Protein Science (1996), *5*:768–774. Cambridge University Press. Printed in the USA.

The importance of larger data sets for protein secondary structure prediction with neural networks

JOHN-MARC CHANDONIA' AND MARTIN KARPLUS[1,2,3]

[1] Biophysics Program, Harvard University, Cambridge, Massachusetts
[2] Department of Chemistry, FAS, Harvard University, Cambridge, Massachusetts
[3] Laboratoire de Chimie Biophysique, Institut Le Bel, Université Louis Pasteur, 67000 Strasbourg, France

(RECEIVED September 14, 1995; ACCEPTED January 23, 1996)

Abstract

A neural network algorithm is applied to secondary structure and structural class prediction for a database of 318 nonhomologous protein chains. Significant improvement in accuracy is obtained as compared with performance on smaller databases. A systematic study of the effects of network topology shows that, for the larger database, better results are obtained with more units in the hidden layer. In a **32-fold** cross validated test, secondary structure prediction accuracy is 67.0%, relative to 62.6% obtained previously, without any evolutionary information on the sequence. Introduction of sequence profiles increases this value to 72.9%, suggesting that the two types of information are essentially independent. Tertiary structural class is predicted with 80.2% accuracy, relative to 73.9% obtained previously. The use of a larger database is facilitated by the introduction of a scaled conjugate gradient algorithm for optimizing the neural network. This algorithm is about 10–20 times as fast as the standard steepest descent algorithm.

Keywords: neural networks; secondary structure prediction; structural class prediction

Secondary structure prediction is one element in understanding how the amino acid sequence of a protein determines its native state. **A** number of attempts have been made to develop tertiary folding algorithms based on a knowledge of the secondary structure (Gunn et al., 1994; Monge et al., 1994). Current methods of secondary structure prediction usually assign one of three states (helix, sheet, or coil) to each residue of a protein. Rules for deriving a prediction from the identities of the surrounding residues have been based on a variety of algorithms. In recent years, neural networks (Rumelhart et al., 1986) have been applied to this problem (reviewed in Sumpter et al., 1994; Barton, 1995). **A** database of known structures (the "training set") is used to train the network. It is then applied to a test set of structures to evaluate its accuracy. Homology between proteins in the two data sets may lead to false indications of greater accuracy (Qian & Sejnowski, 1988). On a typical database, without homology between training and test sets, methods that consider single protein chains can produce a three-state accuracy of 62–63% (Qian & Sejnowski, 1988; Holley & Karplus, 1989, 1991; Chandonia & Karplus, 1995). Accuracy can be improved to 72.5% by training and testing networks on groups of aligned, homologous sequences, rather than on single chains (Rost & Sander, 1994).

Basic information on tertiary structure, such as the folding class (Levitt & Chothia, 1976), can also be predicted using neural network methods. Using characteristics such as sequence length and amino acid composition, a network can assign proteins to one of four broad classes (all-alpha, all-beta, alpha/beta, or other) with 62.3% accuracy (Chandonia & Karplus, 1995). A two-step approach, in which the secondary structure of proteins is first predicted, and then used as additional input to the class prediction network, improves the success rate to 73.9%. Furthermore, such an approach can eliminate one or more classes as possible candidates for a given sequence with near-perfect accuracy. The latter result can be used to set up training sets for secondary structure prediction that more closely match the expected class of the predicted protein, resulting in a slight gain in accuracy (Chandonia & Karplus, 1995).

It has been shown that the particular proteins chosen for the training and test sets can lead to large variation in the results (Zhang et al., 1992; Rost & Sander, 1993a; Chandonia & Karplus, 1995), even in the absence of homology between the sets. Therefore, unbiased results require some method of multiple cross validation; i.e., splitting up the data set into many discrete groups, then testing each group individually (using proteins from the remaining groups as a training set) and averaging the results. This considerably increases the overall calculation time.

In this paper, we apply a single-sequence secondary structure and class prediction neural network algorithm to a recently derived database containing more than five times as many se-

Reprint requests to: Martin Karplus, Department of Chemistry, 12 Oxford Street, Cambridge, Massachusetts 02138; e-mail: marci@tammy.harvard.edu.

quences and residues as that employed previously. To reduce computational time required to cross validate results on this database, a new neural network training method, developed by Møller (1993) and based on conjugate gradient minimization, is used. We describe the methods used in the calculations, and present results and a concluding discussion. We comment on the accuracy limit of such database approaches to secondary structure and class prediction algorithms and its relation to the protein folding mechanism.

