Article No. jmbi.1999.3091 available online at http://www.idealibrary.com on IDEAL® J. Mol. Biol. (1999) 292, 195–202

JMB

COMMUNICATION

Protein Secondary Structure Prediction Based on Position-specific Scoring Matrices

David T. Jones

Department of Biological Sciences, University of Warwick, Coventry CV4 7AL United Kingdom

A two-stage neural network has been used to predict protein secondary structure based on the position specific scoring matrices generated by PSI-BLAST. Despite the simplicity and convenience of the approach used, the results are found to be superior to those produced by other methods, including the popular PHD method according to our own benchmarking results and the results from the recent Critical Assessment of Techniques for Protein Structure Prediction experiment (CASP3), where the method was evaluated by stringent blind testing. Using a new testing set based on a set of 187 unique folds, and three-way cross-validation based on structural similarity criteria rather than sequence similarity criteria used previously (no similar folds were present in both the testing and training sets) the method presented here (PSIPRED) achieved an average Q_3 score of between 76.5% to 78.3% depending on the precise definition of observed secondary structure used, which is the highest published score for any method to date. Given the success of the method in CASP3, it is reasonable to be confident that the evaluation presented here gives a fair indication of the performance of the method in general.

Keywords: protein structure prediction; secondary structure; protein folding; sequence analysis; neural network

As a result of the influx of sequence data from the numerous genome sequencing projects, interest has never been greater in methods for predicting protein structure from amino acid sequence. At present, the prediction of an unknown protein structure by comparative modelling (e.g. Sali, 1995) is by far the most reliable technique, but only when a template protein structure can be found with a very high degree of sequence similarity to the target protein. In the absence of a suitable homologous template structure with which to build a model for a given sequence, fold recognition methods now provide another option for constructing useful tertiary structural models (e.g. Bowie *et al.*, 1991; Jones *et al.*, 1992; Lemer *et al.*, 1995). Beyond methods based on recognizing similarities between proteins are *ab initio* tertiary methods, which attempt to predict the structure of a protein without reference to a template structure. Despite some recent progress in *ab initio* tertieary protein structure prediction (Jones, 1997), by far the most commonly used *ab initio* prediction methods are aimed at the prediction of secondary structural elements in proteins (e.g. Lim, 1974; Chou & Fasman, 1974; Garnier *et al.*, 1978; Zvelebil *et al.*, 1987; Rost & Sander, 1993; Geourjon & Deleage, 1995; Salamov & Solovyev, 1995; Frishman & Argos, 1996; King & Sternberg, 1996). Secondary structure prediction methods are not often used alone, but are instead often used to provide constraints for tertiary structure prediction methods or as part of fold recognition methods (e.g. Russell *et al.*, 1996; Rost, 1997).

Early methods for secondary structure prediction were based on either simple stereochemical principles (Lim, 1974) or statistics (Chou & Fasman, 1974; Garnier *et al.*, 1978). The GOR method (Garnier *et al*, 1978) has been particularly popular due the simplicity of implementing the method in software. Increasingly, however, rather than a single sequence a whole family of related sequences is available for analysis. By constructing a multiple sequence alignment, additional information may be obtained from the observed patterns in sequence variability, and the location of insertions and deletions. Probably the earliest

E-mail address of the corresponding author: jones@globin.bio.warwick.ac.uk

0022-2836/99/370195–8 $30.00/0

attempts at using such multiple sequence information for secondary structure prediction was the successful prediction of the secondary structure (and from this the fold) for the alpha-subunit of tryptophan synthase by Niermann *et al*. (1987), and a general method published by Zvelebil *et al*. (1987). It is fair to say, however, that the use of multiple sequence data was perhaps popularised by the later work by Benner & Gerloff (1991) on the successful secondary structure prediction for the cAMP-dependent kinases. The main source of information in this approach to secondary structure prediction is obtained by observing that the most conserved regions of a protein sequence are those regions which are either functionally important, and/or buried in the protein core. Conversely, the more variable regions can be fairly confidently assumed to be on the surface of the protein, where few constraints are imposed on the type of amino acid residues observed, apart from a bias towards hydrophilic amino acid residues. By clustering the sequences in an aligned family, and assessing the degree of sequence variability observed between very similar pairs, Benner & Gerloff demonstrated that the degree of solvent accessibility of an amino acid residue can be predicted with reasonable accuracy. Secondary structure can then be predicted by comparing the accessibility patterns generally associated with specific secondary structures when packed against a hydrophobic protein core.

