PROTEINS: Structure, Function, and Genetics 41:17–20 (2000)

Prediction of Protein Secondary Structure at 80% Accuracy

Thomas Nordahl Petersen,[1]* Claus Lundegaard,[1] Morten Nielsen,[1] Henrik Bohr,[2] Jakob Bohr,[2] Søren Brunak,[2] Garry P. Gippert,[1] and Ole Lund[1]

[1]*Structural Bioinformatics Advanced Technologies A/S, Hørsholm, Denmark*
[2]*Structural Bioinformatics, Inc., SAB, San Diego, California*

ABSTRACT **Secondary structure prediction involving up to 800 neural network predictions has been developed, by use of novel methods such as output expansion and a unique balloting procedure. An overall performance of 77.2%–80.2% (77.9%–80.6% mean per-chain) for three-state (helix, strand, coil) prediction was obtained when evaluated on a commonly used set of 126 protein chains. The method uses profiles made by position-specific scoring matrices as input, while at the output level it predicts on three consecutive residues simultaneously. The predictions arise from tenfold, cross validated training and testing of 1032 protein sequences, using a scheme with primary structure neural networks followed by structure filtering neural networks. With respect to blind prediction, this work is preliminary and awaits evaluation by CASP4. Proteins 2000;41:17–20.** © 2000 Wiley-Liss, Inc.

Key words: sequence profiles, multiple predictions, neural networks, confidence, output expansion, balloting, cross validation.

INTRODUCTION

Prediction of protein tertiary structure from the amino-acid sequence remains one of the biggest challenges in structural biology. One step toward solving this problem is by increasing the accuracy of secondary structure predictions for subsequent use as input to ab initio calculations or threading algorithms. The predictions are of significant importance to protein fold recognition and homology modeling, and the degree to which the secondary structure can be determined has become a benchmark for protein structure prediction. Studies have shown that an increased performance in secondary structure prediction can be obtained by combining several estimators, e.g., neural networks.[1,2] A combination of up to eight neural networks has been shown to increase the accuracy, but a saturation point was reached in the sense that adding more networks would not increase the performance substantially.[3] Here we report on an improved procedure for generating and combining up to 800 predictions of differently trained neural networks. Early methods for predicting protein secondary structure relied on the use of single protein sequence.[4–7] Several groups have shown that a significant increase in performance can be obtained by using sequence profiles[8–11,1] or position-specific scoring matrices[12] obtained using the PSI-BLAST program.[13] The data presented in this study also shows that a significant increase in performance can be obtained by a neural network algorithm where secondary structure prediction is performed on several consecutive residues simultaneously.

MATERIALS AND METHODS

Preparation of Data Set for Network Training and Evaluation

Data used to train the neural networks was prepared from the Protein Data Bank version of August, 1999.[14] An extensive quality check was applied to the PDB entries. For X-ray structures a resolution cutoff of 2.5 Å was used. For NMR ensembles, subsets of residues with mean Cα-Cα distance RMSD below 1.0 Å were selected. Prior to splitting candidate PDB entries into single chains, assignments of eight-category secondary structure classes were made using the DSSP program.[15] A minimum chain length of 30 residues was required, and chain breaks were not allowed. This reduction resulted in a set of 9926 protein chains. Sequence similarity reduction using the Hobohm algorithm #1,[16] as described previously,[17] reduced the set further to 1,168 chains. Transmembrane proteins were removed manually. Sequences with a high similarity to the RS126 set[1] were subsequently removed using the Hobohm algorithm #1, resulting in a train/test set of 1032 protein chains (TT1032 set) used for training of all neural networks. The 126 protein chains in the RS126 set were used only for final evaluation of network performance. Evaluation of the performance shown in Figure 1 (on the RS126 set) was done using sequences where sequence gaps corresponding to missing residues in the coordinate file were filled in using SWISS-PROT sequences.