Methods

Neural networks

The neural networks are standard feed-forward networks consisting of two or three layers of units (Rumelhart et al., 1986; Chandonia & Karplus, 1995). They are fully connected from one layer to the next. The first and last layers are referred to as the input and output layers, respectively. The middle layer, if present, is referred to as the hidden layer, because its inputs and outputs connect only to other network units. A detailed description of the secondary structure and tertiary class prediction networks may be found in Chandonia and Karplus (1995). The shorthand method for describing neural network topology is also used here. Secondary structure prediction networks are described by three numbers: the width of the input window (in residues), the size of the hidden layer, and the number of units in the output layer. The notation, a "15 X 8 X 2 network" means that a window of 15 consecutive residues (using an input layer of $15*21$, or 315 units), a hidden layer of 8 units, and an output layer of 2 units are being used. Class prediction networks are described simply in terms of the number of units in each layer. For example, a 26 X 8 X 4 network has 26 input units, 8 hidden units, and 4 output units.

Training a network involves minimizing an error function, which is a multivariate function of network weights. If $\mathbf{w}$ is a vector of the weights ($\mathbf{w} = [w_0, w_1, \ldots w_N]^T$), most training strategies consist of picking a random initial vector $\mathbf{w}_0$, and updating it in a series of steps until the error function $E(\mathbf{w})$ is close to zero. In each step, the weights are adjusted by picking a search vector $\mathbf{p}_k$ and a step size a_k, and setting

$$\mathbf{w}_{k+1} = \mathbf{w}_k + a_k\mathbf{p}_k. \quad (1)$$

In feed-forward networks, the first derivative of the error function with respect to any of the weights can be computed with a single loop over the training set data (Rumelhart et al., 1986). The standard back propagation algorithm (BP) (Rumelhart et al., 1986) is called a "steepest descent" algorithm because $\mathbf{p}_k$ is set to the downward gradient $-E'(\mathbf{w})$, whereas a_k is fixed to a constant supplied by the user. **A** variation of BP used in many secondary structure prediction networks (Rost & Sander, 1994; Chandonia & Karplus, 1995) also includes a "momentum" term ($m * \mathbf{p}_{k-1}$), but this adds little to the speed of the algorithm while requiring a second user-supplied constant (Møller, 1993).

A family of algorithms known as "conjugate gradient" (CG) algorithms uses a set of recursively determined conjugate vectors $\mathbf{p}_{0..N}$, with the $a_{0..N}$ values chosen so to minimize the error $E(\mathbf{w}_{k+1})$ along the line of possible $\mathbf{w}_{k+1}$ values. The procedure for finding the optimal a_k value along the line $\mathbf{w}_k + a_k\mathbf{p}_k$ is called the "line search," and is often the most time-consuming step. The initial direction $\mathbf{p}_0$ is set to the gradient $-E'(\mathbf{w})$, and subsequent search vectors $\mathbf{p}_k$ are set to the component of the gradient in a direction conjugate to all the previous vectors $\mathbf{p}_{0..k-1}$. For quadratic error functions, this algorithm is very efficient, requiring a number of steps equal to the number of weights in the network to find the local minimum. For nonquadratic error functions, the procedure must be reset at least once every N (the number of weights) steps, with $\mathbf{p}$ set to $-E'(\mathbf{w})$, to avoid running out of possible conjugate vectors. Although neural networks usually use nonquadratic error functions, CG-based algorithms are still an order of magnitude more efficient than the standard steepest descent procedure (Møller, 1993).

Networks in this study were trained using the recently developed scaled conjugate gradient (SCG) algorithm (Møller, 1993). SCG is slightly faster than other CG based algorithms, because it uses an approximation to find the optimal step size, a_k, rather than doing a time-consuming line search with each step. Furthermore, SCG does not require a direct computation of the Hessian matrix $E''(\mathbf{w})$, but only of the gradient $E'(\mathbf{w})$; in feed-forward neural networks, this can be computed in two passes through the training data (Rumelhart et al., 1986). In tests on the parity problem (described in Rumelhart et al., 1986), SCG has been shown to be an order of magnitude faster than BP, in addition to scaling better with the addition of more weights (networks with 16–100 weights were used in the study). Also, SCG converged to the exact solution more often than BP, and requires no user-supplied parameters (Møller, 1993).