Despite the apparent power of the manual approach described above, it is clearly beneficial to attempt to incorporate these ideas into an automatic method so that a large number of accurate predictions can be generated routinely. The PHD method by Rost & Sander (1993) uses a set of feed-forward neural networks trained by back-propagation (Rumelhart *et al*., 1986) to replace the ''human expert'' components of the Benner & Gerloff approach, and has since become the *de facto* standard secondary structure prediction method.

The method described here also makes use of neural networks, but is greatly simplified. Despite the simplification, the method achieves a very high degree of prediction accuracy, being the most accurate method evaluated in the recent third CASP experiment (Moult *et al*., 1997) and can be easily implemented and run on any common computer system.

Method

The prediction method (illustrated in Figure 1) is split into three stages: generation of a sequence profile, prediction of initial secondary structure, and finally the filtering of the predicted structure.

Generation of sequence profiles

The main design goal of this prediction method was to make the entire system easily ported to any workstation. This aim encompasses both the generation of sequence profiles and the actual prediction of secondary structure. Standard approaches to generating sequence profiles are cumbersome and time-consuming. For example, the PHD server (Rost & Sander, 1993) makes use of a large multiprocessor computer system to generate multiple sequence alignments in a timely fashion, and it is therefore difficult to move the whole PHD prediction server to another site. Furthermore, the prediction accuracy of methods based on multiple sequence alignments has been found to correlate with the degree of divergence present in the aligned set of sequences. Alignments which incorporate sequences with significant yet low sequence similarity to the target protein produce more accurate predictions that those which incorporate sequences which are very closely related to the target. Recently a new method for very sensitive sequence comparison based on the new gapped-version of BLAST has been published (Altschul *et al*., 1997). With suitable choices of parameters and filtering of the search data banks, PSI-BLAST greatly outperforms a standard Smith-Waterman (Smith & Waterman, 1981) search in its ability to detect distant homologues of a query sequence. In addition to this, PSI-BLAST generates sequence profiles as part of the search process, and here we explore the idea of using these intermediate PSI-BLAST profiles as a direct input to a secondary structure prediction method rather than extracting the sequences, and producing an explicit multiple sequence alignment as a separate step. By using the PSI-BLAST profiles directly, the very time-consuming multiple-sequence alignment stage is eliminated, and this leads to a radical reduction in the overall time taken to go from target sequence to predicted secondary structure. On a Silicon Graphics Origin 200 server, the entire prediction process takes only two minutes.

Although PSI-BLAST is a very powerful sequence searching method, it is prone to failure for a number of reasons. The iterative nature of the PSI-BLAST algorithm makes it very sensitive to biases in the sequence data banks. In particular, PSI-BLAST is very prone to erroneously incorporating repetitive sequences into the intermediate profiles. As soon as one or two of these pathological sequences are incorporated, then the whole process goes astray with completely random sequences being matched with apparent high confidence. In order to maximise the effectiveness of PSI-BLAST in producing very sensitive profiles, a custom sequence data bank was constructed for the present application. Firstly, a large non-redundant protein sequence data bank was compiled by extracting non-identical sequences from a number of publicly available data banks. This databank, which currently contains around 340,000 sequences, is then filtered with the SEG program (Wootton & Federhen, 1993) to remove regions with very low information content. A custom program is used to further filter the data bank in order to remove transmembrane segments (Jones

