into approximate percentage identity values using Figure 1. Thirty-one proteins were compared in five families: serine proteinases (nine), immunoglobulin domains (eight), TIM barrels (four), dehydrogenases (four) and viral coat proteins (five) (Table I).

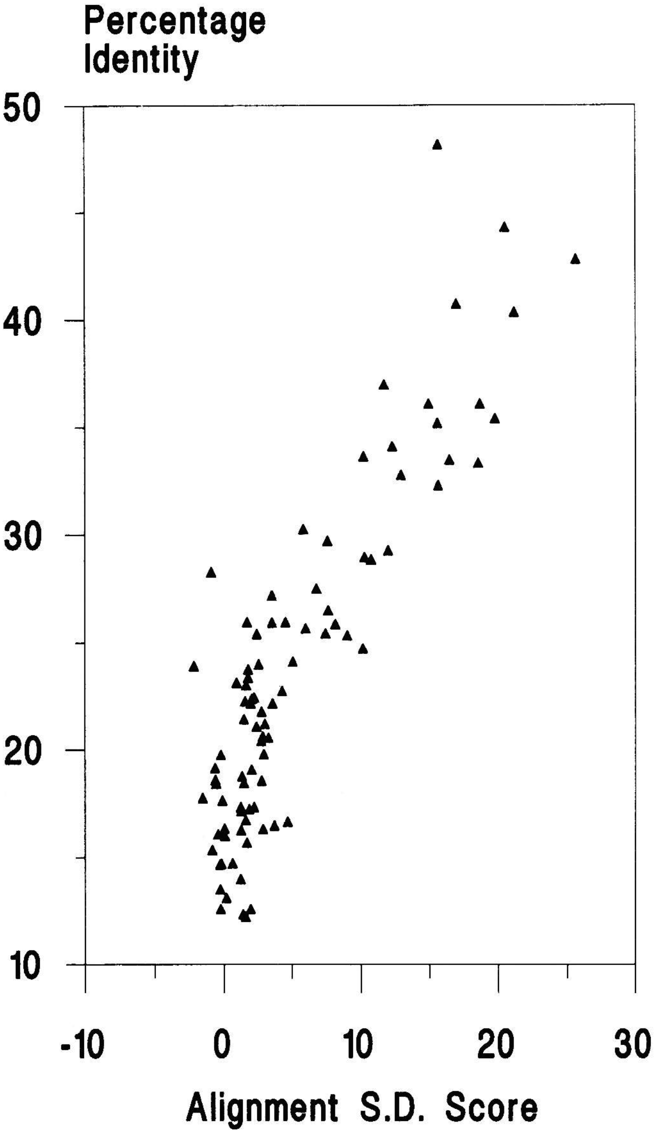


Fig. 1. A plot to show the correlation between significance score (SD) and percentage identity for 182 pairwise alignments of the proteins in Table I.

Results and discussion

Secondary structure prediction from sequence alignment

The sequences were aligned in pairs using the Needleman and Wunsch (1970) algorithm with the 250 PAM matrix (Dayhoff, 1978) using a gap penalty and a constant of eight. The known secondary structure of the two aligned proteins was superimposed onto the alignment and an accuracy calculated for how well the secondary structure of one protein was predicted by alignment to the other.

The accuracy was calculated using equation (1), but with the sequence length replaced by the number of aligned residues. This gives an accuracy 'for the residues predicted' and means that gaps in the alignment are not counted as incorrect predictions. The results (Figure 2 and Table II) show a good correlation between the two properties, in agreement with the more stringent test of Barton and Sternberg (1987).

Below a score of 2.5 SD the accuracy varies from 20 to 65% with a mean of 42% and a standard deviation of 10%. From 2.5 to 5 SD the range improves to between 40 and 65% with a mean of 55% and standard deviation of 8%. By far the most significant change is on either side of the 5 SD score. Between 5 and 15 SD the accuracy range becomes 60–90%, the mean 75% and the standard deviation falls to 6%. Finally, the significance scores above 15 SD occupy an accuracy range from 80 to 95%, having a mean of 85% and a standard deviation of 4%. The above results are also summarized in Table II.