Data set

The proteins used in this study were a set of 318 chains representative of high-resolution structures available in the Brookhaven Protein Data Bank (PDB) in early 1994. This database was prepared by Andrej Sali using the MOLSCRIPT program (Sali & Overington, 1994). First, protein chains from all well-resolved structures in the PDB were classified into groups according to sequence homology. The structure with the highest resolution in each group was taken to represent that group. All structures had been determined by X-ray crystallography to a resolution of 2.3 A or better; some chains for which only NMR-determined structures were available were also used. No pair of protein chains contained more than 30% identical residues.

The program DSSP (Kabsch & Sander, 1983b) was used to classify the secondary structure of all residues in the database. **All** residues that were neither alpha helix (H) or extended beta sheet (E) were considered to be in the "coil" category. The complete database contains a total of 56,966 residues with a composition of 30% helix, 21% sheet, and 49% coil; 3_{10}-helices were treated as coil.

Multiple cross validation trials are necessary to minimize variation in results caused by a particular choice of training or test sets (Rost & Sander, 1993a; Chandonia & Karplus, 1995). Because of the size of the database, jackknife cross validation (individual testing of each protein in the database) was not feasible. Instead, the database was divided into 32 groups, each containing several protein chains (31 sets of 10 chains, 1 set of 8). Networks were trained on sets produced by removing one group of proteins at a time from the database of 318. Each network was then tested on the excluded group of proteins and the results were combined for evaluation of overall prediction accuracy.

Structural class prediction

Quantitative definitions for a four-class system have been proposed by Kneller et al. (1990), based on examination of typical proteins from the Levitt and Chothia (1976) classes. Details are described in Chandonia and Karplus (1995). The data set contained 52 chains in the all-a class, 48 all-β chains, 84 α/β chains, and 134 other chains. All-a proteins contained an average of 59% helix, 2% sheet, and 39% coil. All-β proteins averaged 3% helix, 32% sheet, and **55%** coil. Proteins in the α/β class averaged 34% helix, 20% sheet, and 46% coil. Other proteins averaged 20% helix, 23% sheet, and 56% coil, and were too small or contained insufficient helix and sheet to fit any of the other classes.

Measurements of accuracy

Three-way percent accuracy (Q_3) and correlation coefficients (Matthews, 1975) for prediction of helix (C_H), sheet (C_E), and coil (C_C) were used to evaluate accuracy; details are given in Chandonia and Karplus (1995). For evaluating the accuracy of structural class prediction, correlation coefficients for each class (C_A, C_B, $C_{A/B}$, and C_O) were used; the overall four-way percent accuracy (Q_4) was also calculated for the 4-output network.

Smoothing

The smoothing algorithm is a method in which the raw network outputs for each residue are averaged with those of the immediately adjacent residues; details are given in Chandonia and Karplus (1995).

Results

Secondary structure prediction

Because our previous study had shown a 19 x 2 x 2 network to be optimal for secondary structure prediction on a smaller data set (Chandonia & Karplus, 1995), we first tested this method on the larger database. As in the previous study, one round of smoothing was used. To test for effects of overtraining on prediction accuracy, we trained our secondary structure networks for 1,000 steps using the SCG algorithm, pausing every 5 steps to evaluate accuracy on the prediction set. Results from the 32 cross validation trials were combined using a weighted average based on the number of residues in each of the prediction sets.

There are several notable differences between the results of this test and the results produced by the same topology network on a smaller database (Chandonia & Karplus, 1995). First, it is clear from the rapid rise in both training and prediction set accuracy that the SCG training method is much more efficient than the BP algorithm used previously. Accuracy on both the training and test sets reach nearly optimal levels within 100 steps, and remain nearly constant during the last several hundred steps. Also, whereas accuracy on the prediction set was somewhat higher than in the previous study (63.39% versus 62.26%), the training set accuracy became significantly lower when using the larger database (64.46% versus 70.77%). This result implies that the network is unable to "memorize" as many specific features of the training set when the size of the database becomes much larger than the number of independent variables (weights and biases) in the network. The improvement in prediction set accuracy is not a result of the training method alone. To demonstrate this, the SCG algorithm was tested on the smaller database. There was no significant difference in accuracy from results obtained using the BP algorithm (results not shown).

Effect of changes to the input layer

Although accuracy on the prediction set was somewhat higher than in the previous study, it was unclear whether the accuracy limit was due to the size of the database or the complexity of the network used to investigate it. To examine this, we increased the complexity of the network systematically by adding units, thereby increasing the number of variables (weights and biases). One method for increasing the number of variables in the network, while at the same time increasing the information available to it, expands the window of residues that the network sees in determining the secondary structure of the central residue in the window. When expanding the window, there is a risk that information on residues distant in sequence will be irrelevant to the network's determination of the structure of the central residue. In such cases, the complexity of network training would be increased, without the addition of useful information. We tested windows of 21 and 39 residues, and also a smaller window of 17. The results are shown in Table 1. We also tested networks that looked only for specific residues at specific positions relative to the central residue, rather than at all residues within a given window size. The residues and positions were chosen using the feature selection algorithm RELIEF (Kira & Rendell, 1992). This approach produced slightly less accurate results than the window method (results not shown).