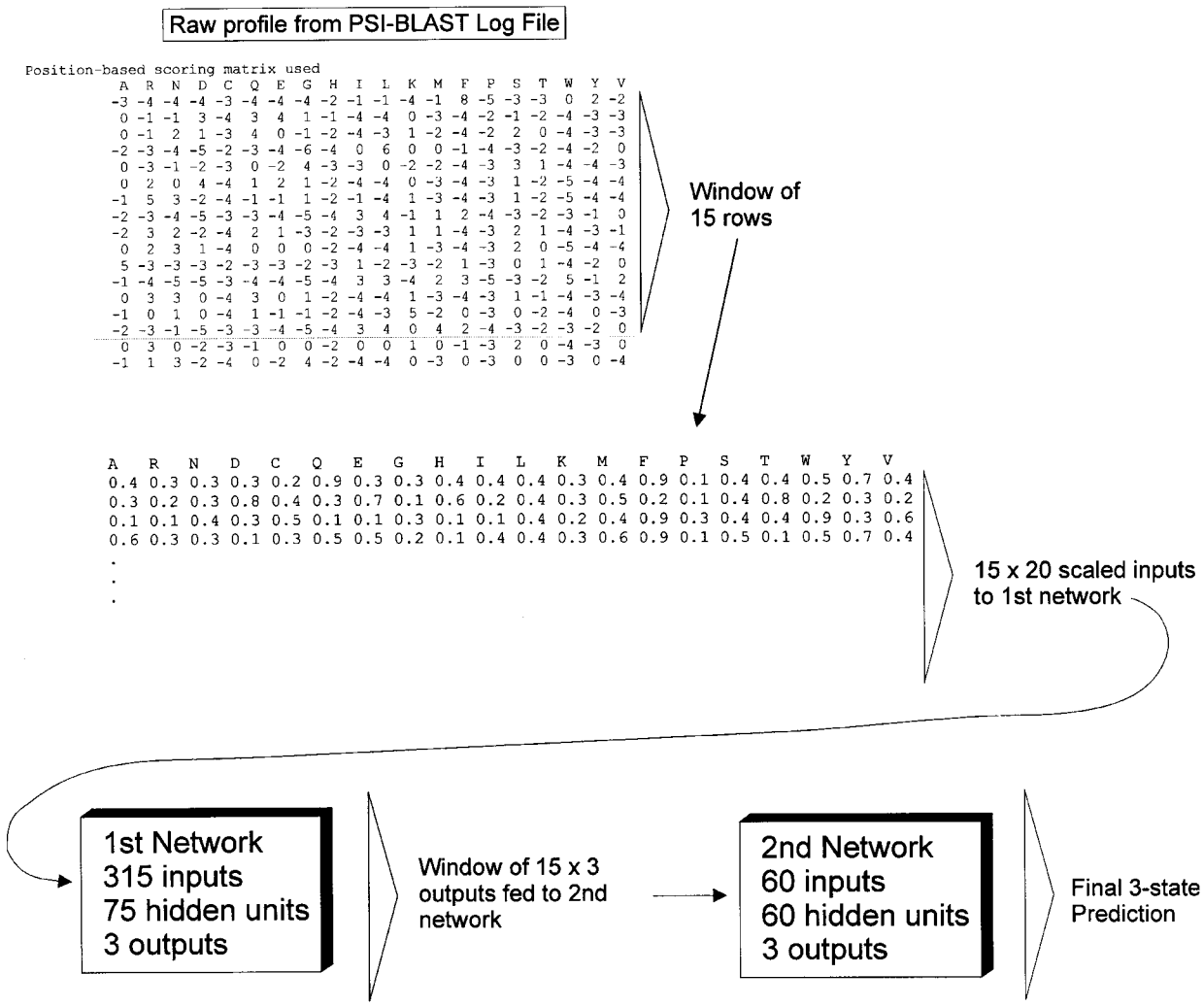

Figure 1. An outline of the PSIPRED method, which shows how the PSI-BLAST score matrices are processed.

et al., 1994), and regions which are likely to form coiled-coil structures.

The final position-specific scoring matrix (log-odds values) from PSI-BLAST (after three iterations) is used as input to the neural network. To obtain this scoring matrix, PSI-BLAST was compiled using the appropriate ''debug'' flag in the POSIT module, which allows the matrix to be parsed easily from the resulting PSI-BLAST log file. This matrix has $20 \times M$ elements, where M is the length of the target sequence, and each element represents the log-likelihood of that particular residue substitution at that position in the template (based on a weighted average of BLOSUM62 matrix scores for the given alignment position). Depending on the coverage of the hits obtained, different parts of this profile may be based on multiple sequences or just the query sequence itself (in which case the profile elements are identical to the appropriate row or column in the BLOSUM62 matrix). PSI-BLAST uses a simple but effective scheme for weighting the contribution of locally different numbers of sequences to the resulting profiles, and here no attempt was made to further adjust for such biases. The profile matrix elements (typically in the range ± seven) are scaled to the required 0-1 range by using the standard logistic function:

$$\frac{1}{1+e^{-x}}$$

where x is the raw profile matrix value. This scaling could also have been achieved by adapting the input units directly to accept input in the given range.