Sequence Profiles

Sequence profiles were generated with the program PSI-BLAST.[13] Profiles were constructed from a log-odds substitution matrix after three BLAST iterations, provided such a matrix was produced by the program. Otherwise the profile was made from a BLOSUM62 matrix.[18] Homologous sequences were searched for in a nonredundant database of sequences extracted from SWISS-PROT and TREMBL.[19] This database was preprocessed such that residues in the protein sequences, annotated as RICH, COIL, REPEAT, HYDROPHOBIC, SIGNAL, or

*Correspondence to: Thomas Nordahl Petersen, Structural Bioinformatics Advanced Technologies A/S, Agern Allé 3, DK-2970 Hørsholm, Denmark. E-mail: tnordahl@strubix.dk

Received 24 February 2000; Acccepted 18 May 2000

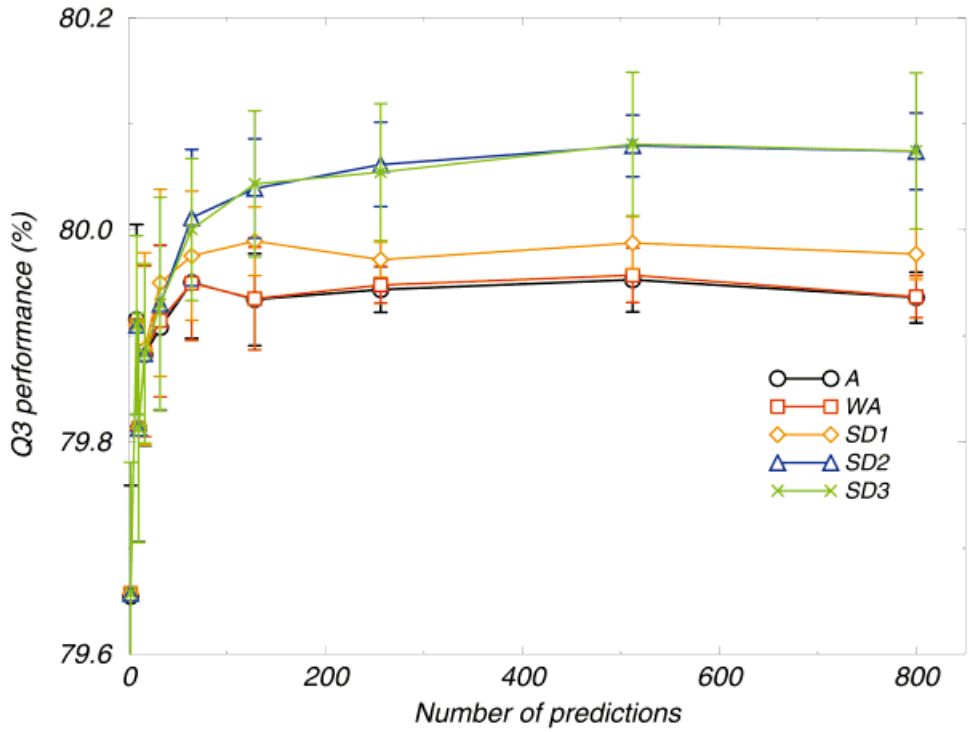

Fig. 1. The Q3 score vs. the number of prediction means N. Each point represents the mean and standard deviation of Q3 performance obtained for 10 random selections of N prediction means included in the balloting procedure. Five combination procedures are shown: Straight averaging (A), weighted average (WA), weighted average where only networks with a per-chain confidence of at least 1–3 standard deviations above the mean are used (SD1, SD2, and SD3). Mutiple uses of predictions were not excluded from the random selection procedure.

TRANSMEMBRANE, were substituted with an X in order to avoid matching unrelated sequences with PSI-BLAST.