Although the best results are obtained when using the 19 x 2 x 2 network, there is little variation with any of the other input layer sizes. The 39 x 2 x 2 network, with almost twice as many weights as the 19 x 2 x 2 net, was able to achieve only a 1% greater accuracy on the training set, and did not improve accuracy on the prediction set. This implies that the larger network was unable to formulate many more rules of secondary structure prediction, even specific rules applying only to the training set. Thus, information from residues more than nine amino acids away in the primary sequence appear to be inconsequential in determining the secondary structure of a given residue.

Table 1. ***Secondary structure prediction accuracy for networks with various sized input windows and a hidden layer of 2 units***[a]

Network topology	Training Q_3	Prediction Q_3	C_H	C_E	C_C
17 × 2 × 2	64.2	63.1	44	35	39
19 × 2 × 2	**64.5**	**63.3**	**44**	**36**	**39**
21 × 2 × 2	64.3	62.9	43	35	39
39 × 2 × 2	65.5	63.1	44	35	39

‘Networks were trained for 500 steps of SCG. One round of smoothing was used. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined results of 32-fold cross validation trials are shown.

Effect of changes to the hidden layer

Another way to increase the number of weights in the network is to increase the size of the hidden layer, in effect give the network more "memory" without increasing the amount of information (the input layer). We tested hidden layer sizes from **2** to 12 units, keeping the input window fixed at 19 residues. Results are shown in Table **2.**

The results show a fairly steady increase in the training set accuracy with hidden layer size. Results on the prediction sets increase significantly as the hidden layer size rises from **2** to **5** units, with little change thereafter. A 19 × **8** × 2 network gives optimum performance on the prediction set. Hidden layer sizes above 8 have a derogatory effect on prediction set performance, possibly because the network uses the additional weights to "memorize" irrelevant features of the training set that are inapplicable to the nonhomologous sequences in the prediction set. A 19 × 8 × 2 network uses approximately 3,200 free variables (weights) to fit a training database of 50,000 residues. Each additional node in the hidden layer adds 401 additional weights to the network.

Smoothing

For results obtained using the 19 × 8 × 2 network, applying the smoothing algorithm improves accuracy (Q_3) from 65.4% to 66.4%. Identical results are seen when using the algorithm twice consecutively. However, applying the algorithm more than twice results in a decrease in accuracy.

Further effects of topology

Keeping the hidden layer size fixed at 8, we again attempted to optimize the size of the input window. Window widths from **15** to 21 were tested; results are shown in Table 3. The **15** × 8 × 2 network was found to outperform others with a larger window size, suggesting that the new rules derived by the network with a hidden layer of 8 units depend only on residues within **7** sequential residues of the one for which structure is being predicted. In fact, little accuracy is lost if the network is only presented with information in an 1 I-residue window (within **5** sequential residues of the one predicted).

Table 2. ***Secondary structure prediction accuracy for networks with an input window of 19 residues and various hidden layer sizes***

Network topology	Training Q_3	Prediction Q_3	C_H	C_E	C_C
19 × 2 × 2	64.4	63.4	44	36	39
19 × 3 × 2	66.3	64.7	48	38	41
19 × 4 × 2	68.0	65.8	50	39	41
19 × 5 × 2	68.9	66.2	51	40	41
19 × 6 × 2	69.7	66.4	52	40	42
19 × 7 × 2	70.4	66.1	52	40	41
19 × 8 × 2	71.1	66.4	52	40	42
19 × 9 × 2	71.8	66.1	52	39	41
19 × 10 × 2	72.5	66.2	52	40	41
19 × 11 × 2	73.2	65.8	52	39	40
19 × 12 × 2	74.0	66.1	52	39	40

’Networks were trained for 1,000 steps of SCG. One round of smoothing was used. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined results of 32-fold cross validation trials are shown.