Neural network architecture

A standard feed-forward back-propagation network architecture (Rumelhart *et al.*, 1986) with a single hidden layer was used for PSIPRED. Although no serious attempt was made to search through the many different possible network topologies (different numbers of input units and different numbers of hidden units), a few alternative architectures were tried, and a set of 16 prediction targets from the second CASP experiment (Moult *et al.*, 1997) was used as a limited benchmark. A

window of 15 amino acid residues was found to be optimal (producing an overall Q_3 score of 80.1 %), and thus the final input layer comprises 315 input units, divided into 15 groups of 21 units. The extra unit per amino acid is used to indicate where the window spans either the N or C terminus of the protein chain. A large hidden layer of 75 units was used, with another three units making the output layer where the units represent the three-states of secondary structure (helix, strand or coil).

As with previous neural network secondary structure prediction methods (Rost & Sander, 1993), a second network is used to filter successive outputs from the main network. As only three possible inputs are necessary for each amino acid position, this network has an input layer comprising just 60 input units, divided into 15 groups of four. Again the extra input in each group is used to indicate that the window spans a chain terminus. For this network, a smaller hidden layer of 60 units was used.

Neural network training

An on-line back-propagation training procedure was used to optimise the weights in the network, i.e. the weights in the network were updated after each pattern presentation, though with a momentum term to prevent oscillation. A momentum term of 0.9 and a learning rate of 0.005 was found to be effective. To prevent over-training of the network, 10 % of the training set was kept aside to evaluate the performance of the network during training. This subset of the training set was not used to calculated the weight changes in the network. Training of the network was halted when the performance of the network on the excluded 10 % of the training data began to degrade.

Testing procedure

Correct evaluation of a secondary structure prediction method requires a properly cross-validated testing procedure. It has been known for a long time that poor cross-validation can produce overly favourable results (see Cuff & Barton (1999) for a detailed discussion), and so it was decided to do a very thorough cross-validation experiment to evaluate the current method. Up until now, secondary structure methods have been tested with training and test sets screened for significant sequence similarity. However, as pointed out by Cuff & Barton (1999), using pairwise sequence alignment methods, weak but significant sequence similarity between members of the two sequence sets can remain. Here we avoid this complication by screening our testing and training sets using a structural similarity criterion. Rather than removing from the training set any protein with a significant degree of sequence similarity to any member of the testing set, we removed any protein with a similar fold to any member of the testing set. We believe that this represents the most stringent cross-validation test that is possible, and avoids the complexities of sensitive sequence comparison.

To produce the training and test sets, each pair of proteins from the testing set and training set was evaluated with respect to the CATH classifications (Orengo *et al.*, 1997) for their constituent domains. Any protein in the training set which had a domain fold in common with any of the domains found in the test set was excluded. A further check was performed using five iterations of PSI-BLAST to detect any missed remote relationships that might not be represented in the CATH classification scheme, but no such pairs were detected, and there appears to be no detectable overlap at all between the training and test sets. It is also worth noting that a PSI-BLAST check is now included in the processing of the current CATH database (C. Orengo, personal communication) to identify distantly related but structurally divergent proteins, and so the chance of any homologous proteins being found in both the training and test sets is negligible. Using this scheme, three independent training and testing set pairs were compiled. The testing sets were based on the CATH T-level, and so comprised a set of unique protein folds. However, only highly resolved structures (resolution <1.8 Å) were included in the final set, giving a total of 187 protein chains in the testing set, divided into three sets of 62, 62 and 63 chains. Note that none of the 16 proteins (or any homologues) used for the limited search for the best network architecture were present in this test set.

The reference secondary structure states (helix, strand and coil states) for each structure in the training and test sets were derived from the definitions produced by DSSP (Kabsch & Sander, 1983). The eight states (H, I, G, E, B, S, T, −) were reduced to three states according to the scheme outlined by Rost & Sander (1993), i.e. H and G are taken to be helix states, E and B are taken to be strand states, and all others considered to be coil. To estimate a higher bound on the expected accuracy, a simpler mapping scheme was also tried where only H states in DSSP are mapped to helix, and E states mapped to strand.