Secondary Structure Class Reduction

The neural networks were trained to predict a three-category (HEC) secondary structure assignment which was reduced from the eight-category assignment produced by the DSSP program.[15] The secondary structure categories were cast into three categories using **either** the plain DSSP secondary structure definition[12]: H (helix) = {H}, E (strand) = {E}, and C (all remaining categories) = {.GIBST} or the definition: H (helix) = {HG}, E (strand) = {EB}, and C = {.IST}.

Neural Networks

A feed-forward neural network architecture was used, with one input layer, one hidden layer, and one output layer. Weights were updated using the conventional back-propagation procedure.[20] Several different input and hidden layer sizes were used. Sequence-to-structure neural networks consisted of input layers corresponding to 15, 17, 19, or 21 residues in a sequence window, in combination with a hidden layer consisting of 50 or 75 units, and nine output units corresponding to HEC class predictions for each of the three central residues *i-1*, *i*, and *i+1* in the input window. In addition to the sequence profile encoding, we used two neurons encoding the relative position in the protein sequence *(i/L)* and *1- (i/L)*, where *L* is the length of the protein chain, two neurons encoding the relative size of the chain *(L/Lmax)* and *1- (L/Lmax),* where *Lmax* is the length of the longest protein chain in the training set, and 20 neurons encoding the frequencies of the 20 amino-acids in the sequence.

Output Expansion

The simultaneous prediction of secondary structure class assignments for residue *i* and several residues in the immediate vicinity of residue *i* is called output expansion. Training with output expansion set to $\pm$ 1, results in nine output categories. The assignment of the central residue *i* in a window, becomes dependent on the three-state assignment of its neighbor residues at positions *i-1* and *i+1*, respectively. An example of the output expansion assignment scheme is shown in Table I.

Output from the sequence-to-structure network is fed to a structure-to-structure network with an input layer consisting of nine output activities from a network with output expansion in a sequence window size of 17, a hidden layer consisting of 40 units, and an output layer consisting of three units corresponding to HEC class predictions for residue *i*.

Network performance was measured using the Q3 score, which is the sum of correct predictions in three-class predictions divided by the total number of residues, and the Matthews correlation coefficients[21] for each secondary structure class.

TABLE I. Assignment Schemes Used for Neural Network Output Categories†

Primary structure	Assignment without output expansion	Assignment with output expansion
1 A	C	−CC
2 G	C	CCH
3 W	H	CHH
4 A	H	HHC
5 L	C	HCE
6 I	E	CE−

†Example of the assignment scheme used for a protein sequence with and without output expansion.

Network Training

The neural networks were trained and tested on the TT1032 data set using a tenfold, cross validation procedure, i.e., using 9/10th of the data (training set) to perform feed-forward and backpropagation, and the remaining 1/10th of the data (test set) to determine the stopping condition. The same partitioning of the training and test sets was used for sequence-to-structure and structure-to-structure networks, thus producing a total number of 100 combinations for each of the eight architectures used. A matrix representing an optimal transformation of network output activities into probabilities was produced for each of the 800 network combinations using the entire TT1032 set as input. Probability transformation matrices were made after training was complete, and were used thereafter to produce HEC probabilities for a query sequence here represented by the RS126 set.

Balloting Probabilities

The balloting procedure is a statistical method that enables an efficient combination of multiple predictions. The procedure consists of two steps: First per residue confidence values, α_{ijk}, are associated with each residue i in chain j for prediction k, as the highest minus the second highest of the three probabilities $P_{ijk}(H)$, $P_{ijk}(E)$, and $P_{ijk}(C)$. A mean confidence for prediction k on chain j is calculated:

$$\alpha_{jk} = 1/N_j \sum \alpha_{ijk}$$

where the sum is over all residues $i = 1. . .N_j$ in chain j. Furthermore a mean and standard deviation per chain confidence is calculated:

$$\langle \alpha_j \rangle = 1/N_k \sum \alpha_{jk}$$

$$\sigma_j = \sqrt{(\langle \alpha_j^2 \rangle - \langle \alpha_j \rangle^2)}$$

where the sum is over all predictions k. The probability P_{ij} (class) for residue i in chain j is calculated:

$$P_{ij}(\text{class}) = \sum \alpha_{jk} P_{ijk}(\text{class}) / \sum \alpha_{jk}$$

where class means H, E, or C, and the sum is over a subset of prediction sets k for which α_{jk} is greater than $\langle \alpha_j \rangle + \sigma_j$, with the constraint that at least 10 prediction sets k are included in the weighted average.

