New Methods for Accurate Prediction of Protein Secondary Structure

John-Marc Chandonia[1,2] **and Martin Karplus**[2,3]*
[1]*Department of Cellular and Molecular Pharmacology, University of California at San Francisco, San Francisco, California*
[2]*Department of Chemistry, FAS, Harvard University, Cambridge, Massachusetts*
[3]*Laboratoire de Chimie Biophysique, Institut Le Bel, Université Louis Pasteur, Strasbourg, France*

ABSTRACT **A primary and a secondary neural network are applied to secondary structure and structural class prediction for a database of 681 non-homologous protein chains. A new method of decoding the outputs of the secondary structure prediction network is used to produce an estimate of the probability of finding each type of secondary structure at every position in the sequence. In addition to providing a reliable estimate of the accuracy of the predictions, this method gives a more accurate Q_3 (74.6%) than the cutoff method which is commonly used. Use of these predictions in jury methods improves the Q_3 to 74.8%, the best available at present. On a database of 126 proteins commonly used for comparison of prediction methods, the jury predictions are 76.6% accurate. An estimate of the overall Q_3 for a given sequence is made by averaging the estimated accuracy of the prediction over all residues in the sequence. As an example, the analysis is applied to the target β-cryptogein, which was a difficult target for ab initio predictions in the CASP2 study; it shows that the prediction made with the present method (62% of residues correct) is close to the expected accuracy (66%) for this protein. The larger database and use of a new network training protocol also improve structural class prediction accuracy to 86%, relative to 80% obtained previously. Secondary structure content is predicted with accuracy comparable to that obtained with spectroscopic methods, such as vibrational or electronic circular dichroism and Fourier transform infrared spectroscopy. Proteins 1999;35:293–306.**

Key words: neural networks; secondary structure prediction; structural class prediction

INTRODUCTION

Secondary structure prediction is a useful first step in understanding how the amino acid sequence of a protein determines the native state. The most accurate secondary structure prediction algorithms are based on neural networks[1,2] and nearest-neighbor algorithms.[7,27] Methods trained and tested on groups of aligned, homologous sequences, as compiled in the HSSP database,[3] are more accurate than methods trained and tested on single sequences.[4] The most accurate network methods currently available have three-state prediction accuracy (cross-validated on large databases of non-homologous sequences) of over 72%.[5,6] Accuracy of almost 75% has been reached on a dataset in which very short helices and strands are considered to be coil;[7] this method does not perform as well when accuracy is measured by correlation coefficients[8] rather than percent of correct predictions.

Other aspects of protein structure, such as the structural class,[9] can also be predicted using neural networks and other methods. Secondary structure predictions alone can be used to assign proteins to one of four broad classes (All-α, All-β, α/β, and Other) with 75% accuracy.[5] Neural networks trained on secondary structure predictions, as well as other information such as the amino acid composition and sequence length, can improve this result to 80%.[6] The network can always eliminate one or more classes as possible candidates for a given sequence with 98% accuracy. Use of the latter result to construct specialized training sets for secondary structure prediction networks according to the predicted structural class results in a slight increase in accuracy.[6]

To use neural network algorithms for the prediction of the secondary structure of a new sequence, it is important to know the reliability of the prediction. Thorough cross validation on a large set of non-homologous proteins provides an estimate of the average accuracy expected for new sequences.[10,11] However, results on individual sequences (or small prediction sets) can vary significantly.[10–12] Therefore, it is desirable to be able to estimate the accuracy of a prediction for a given sequence, or particular regions of a sequence. The neural network outputs have been shown previously to correlate with the accuracy of the predictions at every position in the sequence.[13] This information can be used to assign a "reliability index" to each prediction.[5] In this paper, we describe a new method of decoding network outputs which predicts the probability of each type of secondary structure occurring at every position in the sequence, as well as the overall prediction accuracy. We also provide data on how much variation in accuracy can be expected for individual sequences.