Results

Figure 2(a) and (b) shows the distributions of Q_3 scores and Sov_3 scores (Rost *et al.*, 1994) for the testing set of 187 protein chains. Note that the average Q_3 score for these 187 proteins, calculated by chain, is found to be 76.0 % with a standard deviation of 7.8 %. The average Sov_3 score was 73.5 % with a standard deviation of 12.7 %. Taken by residue (i.e. averaging with weighting by sequence length), the average Q_3 score is 76.5 %. Using the simpler DSSP mapping, which results in a higher proportion of coil states, the by-residue Q_{3s} score was found to be as high as 78.3 %. These results indicate that the method described here, despite the very stringent cross-validation strategy,

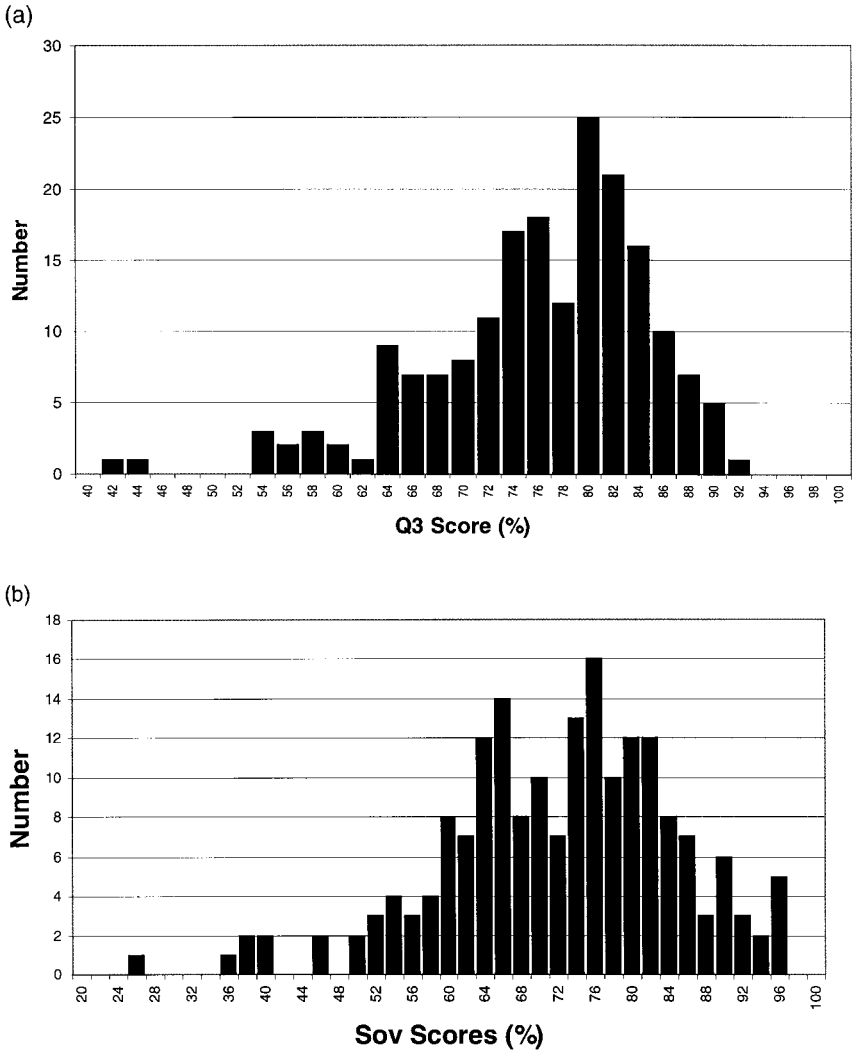

Figure 2. (a) Bar graph showing the distribution of Q_3 scores for the benchmark set of 187 protein chains with unique folds. (b) Bar graph showing the distribution of Sov_3 scores for the same 187 protein chains.

is at the very top of the range of accuracies documented for secondary structure prediction methods.