Table 3. ***Secondary structure prediction accuracy for networks with various sized input windows, and a hidden layer of 8 units***

Network topology	Training Q_3	Prediction Q_3	C_H	C_E	C_C
11 × 8 × 2	69.0	66.2	51	39	42
13 × 8 × 2	69.8	66.5	52	39	42
15 × 8 × 2	70.3	66.5	52	40	42
17 × 8 × 2	70.7	66.2	52	39	41
19 × 8 × 2	71.2	66.2	51	39	41
21 × 8 × 2	71.6	65.9	51	39	41
39 × 8 × 2	75.9	63.7	49	37	39

“Networks were trained for 1,000 steps of SCG. One round of smoothing was used. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined results of 32-fold cross validation trials are shown.

Sequence profiles

Using profiles of aligned homologous sequences as network input instead of single sequences has been shown to improve secondary structure results by about **6%** (Rost & Sander, 1993b). We tested our networks on a database containing profiles for our 318 sequences from the HSSP database (Sander & Schneider, 1991). Preliminary results are given in Table 4. They show that the **6%** increase in accuracy is maintained for our database, and that additional units in the input and hidden layers can improve accuracy beyond that seen in the 15 × 8 × 2 network found to be optimal for prediction of single sequences.

Structural class prediction

Both 4-output and single-output class prediction networks (Chandonia & Karplus, 1995) were tested on the database, using 32-fold cross validation. To obtain the information on secondary structure required by the class prediction network, a 19 ×

Table 4. ***Secondary structure prediction accuracy using sequence profiles***

Network topology	Training Q_3	Prediction Q_3	C_H	C_E	C_C
15 × 8 × 2	75.9	72.6	64	51	50
15 × 9 × 2	76.3	72.8	65	52	50
15 × 10 × 2	76.8	72.9	65	52	50
17 × 8 × 2	76.2	72.9	65	52	50
17 × 9 × 2	76.8	72.9	65	53	50

[a] Networks were trained for 1,000 steps of SCG. One round of smoothing was used. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined results of 32-fold cross validation trials are shown.

5 x 2 network was trained for 500 steps. Averaged over all trials, this network produced an overall accuracy of 65.7% on the prediction sets.

For the class prediction network, we first tested the 25 x 4 and **24** x 1 topologies, which were found to be optimal on the smaller database (Chandonia & Karplus, 1995). A graph of training and prediction set accuracy for the 4-output network is shown in Figure 1.

Although there is much more variation in accuracy with continued training than for secondary structure prediction, there is a clear peak in the prediction set accuracy between about 50 and 300 training steps. Continued training leads to "memorization" of the training set and loss of prediction set accuracy. For comparison of class prediction networks in subsequent tests, we used the accuracy recorded after 300 steps. The results are very similar to those obtained on the smaller database; overall accuracy increases by 3.8% (from 73.9% to 77.7%). Accuracy at eliminating potential classes remains high; only one all-beta protein was misclassified as all-alpha, and no all-beta proteins were misclassified as all-alpha. Six proteins from the alpha/beta and "other" classes were misclassified between the two classes. Therefore, one class can be always eliminated by this 4-output class prediction network with 98% accuracy.

Although the addition of a hidden layer was not effective at increasing class prediction accuracy using the smaller database (Chandonia & Karplus, 1995), we tested networks with hidden layers of 2–10 units on the larger database. Results for 4-output and single-output networks are shown in Tables 5 and 6.

A hidden layer of 9 units appears optimal for class prediction using the 4-output network. Single-output networks for predicting each class can perform as well with smaller hidden layers as single-output networks with a hidden layer size of 9. Accuracy at class elimination remained over 98% for hidden layers of all sizes. As seen with our smaller database (Chandonia & Karplus, 1995), the results obtained using 4-output networks are more accurate than those produced by single-output networks. It is likely that results could be further improved by a larger database, because these class prediction networks contain more than 250 independent weights and biases, and there are only 308 proteins in the training database. Neural networks have been shown to become inefficient if there are not at least twice as many training cases as independent variables (Rumelhart et al., 1986).

Secondary structure prediction with class elimination

The class prediction algorithm can limit the predicted class of a protein to three of the four possible protein classes with 98% accuracy. If the eliminated class is all-α or all-β, the corresponding proteins can be removed from the training set for secondary structure prediction, and the resulting "reduced" training set can be used to re-predict the secondary structure of the protein. More details of this training set reduction algorithm are given in Chandonia and Karplus (1995). This algorithm was applied to the larger database, using 15 x 8 x 2 secondary structure prediction networks and 26 x 8 **x** 4 (or 1) class prediction networks.