When the alignment score is above 15 SD a protein can be confidently modelled from the tertiary structure of an homologous protein; the sequence alignment predicting at least four out of every five residues correctly. In the significance score range 5–15 SD, the secondary structural blocks are normally conserved although their lengths are known to be more variable. This is reflected in the mean alignment accuracy which shows that at least three out of five residues are correctly predicted. Modelling below the 5 SD limit is speculative. Even when the proteins are homologous and their core structures similar, there is a high possibility that the number and arrangement of the secondary structural blocks may have changed. An automatic sequence alignment is no longer adequate to obtain confidently the core structure of the unknown protein. Between 2.5 and 5 SD almost one in two residues is incorrectly predicted. Fortunately, the incorrect predictions are often localized into regions, however it is necessary to identify these regions and correct them before model building.

De novo secondary structure prediction

The most widely used algorithms are those which adopt the statistical approach to secondary structure prediction, such as Garnier *et al.* (1978) (GOR), Chou and Fasman (1978) (CF) and Gascuel and Golmard (1988) (GG). Each algorithm bases its prediction on similar information: the protein primary sequence and, if knowledge of a homologous protein is available, the approximate percentages of each class of secondary structure.

The most accurate of the above methods was that of GOR. The results, obtained using the decision constants suggested in Garnier *et al.* (1978), are given below (Table II). For the 31 proteins used in this study the GOR prediction accuracy was in the range 40–75%, with a mean of 57% and a standard deviation of 8%.

Weighted average structure prediction (WASP)

The WASP program allows the secondary structure prediction information from several standard methods to be combined into a single prediction. In this case Garnier *et al.* (1978) (GOR), Chou and Fasman (1978) (CF), Gascuel and Golmard (1988) (GG) and Hopp and Woods (1981) (HW) were used. Knowing the secondary structure of the homologous protein, it is possible to select the optimum combination of standard algorithms to predict it. Each secondary structure prediction is performed independently, in this case using three standard algorithms per class of secondary structure.

Coil (1)	HW hydrophilicity parameters.
Coil (2)	GOR coil parameters.
Coil (3)	GG coil parameters.
Helix (1)	CF helix parameters.
Helix (2)	GOR helix parameters.
Helix (3)	GG helix parameters.
Sheet (1)	CF sheet parameters.
Sheet (2)	GOR sheet parameters.
Sheet (3)	GG sheet parameters.

It is the statistical information from these standard algorithms which is used to form prediction profiles. The Hopp and Woods

Table I. The proteins used in the study

Protein	Segment	Length	Code	Reference
Trypsin	–	233	1sgt	Read and Games (1988)
Alpha-lytic protease	–	198	2alp	Fujinaga *et al.* (1985)
Proteinase A	–	181	2sga	Moult *et al.* (1985)
Proteinase B	–	185	3sgb	Fujinaga *et al.* (1987)
Tonin	–	235	1ton	Ashley and MacDonald (1985)
Trypsin	–	223	2ptn	Marquart *et al.* (1983)
Native elastase		240	3est	Radhakrishnan *et al.* (1987)
Rat mast cell protease	–	224	3rp2	Reynolds *et al.* (1985)
Hydrolase	–	245	4cha	Blow (1976)
FC fragment	ch2	105	1fc1	Deisenhofer *et al.* (1976)
FC fragment	ch3	101	1fc1	Deisenhofer *et al.* (1976)
FAB fragment	ch1	99	1mcp	Rudikoff *et al.* (1981)
FAB fragment	vh	123	1mcp	Rudikoff *et al.* (1981)
FAB fragment	vl	115	1mcp	Rudikoff *et al.* (1981)
Immunoglobulin FAB	ch1	103	2fb4	Kratzin *et al.* (1989)
Immunoglobulin FAB	vh	126	2fb4	Kratzin *et al.* (1989)
Immunoglobulin FAB	vl	112	2fb4	Kratzin *et al.* (1989)
Glycolate oxidase	–	369	1gox	Lindqvist and Branden (1989)
Triose phosphate isomerase	A	247	1tim	Alber *et al.* (1981)
Typtophan synthase	A	268	1wsy	Hyde *et al.* (1988)
D-Xylose isomerase	–	393	5xia	Kenrick *et al.* (1987)
Glyceraldehyde dehydrogenase	O	334	1gd1	Branlant *et al.* (1989)
Cytoplasmic malate dehydrogenase	A	334	4mdh	Birktoft *et al.* (1989)
Lactate dehydrogenase	–	330	6ldh	Zapatero *et al.* (1987)
Apo-liver alcohol dehydrogenase	–	374	8adh	Colonna *et al.* (1986)
Viral coat protein – mengo encephalomyocarditis	–	277	2mev	Luo *et al.* (1978)
Rhinovirus	–	289	2rs1	Arnold and Rossman (1990)
Tobacco necrosis virus	–	195	2stv	Liljas and Strandberg (1984)
Tomato bushy stunt virus	–	387	2tbv	Hopper *et al.* (1984)
Southern bean mosaic viral coat protein	–	261	4sbv	Rossman *et al.* (1983)

hydrophilicity profile was calculated by taking a moving average of seven residues and the CF profiles by a moving average of five.