Despite the impressive results shown in Figure 2, there always remain nagging doubts about the possibility of some bias remaining from a knowledge of the experimentally determined structures. For example, the architecture of the network and training protocol could be optimised for any given given testing set, so that no matter how rigorously the predictions have been evaluated by cross-validated testing, the results are still better than might be expected for newly characterised proteins. The CASP (Moult *et al.*, 1997) experiment which has been run every two years since 1994 offers a means to evaluate available prediction methods entirely blindly. Although it is possible to criticise the CASP experiment on the basis of small sample sizes, it does act as a very useful adjunct to benchmarking procedures, and offers a ''level playing field'' so that methods can be fairly compared against each other. In view of this, it was vital to participate in the third CASP experiment (CASP3) with the method described here to see if it really was able to make more accurate predictions than the existing popular methods.

Table 1 summarises the results for the predictions which were submitted to the CASP3 prediction server and which were evaluated by the independent assessors. The raw data for this table was extracted from the public CASP3 Web page: http://predictioncenter.llnl.gov/casp3. For the purposes of discussion during the CASP3 meeting held at Asilomar, the assessors decided to restrict their own evaluation to the hardest targets (i.e. those which were most poorly predicted on average by all the groups). The average Q_3 and Sov_3 scores for PSIPRED when evaluated on these targets were 73.4 % and 71.9 %, respectively. On the same targets, the next best method (K. Karplus, unpublished results) achieved an average by residue Q_3 score of 69.0 %, and an average Sov_3 score of 65.7 %. For reference, the widely used PHD method achieved a Q_3 score of 66.7 % and a Sov3 score of 63.8 % on these targets. It is important to point out, however, that the PHD results evaluated at CASP3 were not provided by the authors, but were instead provided independently by the

Table 1. Complete set of PSIPRED prediction results for the 21 CASP3 targets for which a prediction was submitted

Target	$Q_{3(a)}$	$Q_{3(b)}$	Sov_3	N_{seq}	Length
T0043	78.5	79.1	72.8	15	158
T0044	73.1	74.3	78.5	12	335
T0046	71.4	73.9	67.9	7	119
T0053	80.2	80.9	82.3	2	257
T0056	78.9	82.5	85.1	16	114
T0059	73.2	76.1	78.7	40	71
T0061	55.3	61.8	65.6	1	76
T0063	70.4	75.6	71.6	45	135
T0064	91.3	91.3	88.1	119	104
T0065	96.8	96.8	100.0	2	31
T0067	83.2	88.6	81.6	29	185
T0068	71.5	71.5	72.2	56	376
T0071	71.3	70.9	73.7	7	237
T0074	94.7	95.8	98.9	263	95
T0075	74.5	77.3	72.9	31	110
T0077	79.8	79.8	76.0	45	104
T0079	93.1	93.1	99.5	123	116
T0080	70.3	76.7	69.3	120	202
T0081	66.2	70.9	70.8	14	151
T0083	85.3	85.3	85.9	5	156
T0085	64.0	65.9	60.7	1	211
Mean (by chain)	77.3	79.4	78.7		
Mean (by residue)	75.7	77.6			

Two Q_3 scores are shown: (a) where DSSP states HG are considered to be alpha helices and EB states are considered to be strands and (b) where only H states are considered to be alpha helices and E states are strands. Sov_3 scores are also shown, along with the length of the target sequence and the number of sequences included in the final PSI-BLAST profile (N_{seq}).

JPRED server (Cuff & Barton, 1998), and so some improvement in the PHD results might be anticipated if a more recent implementation of the program was tested. However, the margin of improvement for PSIPRED over PHD at CASP3 was not trivial, and so the CASP3 results and the benchmark results presented here are probably a fair reflection of at least the current rankings of available secondary prediction methods.

Figure 3 shows the case by case comparison of the Q_3 score from the PSIPRED CASP3 predictions with predictions from JPRED (Cuff & Barton, 1999), which incorporates PHD results in its consensus prediction, and another popular method, DSC (King & Sternberg, 1996). The average Q_3 scores for the three methods over these targets are 76.3 % (PSIPRED), 72.4 % (JPRED) and 67.3 % (DSC over 16 targets). Furthermore, in all but one case, PSIPRED achieved an accuracy of at least 60 %, and produced no predictions with an accuracy below 55 %. A total of 17 of the 23 PSIPRED predictions had an accuracy of at least 70 % compared to 12 out of 23 for JPRED. Despite the higher overall performance of PSIPRED over the other methods, in seven cases one of the other prediction methods produced a more accurate prediction.