RESULTS

Output Expansion and Balloting Procedure

Predictions were made using neural networks trained on a set of 1032 high-quality protein chains non-sequence similar to the RS126 set.[1] The RS126 evaluation set was not used at any point during training and testing of the networks. A total of 800 prediction means were generated with diverse network architectures, and cross validated on test and training sets both as sequence-to-structure and structure-to-structure networks. Window sizes in the range of 15 to 21 amino acids were applied in combination with 50 and 75 hidden neurons. Using the plain secondary structure assignment, an overall three-state prediction performance (Q3) of 80.2% was measured for the evaluation set of 126 protein chains, with a mean per-chain performance of 80.6%. With output expansion, the percentage of correct predictions were 84.6%, 69.0%, and 82.2% with correlation coefficients of 0.778, 0.639, and 0.623 for the DSSP secondary structure categories H, E, and C, respectively. Without output expansion, the Q3 performance was 0.5% lower. Combining predictions using a balloting scheme increases performance significantly over straight-forward averaging (Fig. 1) and the performance continues to increase as more networks are included in the balloting process. Thus, it is possible to benefit even from suboptimal networks when predictions are combined efficiently. In the earlier work of Chardonia and Karplus,[3] extension to eight networks was indicated as an upper limit for increased performance. The reason for this is that predictions produced by suboptimal networks have a larger detrimental influence on the performance obtained by taking a simple average compared to the balloting procedure employed in the present study.

Networks were also trained and tested using the other secondary structure definition. With this definition a per residue Q3 performance of 77.2% (77.9% mean per-chain) was obtained with correlation coefficents of 0.733, 0.629, and 0.574, for H, E, and C, respectively.

Comparison With Other Secondary Structure Prediction Methods

In the recent CASP3 experiment[22] the PSI-PRED method[12] was shown to have superior performance on 23 out of 35 sequences included in the experiment. We therefore trained 10 sequence-to-structure and 10 structure-to-structure neural networks with the same architecture as used by Jones.[12] The average performance of these networks was 78.0% when tested on the RS126 set, using the plain DSSP secondary structure definition. This is comparable to the evaluation set performance of 78.3% reported in the study by Jones.[12]

DISCUSSION

We have found that an increase in the accuracy of secondary structure prediction can be obtained by combining many neural network predictions. Previously up to eight networks have been used[3] in part because it was too computationally demanding to include many network predictions, and in part because it seemed that a saturation

point had already been reached. Furthermore, we have also shown that an increase in performance can be obtained using a novel procedure called output expansion in which the secondary structure is predicted for a given residue and its neighbors simultaneously. The additional output units give hints to the neural networks thereby restraining the weights leading to improved generalization.

The PHD method developed by Rost and Sander performed best in the CASP2 experiment with a mean Q3 of 74%.[23] In a recent comparative study, the PHD method had the best Q3 (71.9%) of all individual methods tested, while a consensus method scored 72.9%.[2] In CASP3 the PSI-PRED method[12] performed best with Q3 performances of 73.4% and 74.6%, respectively, on the two small test sets used by the evaluators. The PSI-PRED method was approximately seven percentage points better than a version of the PHD method similar to the one used in CASP2.[22]