The WASP profiles are then formed by summing a given percentage of each standard profile. In Figure 3, four residues of a protein primary sequence are shown with the associated prediction profile values from each standard algorithm scaled from 0 to 100. Below this is a WASP profile constructed from 25% HW, 50% GG and 25% GOR.

When the WASP profiles are trained on the homologous protein they are generated by combination of the three algorithms in user defined increments. For the results given here that increment was 4%, meaning the first WASP profile comprised 96% HW, 4% GOR and 0% GG and the second 92% HW, 8% GOR and 0% GG etc. This means that a total of 25^3 (or 15 625) different WASP profiles will be generated.

Having formed a WASP profile, a cut-off value is varied between 0 and 100 in steps of 2. Residues that have a WASP profile value greater than the cut-off value are predicted as adopting the given secondary structure, those with a WASP profile value equal or less are not predicted. Each WASP profile therefore gives rise to 50 predictions, meaning that the entire process will generate 50 × 15 625 (or 781 250) predictions of the protein of known secondary structure.

Each WASP prediction is described by four parameters: percentage of algorithm-1, percentage of algorithm-2, percentage of algorithm-3 and a cut-off value. From these numbers a prediction of a given secondary structure can be made. The 781 250 predictions are compared to the known structure of the protein and the function given in equation (2) is evaluated.

$$\begin{array}{l}\text{training} \\ \text{accuracy}\end{array} = \frac{(\mathit{correct} - \mathit{incorrect})}{\mathit{total}} \times 100 - \frac{(\mathit{predicted} - \mathit{total})}{\mathit{total}} \times 50 \qquad (2)$$

where *correct* is the number of residues correctly predicted (or true-positive predictions), *incorrect* is the number of residues incorrectly predicted (false-negative and true-negative), *total* is the number of residues known to have the given secondary structure, and *predicted* is the number of residues predicted to have the given secondary structure. The first part of the equation returns a value of 100 for a completely correct prediction and −100 for one that is completely incorrect. As each secondary structure is predicted independently the accuracies are biased by the number of residues predicted to discourage total and zero predictions. If the number predicted is correct *predicted = total* and the bias is zero, otherwise a scaled value is subtracted based on the modulus of the difference, making it the same for under- and over-prediction. The equation was developed by visually comparing known and predicted secondary structures on a customized graphical interface (Boscott, 1990).

At the end of training there is a WASP profile and cut-off value for each class of secondary structure and an associated training accuracy from equation (2). After training, the WASP algorithms are by definition as good as, or better than, the best composite algorithm prediction. A qualitative example of the result of training is shown in Figure 4.