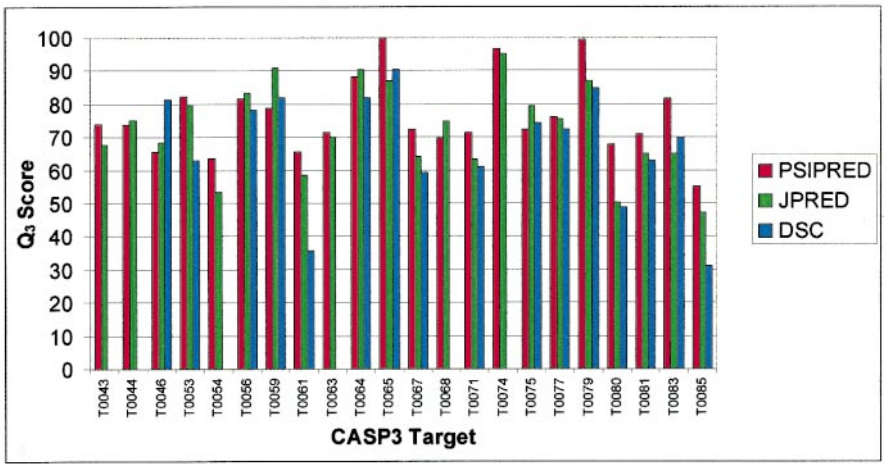

Figure 3. Bar graph showing a comparison of prediction results for 22 CASP3 targets by PSIPRED and two other popular methods: JPRED (including PHD) and DSC.

This suggests that there might be some scope for improving the prediction accuracy of PSIPRED by calculating a consensus prediction with other methods, as is currently done by the JPRED method.

Conclusions

At this stage it is not yet clear which factors contribute most to the success of the PSIPRED method, and work is currently underway to compare the results obtained from PSIPRED with those obtained from other methods, but using the same input profiles. There are three aspects of the PSI-BLAST program that no doubt contribute, perhaps equally, to the success of PSIPRED. Firstly the alignments produced by PSI-BLAST are based on pairwise local alignments. Previous work (Frishman & Argos, 1997; Salamov & Solovyev, 1997) has suggested that the use of reliable local alignments produces a definite improvement in the accuracy of resulting secondary structure predictions. Secondly, the use of iterated profiles greatly enhances the sensitivity of PSI-BLAST. It has been shown that PSI-BLAST can identify twice as many pairwise relationships than an equivalent pairwise comparison method. Thirdly, in my own laboratory (M. Tress, unpublished results) we have found that the accuracy of PSI-BLAST alignments (when compared to alignments based on structure comparison) are significantly higher than any other method we have tried for automatic multiple sequence alignment (though again this relates to the effect of reliable local alignments).

Perhaps the most significant conclusion that can be reached from the presented results, is that a very simple method for secondary prediction based on a straightforward neural network evaluation of PSI-BLAST-generated profiles is capable of producing results which rank the method at the very top of the current crop of prediction methods. The ideas described can be easily applied to other prediction schemes, and it might well be expected that by using PSI-BLAST profiles, as opposed to profiles generated from traditional multiple sequence alignment approaches, other secondary structure prediction methods will show measurable improvements in accuracy.

Availability

The PSIPRED Web server, along with the software and test sets used here may be obtained electronically from the following address: http://globin.bio.warwick.ac.uk/psipred.

Acknowledgements

This work was supported by The Royal Society.

References

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

Benner, S. A. & Gerloff, D. (1990). Patterns of divergence in homologous proteins as indicators of secondary and tertiary structure: a prediction of the structure of the catalytic domain of protein kinases. *Advan. Enzyme Reg.* **31**, 121-181.

Bowie, J. U., Luethy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science,* **253**, 164-170.

Chou, P. Y. & Fasman, G. D. (1974). Conformational parameters for amino acids in helical, -sheet, and random coil regions calculated from proteins. *Biochemistry,* **13**, 211-222.