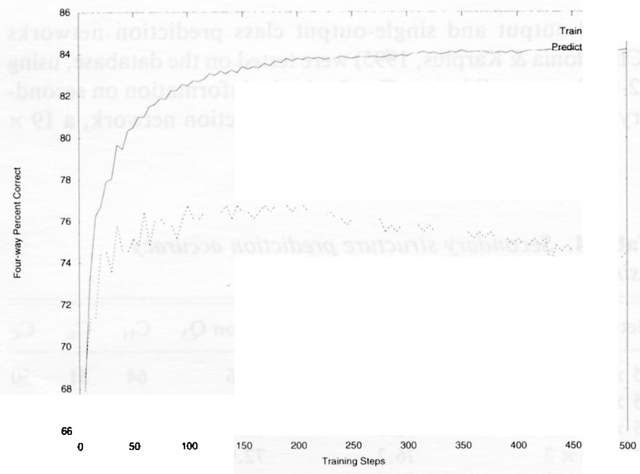

Fig. 1. Class prediction on P318 data set. 26 x 4 class prediction networks were trained for 500 steps using the SCG algorithm. Training was interrupted after every 5 steps to measure accuracy on both the training and prediction sets. Results from 32 cross validation trials are combined. Training set accuracy peaks at 84.2%; accuracy on the prediction set peaks at 76.7%, then decreases to 74.5% after 500 steps.

Table 5. ***Structural class prediction using 4-output networks***

Network topology	Training Q_4	Prediction Q_4	C_A	C_B	$C_{A/B}$	C_O
26 × 4	85.9	77.7	53	34	***66***	71
26 × 2 x 4	86.4	*15.1*	50	42	59	12
26 × 3 x 4	88.8	***17.5***	54	44	63	72
26 × 4 x 4	90.3	78.9	58	44	64	75
26 × 5 x 4	90.6	19.6	60	45	***65***	76
26 × 6 x 4	90.9	19.8	59	48	***66***	75
26 × 7 × 4	91.0	19.8	59	48	***65***	75
26 × 8 x 4	91.0	80.0	59	48	***66***	75
26 × 9 x 4	90.7	80.2	58	**48**	**65**	76
26 × 1 0 x 4	90.6	19.9	58	41	***66***	75

[a] Secondary structure predictions were produced using 1,000 steps of training on a 15 x 8 x 2 network. Networks were trained for 300 steps of SCG. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined and averaged results of seven 32-fold cross validation trials are shown.

Table 6. ***Structural class prediction using single-output networks***

Network topology	C_A	C_B	$C_{A/B}$	C_O
26 × 1	39	19	**57**	68
26 × 2 x 1	50	35	64	**70**
26 × 3 x 1	53	39	64	73
26 × 4 x 1	55	**42**	66	72
26 × 5 x 1	53	39	***66***	12
26 × 6 x 1	53	40	***65***	12
26 × 7 × 1	54	39	64	12
26 × 8 x 1	54	38	***66***	72
26 × 9 x 1	53	38	***65***	71
26 × 10 × 1	54	31	*65*	72

[a] Networks were trained for 300 steps of SCG. Correlation coefficients are shown for the prediction set and are multiplied by 100. Combined results of seven 32-fold cross validation trials are shown.

Results obtained using full training sets are shown in Table *7*, and results obtained using reduced training sets are shown in Table 8. A graph of training and prediction set accuracy versus training time, for both the full and reduced training sets, is shown in Figure *2*.

Results obtained by applying the algorithm are similar to those seen when using the smaller database. Accuracy on both the all-a and all-β proteins increased by about 1% when using the reduced training sets. Accuracy on α/β proteins decreased slightly because several of these were misclassified as all-a or all-β. Accuracy on the "other" proteins increased slightly. The average accuracy on the entire database (weighted by protein length) increased by 0.4%, from *66.6%* to *67.0%.* The average increase is approximately the same as that observed on the smaller database, where accuracy increased from *62.2%* to *62.6%.*

Concluding discussion

Implementation of an improved neural network training algorithm has made possible the application of existing neural network-based procedures to a larger protein database. A conjugate gradient-based algorithm such as SCG allows a database that is an order of magnitude larger to be investigated in roughly the same time as required for a standard back propagation algorithm. A database of 318 protein chains (containing *56,966* amino acid residues) was used, roughly five times the size of the database studied previously. The effects of network topology on prediction accuracy are investigated systematically. Both secondary structure and class prediction results improve by about *4%* with the larger database; i.e., the secondary structure prediction accuracy is 67.0%, and the structural class prediction accuracy is *80.2%.*

The methods described here focus on the prediction of the secondary structure of single sequences. The improvements realized by a larger database and a more efficient training algorithm can be applied successfully to algorithms that operate on a profile of multiple, related sequences (Rost & Sander, 1993a), as such algorithms are based on a similar underlying network. We find that the use of profiles results in a *6%* increase in accuracy with our database, suggesting that the information in a larger database of nonhomologous sequences is independent of the evolutionary information in profiles. Other methods used by Rost and Sander (i.e., second level networks, insertion and deletion information) are being investigated currently, and an additional increase in accuracy is expected.