The WASP prediction accuracy is naturally limited by that of

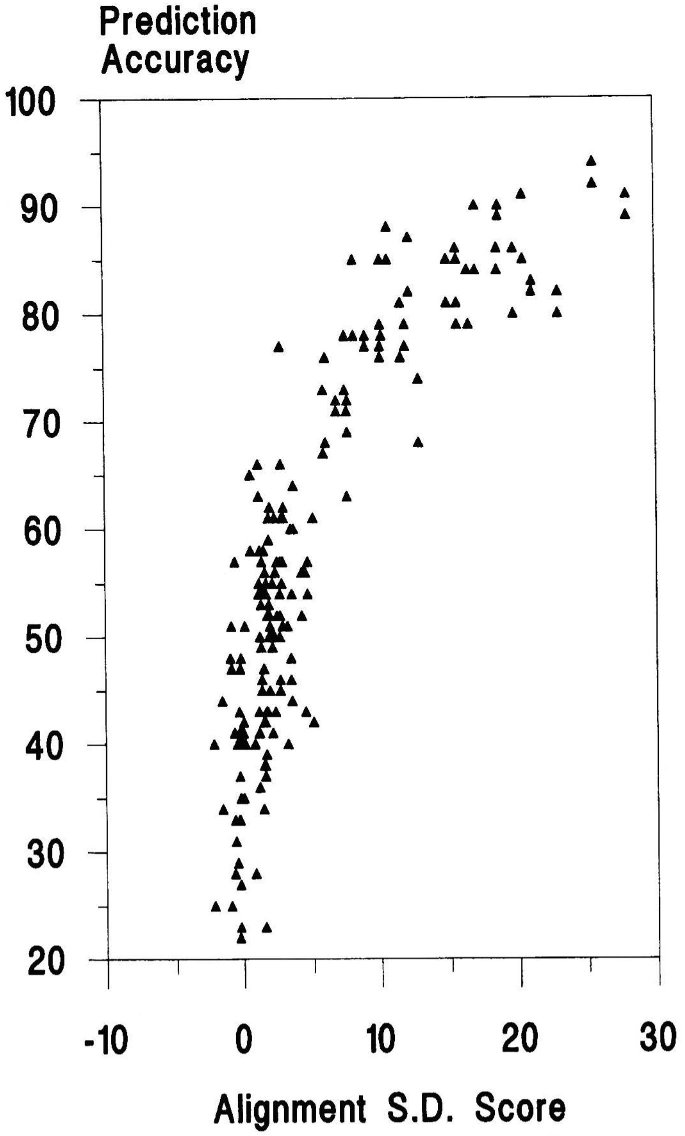


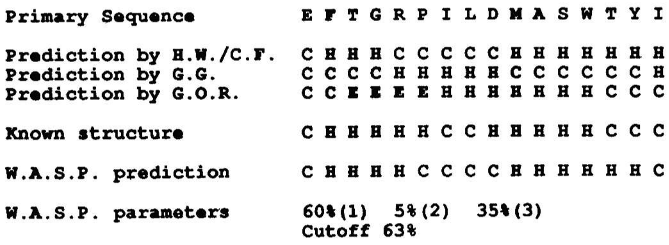
Fig. 2. The accuracy of secondary structure prediction by sequence alignment plotted against the alignment SD score to the homologous protein. One hundred and eighty-two predictions were made from pairwise alignment of the proteins in Table I.

Table II. The results of the secondary structure prediction methods

Method	Significance score to training protein (SD)	Accuracy mean (%)	Accuracy range (%)	Accuracy standard deviation (%)
GOR	–	57	40–70	8
Theroretical maximum from WASP	–	66	50–90	7
WASP	<2.5	61	40–85	7
	2.5–5	59	35–85	10
	5–15	64	45–90	8
	>15	64	50–75	4
	All values	62	35–90	8
Alignment	<2.5	45	20–65	10
	2.5–5	55	40–65	8
	5–15	75	60–90	9
	>15	85	80–95	4

Protein Sequence (One letter code)	**A**	**P**	**C**	**D**
(1) Coil H.W.	**55**	**59**	**47**	**55**
(2) Coil G.G.	**70**	**75**	**88**	**84**
(3) Coil G.O.R.	**41**	**65**	**72**	**81**
25% - (1)	**55/4+**	**59/4+**	**47/4+**	**55/4+**
50% - (2)	**70/2+**	**75/2+**	**88/2+**	**84/2+**
25% - (3)	**41/4+**	**65/4+**	**72/4+**	**81/4+**
W.A.S.P. - coil	**59**	**69**	**74**	**75**

Fig. 3. The formation of a WASP profile over four residues of a sequence by combining the prediction profiles from three standard algorithms in the ratio 1:2:1.



```
Primary Sequence          E F T G R P I L D M A S W T Y I

Prediction by H.W./C.F.   C H H H C C C C C H H H H H H H
Prediction by G.G.        C C C C H H H H H C C C C C C H
Prediction by G.O.R.      C C E E E E H H H H H H H C C C

Known structure           C H H H H H C C H H H H H C C C

W.A.S.P. prediction       C H H H H C C C C H H H H H H C

W.A.S.P. parameters       60%(1)  5%(2)  35%(3)
                          Cutoff 63%
```

Fig. 4. A typical optimized WASP prediction by combination of three algorithms.

the standard algorithms themselves. For example, if none of the three algorithms predict a region to be helical, it is very unlikely that the WASP prediction will be helical. Where the method gains, is when the different algorithms are correct for different regions of the protein, as is shown by example in Figure 4. The four WASP parameters for each class of secondary structure together with their respective training accuracies are then used for prediction of the target protein.