Studies have shown that jury techniques combining several neural networks[11] or other prediction methods[10] can be more accurate than methods based on a

*Correspondence to: Martin Karplus, Department of Chemistry, 12 Oxford Street, Cambridge, MA 02138. E-mail: marci@tammy.harvard.edu

Received 18 June 1998; Accepted 17 December 1998

single prediction. However, it is unclear how to best weight the predictions resulting from several methods.[10] With estimated accuracies at every position in the sequence, predictions can be readily combined in a jury decision by averaging the predicted probabilities resulting from two or more network trials. In this paper, we investigate the effect of jury size on the accuracy of the combined prediction, and discuss possible reasons for the increased accuracy seen as a result of jury decisions.

Previous studies have shown that an increase in the size of the database of known structures can result in an increase in the accuracy of both secondary structure and class predictions, provided the networks are supplied with more independent variables (weights) with which to learn additional patterns present in the larger database.[6] The methods described in this paper are applied to a database of non-homologous sequences almost three times larger than the ones previously studied using similar techniques. A discussion is given of the resulting increase in accuracy and the implications for the future of secondary structure prediction as database sizes continue to increase. We compare our results with other recent methods[5,7,27] using a database of 126 proteins often used to benchmark secondary structure prediction methods. Finally, we consider briefly the accuracy of the present results relative to those reported for secondary structure prediction in CASP2.[14]

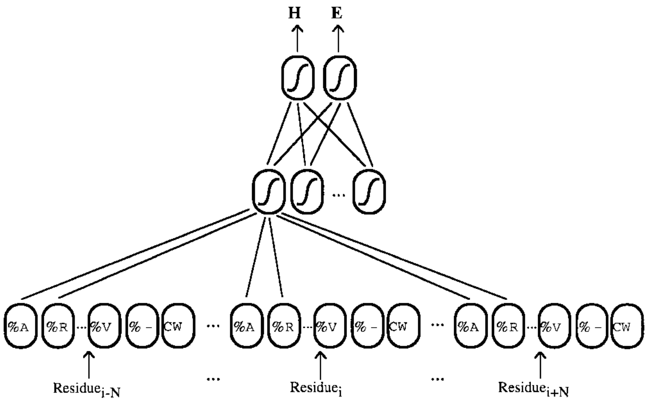

Fig. 1. First level secondary structure prediction network. Units in the network are represented by ellipses, connections between units by solid lines. In the input layer, shown at the bottom of the figure, clusters of 22 units are used to input information about each residue in a continuous stretch of sequence surrounding a given residue, i, for which the secondary structure is being predicted. Twenty units are used to enter the percentage of each amino acid in the multiple sequence alignment at that position. The 21st unit in each cluster, labelled %-, is turned on when the input window overlaps the ends of the sequence alignment. The 22nd unit in each cluster is used to enter the conservation weight.[3] All input units are connected to every unit in the hidden layer, each of which is connected to both output units, H and E. Most units and connections are not shown for clarity.

METHODS

Data Set

The proteins used in this study were a set of 681 chains representative of high-resolution structures available in the Brookhaven Protein Data Bank (PDB) in late 1996. This database was prepared by Andrej Šali using the MODELLER program.[15] 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 determined by X-ray crystallography have a resolution of 3.0 Å or better; 27 chains for which only NMR-determined structures were available were used in addition. Filtering was done to ensure that no pair of protein chains had more than 25% sequence identity. Pairwise alignments were done using global dynamic programming[26] with the identity substitution matrix and a constant gap penalty of 3.

The HSSP database[3] was used to obtain multiple sequence information for each of the structures in the database. An average of 40 aligned sequences for each structure was found, and there was an average of 3.4 different residue types at each position in the known structures. The program DSSP[16] was used to classify the secondary structure of residues in the database. All residues that were neither alpha helix (H) or extended beta strand (E) were considered to be in the "coil" category. The complete database contains a total of 158,428 residues with a composition of 30% helix, 22% strand, and 48% 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.[4,12] 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 randomly divided into 15 groups, each containing several protein chains (14 sets of 46 chains, and 1 set of 37). Networks were trained on sets produced by removing one group of proteins at a time from the database of 681. Each network was then tested on the excluded group of proteins and the results were combined for evaluation of overall prediction accuracy. All results reported in this paper were obtained using this method of cross validation.