The secondary structure and class prediction results demonstrate that the most efficient network topologies for solving a given problem can change as the size of the database increases. In particular, as more information is presented, the hidden layer becomes more important. The results demonstrate for the first time that there are improvements in accuracy due to the addition of units to the hidden layer. Additional units in the hidden layer of neural networks allow the formulation of more complex (and accurate) rules for solving a given problem. However, there must be sufficient data in the training set to compensate

Table 7. ***Secondary structure predictions using the full training set summarized by the actual class of the proteins tested (the test set)***

Test set	Prediction accuracy (Q_3)	C_H	C_E	C_C
All-or proteins	70.2	53	15	47
All-β proteins	65.3	20	41	40
α/β proteins	65.7	51	41	42
"Other" proteins	67.9	53	40	42
All **proteins**	***66.6***	**54**	**41**	**43**

[a] Tests were done using 15 x 8 x 2 networks with 32-fold cross validation.

Table 8. ***Secondary structure prediction using reduced training sets, from which all-α and all-β proteins were potentially eliminated***

Test Set	Reduction	Prediction accuracy (Q_3)	C_H	C_E	C_C
All-or proteins	12%	71.1	53	15	48
All-β proteins	12%	66.3	19	42	41
α/β proteins	3%	65.6	51	40	42
"Other" proteins	4%	68.1	53	41	42
All **proteins**	***6%***	***67.0***	**54**	**42**	**43**

[a] Results are summarized by the actual class of the proteins tested (the test set). The reduction measures the decrease in the number of proteins in the reduced training sets, relative to the full training set, averaged over all proteins in each class; e.g., for proteins in the all-α class, an average of 12% of the proteins in the full training set were eliminated to produce the reduced training sets. Tests were done using 15 x 8 x 2 networks with 32-fold cross validation. Correlation coefficients are multiplied by 100.

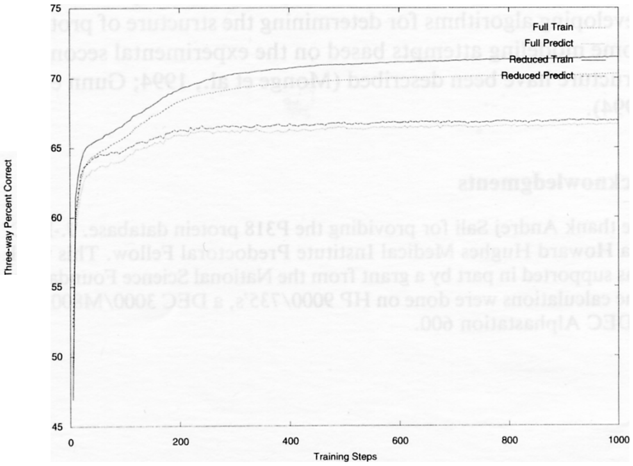

Fig. 2. Reduced versus full training sets. 15 x 8 x 8 networks were trained for 1,000 steps using the SCG algorithm. Training was interrupted after every 5 steps to measure accuracy. Results from 32 cross validation trials are combined. Results using the entire training set are compared to those produced using reduced training sets (networks with a hidden layer size of 8 were used for class prediction). The reduced results consistently remain 0.4–1% higher than the unreduced results over the entire training period.

for the large number of independent variables (weights and biases) that are determined during training. It is likely that, as the number of nonhomologous well-defined protein structures increases, networks with additional hidden units will become more effective at predicting secondary structure and structural class.

Despite the changes in database size and training procedure, several methods developed in the previous paper (Chandonia & Karplus, 1995) are still useful. Both the smoothing filter and the class-based training set selection method provide small but significant improvements in secondary structure prediction accuracy.