The information used in prediction is as follows.

(i) The primary sequence of the unknown protein.
(ii) The percentages of helical, extended and coil residues in the homologous protein.
(iii) The optimum WASP parameters for predicting the homologous protein.
(iv) The training accuracy values from predicting the homologous protein.

To predict the secondary structure of the unknown protein, each class of secondary structure is independently predicted using the WASP parameters generated by training on the homologous protein. This prediction is not mutually exclusive and a residue can be predicted as adopting helical, extended and coil conformations. Having profiles of all predicted states available is useful when interpreting the prediction by eye. However, for the purpose of evaluation, the following strategy was adopted to give a unique prediction at each position. When a multiple prediction occurred, the quantity in equation (3) was calculated for each of the predicted classes of secondary structure. The class with the maximum decision score was that chosen.

$$\text{decision score} = (\textit{profile}\text{ value} - \textit{cut-off}\text{ value}) \times \textit{percentage} \qquad (3)$$

where *profile* is the value of the WASP profile at a given residue, *cut-off* is the prediction cut-off value and *percentage* is the percentage of the given secondary structure in the training protein.

Having obtained a mutually exclusive secondary structure prediction for the unknown protein the prediction can be further optimized. This is done by comparing the percentage of each

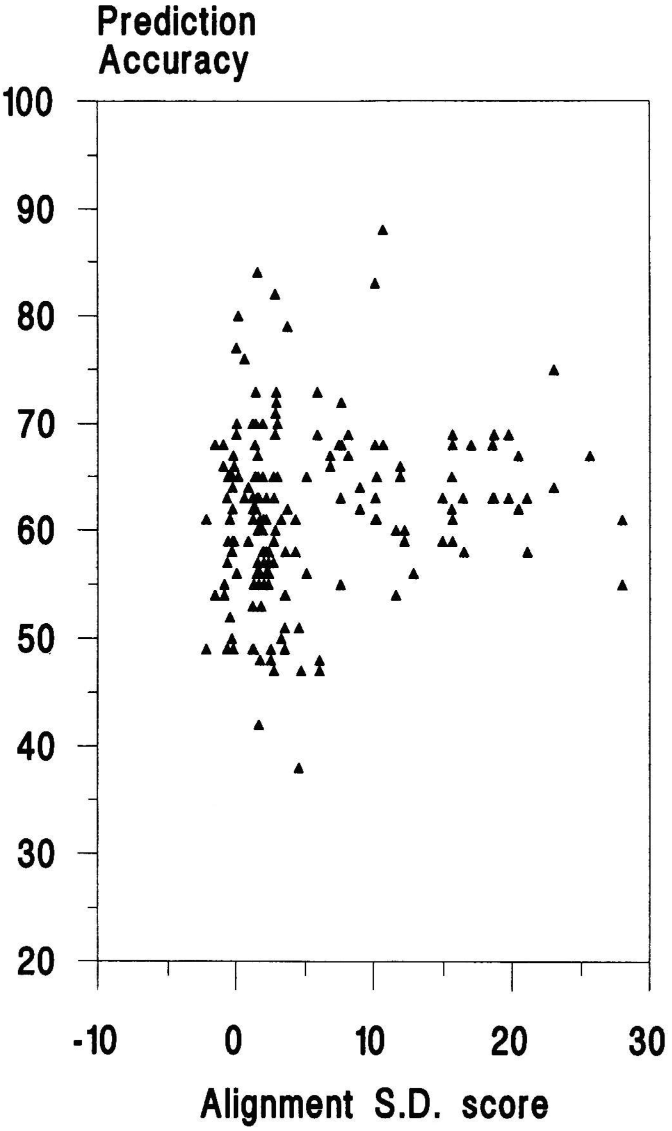


Fig. 5. The accuracy of secondary structure prediction by the WASP method plotted against the alignment SD score to the homologous protein. One hundred and eighty-two predictions were made from pairwise alignment of the proteins in Table I.