For comparison of results with those obtained on a smaller database, the same procedure was performed on the PDB database from 1994. This resulted in a set of 258 chains containing a total of 50,718 residues, with a composition of 30% helix, 22% strand, and 48% coil. For cross-validated testing, these were divided into 12 sets of 20 proteins and one set of 18.

Neural Networks

The primary secondary structure prediction network used in this study is similar to several described previously.[4,12,17] The network uses a "sliding window" approach to iteratively predict the secondary structure of each residue in the protein. At a given time, the network is presented with 15 to 27 (the *window width*) sequential residues of the protein. When training or testing the network, this input window is centered on each of the residues in the protein in turn, and produces a secondary structure prediction for that residue. An overview of the network is shown in Figure 1, and each layer is discussed below.

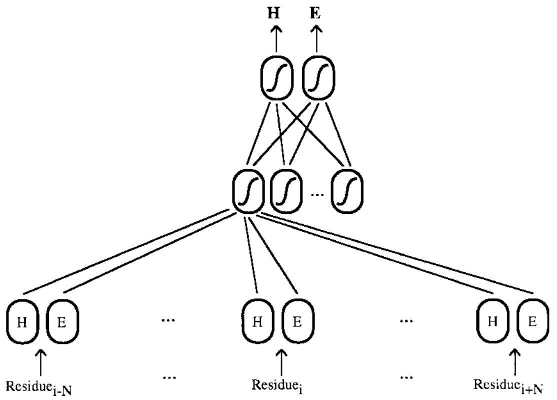

Fig. 2. Second level secondary structure prediction network. Units in the network are represented by ellipses, connections between units by solid lines. In the input layer, shown at the bottom of the figure, clusters of 2 units are used to input information about each residue in a continuous stretch of sequence surrounding a given residue, i, for which the secondary structure is being predicted. These units are used to enter the H and E values output by the first level network at each positon. All input units are connected to every unit in the hidden layer, each of which is connected to both output units, H and E. Most units and connections are not shown for clarity.

For each residue in the input window, the residue type is encoded and presented to the network in 22 separate units of the input layer. Twenty of the units represent a single amino acid residue and are encoded with the frequency with which that residue appears in the profile at that position in the window. The 21st unit is used to indicate that no amino acid appears at the position; this occurs when the window overlaps the ends of all sequences in the profile. The 22nd unit encodes the conservation weight[3] at the position; this is a weighted sum of residue similarities[18] over all sequence pairs in the profile.

The output layer of the networks consisted of two units, *H* and *E*, whose outputs correspond to helix and strand prediction, respectively. In previous studies,[6,12] these were converted to a prediction by comparing the outputs to a cutoff; if neither value was greater than the cutoff, coil is predicted as the secondary structure. Otherwise, the secondary structure corresponding to the higher of the two values was predicted. The cutoff was experimentally determined for each training set, in order to maximize the sum of the Matthews correlation coefficients[8] for the prediction of helix, strand, and coil. In this study, the cutoff method is compared to the estimated accuracy method of decoding the outputs, described below.

Hidden layers of several sizes were tested. For the largest hidden layer tested, containing 40 units, the fully connected network contains 15,082 independent variables (weights and biases). Because this number is small compared to the size of the training set, there were no problems with over-training, or "memorizing" specific characteristics of the training set at the expense of prediction set accuracy.[19]

A second-level network is used to refine the results produced by the primary network. Like the primary network, the second-level network examines a window on the primary sequence and predicts the secondary structure of the central residue in the window, encoding the output in two units H and E. At each position in the window, the H and E outputs from the primary network are used as input, rather than supplying the network with sequence information directly. In a previous study on a database of 318 proteins, an input layer which examined 19 consecutive residues was found to be optimal (Chandonia and Karplus, unpublished results). Hidden layers of several sizes were tested to determine the optimal size for the new database. The H and E outputs can be translated directly into predictions using the cutoff method, or they can be used in the estimated accuracy algorithm described below. An overview of the second level network is shown in Figure 2.