A question of practical and theoretical interest is the upper limit of the secondary structure prediction accuracy that can be achieved by this type of data base approach. Certainly, one important factor is the lack of long-range interaction information due to tertiary contacts that can perturb the local secondary structural tendencies. It is likely that additional improvement will be achieved by a further increase in the number of nonhomologous structures in the data base. Rooman and Wodak (1988) suggest that such improvement will be found until at least 1,500 structures are included. The exact positions of helix and sheet ends vary within homologous families (Rost et al., 1994), and sometimes even change upon ligand binding (Jurnak et al., 1990). A recent study (Rost et al., 1994) suggests that the upper limit on secondary structure prediction accuracy wil be 88%, the average accuracy found by homology modeling. For individual sequences of significant length (>100 residues), our current accuracy ranges from around 60 to 90%. It will be difficult to improve on the upper limit without more detailed modeling. However, given a larger training database, neural networks should be able to learn additional prediction patterns that will improve accuracy for proteins on which prediction performance is poor currently, without the need for tertiary structural information.

The rather high accuracy of secondary structure prediction based on sequence alone suggests that the early formation of secondary structure is likely to play a role in protein folding. In particular, it may be important for reducing the search problem in finding the unique native state (Karplus & Weaver, 1994). Also, introducing the predicted secondary structure may be useful in developing algorithms for determining the structure of proteins. Some modeling attempts based on the experimental secondary structure have been described (Monge et al., 1994; Gunn et al., 1994).

Acknowledgments

We thank Andrej Sali for providing the P318 protein database. J.-M.C. is a Howard Hughes Medical Institute Predoctoral Fellow. This work was supported in part by a grant from the National Science Foundation. The calculations were done on HP 9000/735's, a DEC 3000/M800, and a DEC Alphastation **600.**

References

Barton GJ. *1995.* Protein secondary structure prediction. *Curr Opin Struct Biol 5*:372–376.

Chandonia JM, Karplus M. 1995. Neural networks for secondary structure and structural class predictions. *Protein Sci 4*:275–285.

Gunn JR, Monge A, Friesner R, Marshall C. *1994.* Hierarchical algorithm for computer modeling of protein tertiary structure: Folding of myoglobin to *6.2* A resolution. *J Phys Chem 98*:702–711.

Holley H, Karplus M. *1989.* Protein secondary structure prediction with a neural network. *Proc Natl Acad Sci USA 86*:152–156.

Holley H, Karplus M. *1991.* Neural networks for protein structure prediction. *Methods Enzymol 202*:204–224.

Jurnak F, Heffron S, Bergmann E. *1990.* Conformation changes involved in the activation of *ras* p21: Implications for related proteins. *Cell 60: 525–528.*

Kabsch W, Sander C. 1983b. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers 22*:2577–2637.

Karplus M, Weaver DL. *1994.* Protein folding dynamics: The diffusion-collision model and experimental data. *Protein* Sci *3*:650–668.

Kira K, Rendell L. *1992.* The feature selection problem: Traditional methods and a new algorithm. In: *Proceedings of the National Conference on Artificial Intelligence 10.* Cambridge, Massachusetts: MIT Press. pp *129–134.*

Kneller DG, Cohen FE, Langridge R. *1990.* Improvements in protein secondary structure prediction by an enhanced neural network. *J Mol Biol 214*:17 *1–182.*

Levitt M, Chothia C. *1976.* Structural patterns in globular proteins. *Nature (Lond) 261*:552–558.

Matthews BW. *1975.* Comparison of the predicted and observed secondary structure of *T4* phage lysozyme. *Biochim Biophys Acta 405*:442–451.

Møller M. *1993.* A scaled conjugate gradient algorithm for fast supervised learning. *Neural Networks 6*:525–533.

Monge A, Friesner R, Honig B. *1994.* An algorithm to generate low-resolution protein tertiary structures from knowledge of secondary structure. *Proc Natl Acad Sci USA 91*:5027–5029.

Qian N, Sejnowski TJ. *1988.* Predicting the secondary structure of globular proteins using neural network models. *J Mol Biol 262*:865–884.

Rooman M, Wodak SJ. *1988.* Identification of predictive sequence motifs limited by protein structure data base size. *Nature 335*:45–49.

Rost B, Sander C. 1993a. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc Natl Acad Sci USA 90*:7558–7562.

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

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

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

Rumelhart D, Hinton GE, Williams RJ. *1986.* Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, eds. *Parallel distributed processing, vol* I. Cambridge, Massachusetts: MIT Press. pp *318–362*

Sali A, Overington JP. *1994.* Derivation of rules for comparative protein modeling from a database of protein structure alignments. *Protein Sci 3*:1582–1596.

Sumpter BG, Getino C, Noid DW. *1994.* Theory and applications of neural computing in chemical science. *Annu Rev Phys Chem 45*:439–481.

Zhang X, Mesirov JP, Waltz DL. *1992.* Hybrid system for protein secondary structure prediction. *J Mol Biol 225*:1049–1063.