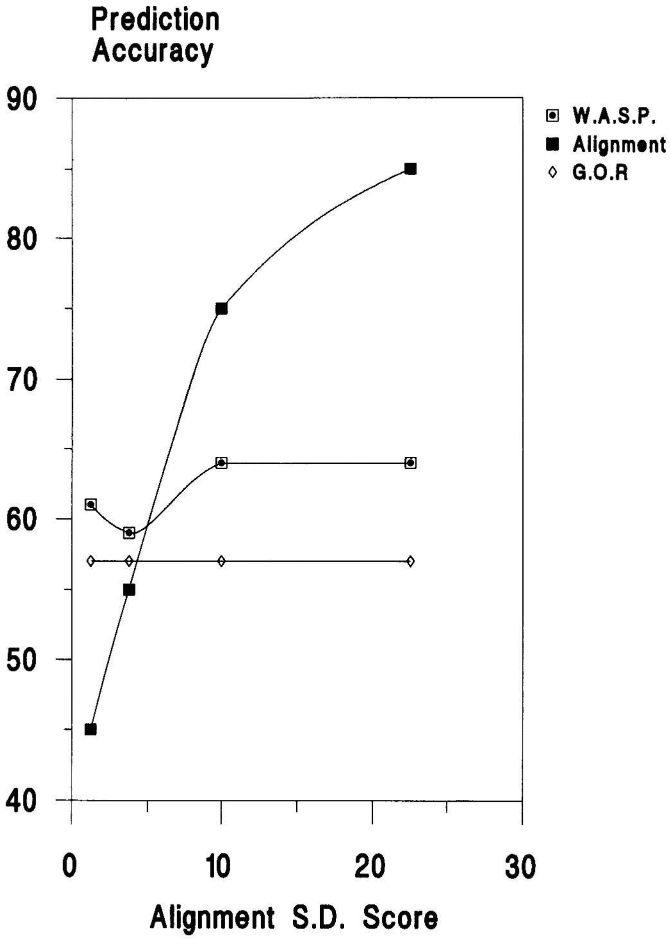


Fig. 6. An interpolation through mean accuracy values for secondary structure prediction by sequence alignment, WASP and GOR. The proteins are given in Table I and the accuracy values in Table II.

secondary structure predicted to the known percentages in the homologous protein. If there is a large discrepancy in any of the values the amounts of secondary structure predicted can be easily varied by movement of the cut-off values. This was again automated by the following rules in which the 'required percentage' of each secondary structure is that of the homologous protein.

(i) If a secondary structure is over-predicted by more than 10% of its required value the cut-off is incrementally increased.

(ii) If a secondary structure is under-predicted by more than 10% of its required value and the required value is more than 30%, the cut-off is incrementally reduced.

(iii) If a secondary structure is under-predicted by more than 10% of its required value, the required value is more than 15% and the training accuracy is positive, the cut-off is incrementally reduced.

Rule (iii) above is used to determine how accurately each of the secondary structural classes is being predicted. It is often the case where the percentage of a given secondary structure is low that its entire prediction is incorrect, e.g. helix prediction in some serine proteinases. Where the accuracy of prediction for the homologous protein is low the program will be less inclined to over-predict that class of secondary structure.

The maximum accuracy that the WASP system can achieve for any given protein is the prediction of the protein when trained on itself. As stated above, predicting 31 proteins by training the WASP system on homologous proteins gave a mean accuracy of 62% and a standard deviation of 8%. The maximum possible accuracy for WASP applied to these proteins is a mean of 66% with a standard deviation of 7%.

The results of the WASP predictions are shown plotted against significance score in Figure 5 and summarized in Table II. Below an alignment score of 2.5 SD the WASP prediction accuracy ranges from 40 to 85% with a mean of 61% and a standard deviation of 7%. These values become slightly worse in the 2.5–5 SD bracket having a mean of 59% and a standard deviation of 10%. It is reasonable to conclude that below 5 SD the range is 35–85%, the mean 60% and the standard deviation 8%. Modelling between 5 and 15 SD gives accuracy values in the range 45–90%, with a mean of 64% and a standard deviation of 8%. Finally, above an alignment score of 15 SD the range contracts to between 50 and 75%, the mean remains at 64% and the standard deviation falls to 4%. Code for the WASP system is available by application to the authors.

Comparison of the WASP method to that of GOR

The WASP system first trains the prediction data from the GOR, CF, GG and HW methods on a homologous protein. It then assesses how well the homologous protein was predicted and uses this information to modify the prediction of the target protein. Based on the extra information used it is expected that WASP should out-perform the component algorithms, as is the case.