Structural Class Prediction

Quantitative definitions for a system of four structural classes have been proposed by Kneller and colleagues,[20] based on examination of typical proteins from the Levitt and Chothia[9] classes. Details are given in a previous paper.[12] The data set contained 102 chains in the All-α class, 104 All-β chains, 274 α/β chains, and 201 other chains. All-α proteins contained an average of 56% helix, 3% strand, and 41% coil. All-β proteins averaged 4% helix, 40% strand, and 56% coil. Proteins in the α/β class averaged 34% helix, 20% strand, and 46% coil. Other proteins averaged 18% helix, 28% strand, and 54% coil, and were too small or contained insufficient helix and strand to fit any of the other classes.

The class prediction networks used here have been described previously.[12] The input layer consists of 20 units to encode the amino acid content of the protein, 1 unit for sequence length, 4 units to encode predicted secondary structure content, and 1 unit to encode the predicted number of alternations between alpha helix and beta strand. Each sequence in the profile of aligned sequences contributed equally to the amino acid content. As before, two types of networks were studied: networks that contain four outputs, one for each class; and specialized networks with only one output for predicting each class individually.

To compare the importance of different inputs, several modified networks were tested. Three groups of networks were created in which a set of inputs was removed. In one group, the 20 units encoding amino acid content were removed. In the second group, all 5 units containing predicted information on secondary structural content were removed, leaving networks with information only on length and amino acid content. Similar networks have been shown to be highly accurate at distinguishing proteins belonging to 4 specific folds: 4-helix bundles, parallel $(\alpha/\beta)_8$ barrels, nucleotide binding fold, and immunoglobulins.[21] However, when tested on a database of 62 proteins, such networks could distinguish proteins belonging to 4 broad folding classes with only 62% accuracy.[12] In the third group of networks tested, the two units encoding "strong" helix and strand predictions were removed, to determine whether this additional information is redundant for the purpose of class prediction. Finally, a fourth group of networks was tested, in which the units encoding strong helix predictions and the units encoding amino acid content were removed.

Measurements of Accuracy

Three-way percent accuracy (Q_3) and correlation coefficients[8] for prediction of helix (C_H), strand (C_E), and coil (C_C) were used to evaluate accuracy; details are given in a previous paper.[12] For evaluating the accuracy of structural class prediction, correlation coefficients for each class (C_α, C_β, $C_{\alpha/\beta}$, and C_O) were used; the overall four-way percent accuracy (Q_4) was also calculated for the 4-output network.

Determining the Optimal Stopping Point for Training Class Prediction Networks

Although the secondary structure prediction networks are provided with enough training data to prevent over-training, the class prediction networks contain a number of weights (252 for a four-output network with a hidden layer size of 8 units) comparable to the number of training examples (about 640). Because over-training of class prediction networks has been observed in previous trials,[6] an unbiased procedure to estimate the optimal stopping point was introduced.

To estimate the optimal stopping point for a given training set, a second jackknife procedure is used. The original training set is divided into multiple training and prediction subsets, and separate networks are trained on each training subset, then tested on the corresponding prediction subset. The optimal stopping point is chosen as the number of steps at which average accuracy of these prediction subsets is highest. This is used as the stopping point for networks trained on the full training set. As long as sequences in the training and prediction sets are unique, the procedure is unbiased, because no information about the prediction set is used in determining when to stop training. Because the training subsets created in this procedure contain fewer proteins that the original training set, and because proteins in the prediction subsets are not homologous to any in the original prediction set, this procedure can only obtain an estimate of the optimal stopping point. However, it has been shown that the accuracy of class prediction falls off gradually near the optimal stopping point,[6] so such an estimate is useful for preventing significant under-training or over-training of class prediction networks. It should be noted that because the original training set itself results from a jackknife over the entire data set, this procedure is computationally expensive. It requires a total number of trained networks on the order of the square of the number of sequences in the data set.