Cuff, J. A. & Barton, G. J. (1999). Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **34**, 508-519.

Frishman, D. & Argos, P. (1996). Incorporation of non-local interactions in protein secondary structure prediction from the amino acid sequence. *Protein Eng.* **9**, 133-142.

Frishman, D. & Argos, P. (1997). Seventy-five percent accuracy in protein secondary structure prediction. *Proteins: Struct. Funct. Genet.* **27**, 329-335.

Garnier, J., Osguthorpe, D. J. & Robson, B. (1978). Analysis and implications of simple methods for predicting the secondary structure of globular proteins. *J. Mol. Biol.* **120**, 97-120.

Geourjon, C. & Deleage, G. (1995). SOPMA: significant improvements in protein secondary structure prediction by consensus prediction from multiple alignments. *Comp. Appl. Biosci.* **11**, 681-684.

Jones, D. T. (1997). Successful *ab initio* prediction of the tertiary structure of NK-lysin using multiple sequences and recognized supersecondary structural motifs. *Proteins: Struct. Funct. Genet.,* **S1**, 185-191.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1992). A new approach to protein fold recognition. *Nature,* **358**, 86-89.

Jones, D. T., Taylor, W. R. & Thornton, J. M. (1994). A model recognition approach to the prediction of all-helical membrane protein structure and topology. *Biochemistry,* **33**, 3038-3049.

Kabsch, W. & Sander, C. (1983). A dictionary of protein secondary structure. *Biopolymers,* **22**, 2577-2637.

King, R. D. & Sternberg, M. J. E. (1996). Identification and application of the concepts important for accurate and reliable protein secondary structure prediction. *Protein Sci.* **5**, 2298-2310.

Lemer, C., Rooman, M. J. & Wodak, S. J. (1995). Protein structure prediction by threading methods: evaluation of current techniques. *Proteins: Struct. Funct. Genet.* **23**, 337-355.

Lim, V. I. (1974). Algorithms for prediction of alpha helices and structural regions in globular proteins. *J. Mol. Biol.* **88**, 873-894.

Moult, J., Hubbard, T., Bryant, S. H., Fidelis, K. & Pedersen, J. T. (1997). Critical assessment of methods of protein structure prediction (CASP): round II. *Proteins: Struct. Funct. Genet.* **S1**, 2-6.

Niermann, T., Kirschner, K. & Crawford, I. P. (1987). Prediction of tertiary structure of the alpha-subunit

of tryptophan synthase. *Biol. Chem. Hoppe-Seyler,* **368**, 1087-1088.
Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). CATH-a hierarchic classification of protein domain structures. *Structure,* **5**, 1093-1108.
Rost, B. (1997). Protein fold recognition by prediction-based threading. *J. Mol. Biol.* **270**, 1-10.
Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70 % accuracy. *J. Mol. Biol.* **232**, 584-599.
Rost, B., Sander, C. & Schneider, R. (1994). Redefining the goal of protein secondary structure prediction. *J. Mol. Biol.* **235**, 13-26.
Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature,* **323**, 533-536.
Russell, R. B., Copley, R. R. & Barton, G. J. (1996). Protein fold recognition by mapping predicted secondary structures. *J. Mol. Biol.* **259**, 349-365.
Salamov, A. A. & Solovyev, V. V. (1995). Prediction of protein secondary structure by combining nearest- neighbor algorithms and multiple sequence alignments. *J. Mol. Biol.* **247**, 11-15.
Salamov, A. A. & Solovyev, V. V. (1997). Protein secondary structure prediction using local alignments. *J. Mol. Biol.* **268**, 31-36.
Sali, A. (1995). Modelling mutations and homologous proteins. *Curr. Opin. Biotechnol.* **6**, 437-451.
Smith, T. F. & Waterman, M. S. (1981). Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195-197.
Wootton, J. C. & Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Comput. Chem.* **17**, 149-163.
Zvelebil, M. J. J. M., Barton, G. J., Taylor, W. R. & Sternberg, M. J. E. (1987). Prediction of protein secondary structure and active sites using the alignment of homologous sequences. *J. Mol. Biol.* **195**, 957-961.

Edited by G. Von Heijne

(Received 7 May 1999; received in revised form 29 July 1999; accepted 29 July 1999)