The results of the prediction by CF, GG and HW were, on average, worse than those of GOR. A comparison has therefore been made between the WASP system and the best of its component algorithms. The GOR predictions were further improved by using decision constants given in the paper; these were not used in the GOR data of the WASP system. The overall mean for the WASP method, predicting the 31 proteins in Table I from each other, is 62% compared to 57% for GOR.

The benefit of using the trained method is naturally dependent on the sequence similarity of the homologous protein. When the protein pair has a significance score less than 5 SD the mean accuracy is 60%, 3% higher than GOR. From Figure 5 this difference is not dependent on the six highest scoring points in the bracket. Removal of these data points lowers both the GOR and WASP accuracy by 1%. When the alignment score exceeds 5 SD the average accuracy for the WASP method rises to 64%, 7% higher than that for GOR. The WASP system could not achieve a mean accuracy greater than 66% using the combined data for GOR, GG, CF and HW as this is a limit currently imposed on the system by the accuracy of the composite data.

Table II gives the mean accuracies, accuracy ranges and standard deviations of the results from the GOR method, sequence alignment and WASP system. Plotting the mean values of each method (Figure 6) shows that WASP offers a small improvement in prediction accuracy over both GOR and sequence alignment when the training and test proteins have a similarity lower than 5 SD.

Summary

Our study of secondary structure prediction accuracy by sequence alignment shows a good correlation with the significance score of the alignment. These results can be used to support confidence in homology modelling based on the result of the alignment. From Figure 2 the threshold for a confident alignment is approximately 5 SD.

The accuracy of sequence alignment and that of secondary structure prediction can be used to divide homology modelling into two classifications. A 'confident' homology model can be built with an alignment score greater than 5 SD which relies primarily on the results of the sequence alignment. Between 2.5 and 5 SD a 'speculative' model can be inferred from a combination of the two predictive methods, with greater weight now being placed on the results of the secondary structure prediction method.

The primary conclusion of this work is that the accuracy of *de novo* secondary structure prediction for homology modelling can be improved by training the prediction method on the homologous protein, the mean increase in accuracy being 7% in the confident region (alignment scores greater than 5 SD) and 3% in the speculative region (alignment scores between 2.5 and 5 SD).

The weighted average structure prediction (WASP) method appears capable of combining algorithms, training on a protein of known structure and then predicting the secondary structure of a homologous protein with a greater accuracy than any of the constituent algorithms. It is likely that the accuracy of this method could be considerably increased by combining algorithms developed for specific protein families, or groups of families.

Acknowledgements

We would like to thank Dr Guy Grant and Dr Richard Gilbert for helpful advice and Mr Steve Gardener for access to the database program IDITIS (Oxford Molecular Ltd). Paul Boscott is supported by an SERC–CASE award with Oxford Molecular Ltd.

References

Alber,T., Banner,D.W., Bloomer,A.C., Petsko,G.A., Phillips,D., Rivers,P.S. and Wilson,I.A. (1981) *Phil. Trans. R. Soc. London Ser. B*, **293**, 159–171.

Arnold,E. and Rossmann,M.G. (1990) *J. Mol. Biol.*, **211**, 763–801.

Ashley,P.L. and MacDonald,R.J. (1985) *Biochemistry*, **24**, 4512–4520.

Barton,G.J. (1990) *Methods Enzymol.*, **183**, 403–428.

Barton,G.J. and Sternberg,M.J.E. (1987) *J. Mol. Biol.*, **198**, 327–337.

Birktoft,J., Fu,Z., Carnahan,G.E., Rhodes,G., Roderick,S.L. and Banaszak,L.J. (1989) *Biochemistry*, **28**, 6065–6081.

Blow,D.M. (1976) *Acc. Chem. Res.*, **9**, 145–152.

Blundell,T.L., Sibanda,B.L., Sternberg,M.J.E. and Thornton,J.M. (1987) *Nature*, **326**, 347–352.

Boscott,P.E. (1990) Part II Thesis. Final Honour School Natural Science Chemistry, University of Oxford.

Branlant,C., Oster,T. and Branlant,G. (1989) *Gene*, **75**, 145–155.

Chou,P.Y. and Fasman,G.D. (1978) *Adv. Enzymol. Rel. Mol. Biol.*, **47**, 54–148.