Reduced Training Set Method

The method for producing optimized training sets based on class prediction has been described in a previous paper.[12] Because the class prediction networks rarely mispredict All-α proteins as All-β, or vice versa, a prediction of either for a tested protein means the other class is eliminated as a possible candidate. Proteins belonging to the eliminated class can then be removed from the training set for secondary structure prediction, and networks trained on the resulting "reduced" training set can be used to re-predict the secondary structure of the protein.

Estimated Accuracy Method

The estimated accuracy method allows network outputs to be translated into predictions, while providing an estimate of the probability of finding the three types of secondary structure at each position in a sequence. A neural network is tested on a large set of proteins of known structure, not including the prediction set. Two dimensional density matrices N_{helix} (H,E), N_{strand} (H,E), and N_{coil} (H,E) are then constructed containing the number of residues in the set with a given secondary structure, as a function of the predicted H and E. A grid size of 0.01 × 0.01 is used to separate the predictions into bins, resulting in 100 × 100 density matrices. For sparse regions of the density matrices, predictions from neighboring bins are combined to provide at least 100 samples in each bin. This is done to ensure that bins contain a statistically significant number of predictions without having to reduce the level of detail of the matrices. From the density matrices, the frequency of each type of secondary structure can be expressed as a function of H and E, by dividing the number of occurrences of each structural type by the total number of residues:

$$p_{helix}(H, E) = \frac{N_{helix}(H, E)}{N_{helix}(H, E) + N_{strand}(H, E) + N_{coli}(H, E)}$$

and correspondingly for the matrices p_{strand} (H,E) and p_{coil} (H,E). Probabilities of finding a given type of secondary structure at every residue in the prediction set are found by looking up the network outputs H and E for that residue in the corresponding frequency matrix. The method implicitly produces a structure prediction, the category corresponding to the highest of the 3 probabilities.

The predictions used to build the density matrices N_{helix} (H,E), N_{strand} (H,E), and N_{coil} (H,E) (the "matrix set") can be made by testing the neural network on its own training set. This method is referred to as the "training set estimated accuracy method." Another method which can be used when performing multiple cross-validated trials is to construct the matrix set from the union of all prediction sets other than the one currently being tested. Although this encompasses the same set of proteins as the training set, the H and E values used for each protein are obtained when that protein is part of a prediction set rather than a training set. This is referred to as the "prediction set estimated accuracy method." Both methods were tested to determine how accurately the estimated frequencies corresponded to observed probabilities of finding each type of secondary structure.

In multiple cross-validated trials, proteins in every prediction set are used in training the neural networks used to make predictions on the other sets. Therefore, in the prediction set estimated accuracy method, proteins from the prediction set were included in all of the training sets of networks used to produce the matrix set. To test whether this biases the method, we did a double jackknife in which two sets of proteins were excluded from the training set at all times. Results in which the prediction set proteins were included in the training sets of networks used to create the matrix set did not vary significantly from results in which they were excluded (data not shown).

Estimated Accuracies for Sequences

The estimated accuracy method can be used to estimate the overall accuracy for a prediction of a given sequence. To do this, the estimated accuracy of the prediction at each residue in the sequence (corresponding to the highest of the three predicted probabilities) is averaged over the length of the chain.

Jury Decisions Using Estimated Accuracies

At each residue, the helix, strand, and coil probabilities are predicted independently by every network in the jury, using the prediction set estimated accuracy method. The jury decision is made by taking an unweighted arithmetic average of the probabilities for helix, strand, and coil over all the networks in the jury. The secondary structure corresponding to the highest of the three resulting probabilities is taken as the combined prediction.