For a comparison to the PSI-PRED approach we have kept our training, test, and evaluation data sets identical, and only changed the methodological aspects in order to show how the improvement was obtained. The comparison clearly demonstrates that the 800 network scheme, with prediction on several residues simultaneously, outperforms the single network approach used in PSI-PRED. In his article, Jones[12] reports a Q3 performance of 76.5% using a CASP-like secondary structure definition, and a Q3 performance of 78.3% with the plain DSSP secondary structure definition identical to the one used in the first part of our study. In our hands an implementation of the Jones method produces a Q3 performance of 78.0% using the plain secondary structure definition. The comparison shows that there is only a small difference in performance using the PSI-PRED training/evaluation set redundancy reduction based on CATH fold classes[12] and the homology reduction used here. In our work we use a reduction scheme which is essentially identical to that used in the PHD approach.[24] The additional 2% gain in performance we observe using methods described here thus appears to be a significant improvement in secondary structure prediction performance, compared to that obtained using the PSI-PRED method, which was previously reported to be the best available.

ACKNOWLEDGMENTS

Dr. Kenneth Geisshirt is thanked for substantial computer system and database support.

REFERENCES

1. Rost B, Sander C. Prediction of protein secondary structure at better than 70 % accuracy. J Mol Biol 1993;323:584–599.
2. Cuff JA, Barton GJ. Evaluation and improvement of multiple sequence methods for protein secondary structure prediction. Proteins 1999;34:508–519.
3. Chandonia J-M, Karplus M. New methods for accurate prediction of protein secondary structure. Proteins 1999;35:293–306.
4. Chou PY, Fasman GD. Conformational parameters for amino acids in helical, sheet and random coil regions, calculated from proteins. Biochemistry 1974;13:211–222.
5. Garnier J, Osguthorpe DJ, Robinson B. Analysis and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 1978;120:97–120.
6. Qian N, Sejnowski TJ. Predicting the secondary structure of globular proteins using neural network models. J Mol Biol 1988; 202:865–884.
7. Bohr H, Bohr J, Brunak S, et al. Protein secondary structure and homology by neural networks. The alpha-helices in rhodopsin. FEBS Lett 1988;241:223–228.
8. Gribskov M, McLachlan AD, Eisenberg D. Profile analysis: detection of distantly related proteins. Proc Natl Acad Sci 1987;84:4355–4358.
9. Zvelebil MJ, Barton GJ, Taylor WR, Sternberg MJ. Prediction of protein secondary structure and active sites using the alignment of homologous sequences. J Mol Biol 1987;195:957–961.
10. Crawford IP, Niermann T, Kirschner K. Prediction of secondary structure by evolutionary comparison: application to the alpha subunit of tryptophan synthase. Proteins 1987;2:118–129.
11. Benner SA, Gerloff D. 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. Adv Enzyme Regul 1991;31:121–181.
12. Jones DT. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 1999;292:95–202.
13. Altshul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
14. Bernstein FC, Koetzle TF, Williams GJB, et al. The protein data bank: a computer based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
15. Kabsch W, Sander C. A dictionary of protein secondary structure. Biopolymers 1983;22:2577–2637.
16. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
17. Lund O, Frimand K, Gorodkin J, et al. Protein distance constraints predicted by neural networks and probability density functions. Protein Eng 1997;10:1241–1248.
18. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
19. Bairoch A, Apweiler R. The SWISS-PROT protein sequence data bank and its new supplement TREMBL. Nucleic Acids Res 1996;24:21–25.
20. Rumelhart D, Hinton G, Williams, R. Learning internal representations by error propagation. In: Rumelhart D, McClelland J, editors. Parallel distributed processing. Cambridge, MA: MIT Press; 1986. p 1:318–363.
21. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
22. Orengo CA, Bray JE, Hubbard T, LoConte L, Sillitoe I. Analysis and assessment of ab initio three-dimensional prediction, secondary structure, and contacts prediction. Proteins 1999;Suppl 3:149–170.
23. Lesk AM. CASP2: report on ab initio predictions. Proteins 1997; Suppl 1:151–166.
24. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.