Colonna,F., Perahia,D., Karplus,M., Eklund,H., Branden,C.I. and Tapia,O. (1986) *J. Biol. Chem.*, **261**, 15273–15280.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, Vol. 5, Suppl. 3. National Biomedical Research Foundation.

Deisenhofer,J., Colman,P.M., Epp,O. and Huber,R. (1976) *Hoppe Seyler's Z Physiol.*, **357**, 1421–1434.

Fujinaga,M., Delbaere,L.T.J., Brayer,G.D. and James,M.N.G. (1985) *J. Mol. Biol.*, **184**, 479–502.

Fujinaga,M., Sielecki,A.R., Read,R.J., Ardelt,W., Laskowski,M. and James,M.N.G. (1987) *J. Mol. Biol.*, **195**, 397–418.

Garnier,J., Osguthorpe,D.J. and Robson,B. (1978) *J. Mol. Biol.*, **120**, 97–120.

Gascuel,O. and Golmard,J.L. (1988) *CABIOS*, **4**, 357–365.

Hopp,T.P. and Wood,K.R. (1981) *Proc. Natl Acad. Sci. USA*, **78**, 3824–3828.

Hopper,P., Harrison,S.C. and Sauer,R.T. (1984) *J. Mol. Biol.*, **177**, 701–713.

Hyde,C.C., Ahmed,S.A., Padlan,E.A., Miles,E.W. and Davies,D.R. (1988) *J. Biol. Chem.*, **263**, 17857–17871.

Kabsch,W. and Sander,C. (1983) *Biopolymers*, **22**, 2577–2637.

Kenrick,K., Blow,D.M., Carrell,H.L. and Glusker,J.P. (1987) *Protein Engng*, **1**, 467–469.

Kratzin,H.D., Palm,W., Stangel,M., Schmidt,W.E., Friedrich,J. and Hilschmann,N. (1989) *Biol. Chem. Hoppe Seyler*, **370**, 263–270.

Liljas,L. and Strandberg,B. (1984) *Biol. Macromol. Ass.*, **1**, 97–119.

Lindqvist,Y. and Branden,C.I. (1989) *J. Biol. Chem.*, **264**, 3624–3628.

Luo,M., Vriend,G., Kamer,G., Minor,I., Arnold,E., Rossman,M.G., Boege,U., Scraba,D.G. and Duke,G.M. (1978) *Science*, **235**, 182–191.

Marquart,M., Walter,J., Deisenhofer,W., Bode,W. and Huber,R. (1983) *Acta Crystallogr. Sect. B*, **39**, 480–490.

Moult,J., Sussman,F. and James,M.N.G. (1985) *J. Mol. Biol.*, **182**, 555–566.

Needleman,S.B. and Wunsch,C.D. (1988) *J. Mol. Biol.*, **48**, 443–453.

Radhakrishnan,R., Presta,L.G., Meyer,E.F. and Wildonger,R. (1987) *J. Mol. Biol.*, **198**, 417–424.

Read,R.J. and Games,M.N.G. (1988) *J. Mol. Biol.*, **200**, 523–551.

Reynolds,R.A., Remington,S.J., Weaver,L.H., Fisher,R.G., Anderson,W.F., Ammon,H.L. and Mathews,B.W. (1985) *Acta Crystallogr. Sect. B*, **41**, 139–147.

Richards,W.G. (1989) *Computer Aided Molecular Design*. VCH Publishers Inc., New York, USA.

Rossman,M.G., Zapatero,C.Abad, Hermodson,M.A. and Erickson,J.W. (1983) *J. Mol. Biol.*, **166**, 37–83.

Rudikoff,S., Satow,Y., Padlan,E., Davies,D. and Potter,M. (1981) *Mol. Immunol.*, **18**, 705–711.

Sander,M. and Smith,J.H. (1989) *Design of Enzyme Inhibitors*. Oxford Science Publications.

Swindells,M.B. and Thornton,J.M. (1991) *Curr. Opinion Struct. Biol.*, **1**, 219–223.

Zapatero,C.Abad, Griffith,J.P., Sussman,J.L. and Rossman,M.G. (1987) *J. Mol. Biol.*, **198**, 445–467.

Zvelebil,M.J., Barton,G.J., Taylor,W.R. and Sternberg,M.J.E. (1987) *J. Mol. Biol.*, **195**, 957–961.

Received on October 13, 1992; revised on December 22, 1992; accepted on January 4, 1993