To determine how the jury accuracy scales with the number of separate networks included in the jury, juries composed of all 2^N-1 possible subsets of N networks were tested. The results were sorted by the number of networks included in the jury, and the mean and standard deviation in results (Q_3 and correlation coefficients) were computed for juries of all sizes from 1 to N.

Juries of structural class prediction networks were also tested. Because there are insufficient data to apply the estimated accuracy method and obtain estimated probabilities of a protein belonging to each class, jury decisions were made by averaging the raw network outputs of all networks in the jury. The averaged outputs were converted to predictions in the same manner as for individual networks: by choosing the class corresponding to the highest of four outputs for the 4-output networks, and by comparison with a cutoff for the single-output networks.

RESULTS

Secondary Structure Prediction

Accuracy of the primary network with a larger database

The optimal neural network topology for a representative set of proteins available in the structural database from 1994 was examined in a previous study.[6] It was shown that as the size of the database increases, increasing the size of the input layer had little effect, and could decrease prediction performance. We therefore tested primary networks with the same size residue window (17) as was found to be optimal in the previous study. Since the current procedure for selecting a representative set of sequences uses a more stringent sequence identity cutoff (25% instead of 30%) than that used before, we re-tested primary networks with various hidden layer sizes (HLS) on a representative set of proteins from the 1994 study that were selected according to the new procedure. Results are shown in Table I.

TABLE I. First Level Networks on the 258 Protein Database†

HLS	Training Set Q_3	Prediction Set Q_3	C_H	C_E	C_C
2	69.78%	68.44%	0.552	0.445	0.466
5	74.20%	71.64%	0.634	0.506	0.498
6	74.74%	71.74%	0.640	0.506	0.498
7	75.45%	71.74%	0.641	0.508	0.495
8	76.19%	72.04%	0.649	0.512	0.497
10	77.12%	72.12%	0.652	0.514	0.498
15	79.47%	72.44%	0.661	0.519	0.500
20	80.35%	72.65%	0.663	0.523	0.503
25	80.95%	72.74%	0.665	0.519	0.505
30	82.42%	72.51%	0.663	0.521	0.498
35	82.52%	72.54%	0.666	0.518	0.502
40	82.38%	72.87%	0.666	0.527	0.506

†First level secondary structure prediction networks were tested on the database of 258 proteins. Networks with several hidden layer sizes (HLS) were tested. Networks were trained for 1,000 steps using the Scaled Conjugate Gradient method. Combined results of 13-fold cross validation trials are shown.

On the smaller database, prediction set accuracy usually increases as additional units are added to the hidden layer. Training set accuracy increases more rapidly, indicating that learning of patterns present only in the training set is increasingly likely with larger hidden layer sizes. These effects have also observed on a small database of proteins without multiple sequence information.[6] Because accuracy on the prediction set is observed to be greatest at the largest hidden layer size tested, these results suggest that networks should be given the largest hidden layer which is computationally feasible, at least until the number of weights and biases in the network approaches the number of examples in the training set. Therefore, for tests on the larger database, a hidden layer of 30 units was used; training a single network required approximately 30 CPU hours on a IRIX R10000 or 500 Mhz DEC Alpha machine, or 100 CPU hours on a HP 735/100. Networks were trained for 1,000 steps using the Scaled Conjugate Gradient procedure,[22] pausing every 5 steps to check accuracy on the prediction sets. Prediction set accuracy (Q_3) increased from 72.87% to 73.34% relative to the best results on the smaller set, while Matthews correlation coefficients for helix (0.666 to 0.671), strand (0.527 to 0.533), and coil (0.506 to 0.517) all increased slightly.

Training set accuracy (Q_3) on the larger protein set was 75.82%. The fact that training set Q_3 is only 2.5% higher than prediction set Q_3 indicates that little “memorization” of specific features of the training set took place. The results on the smaller database suggest that, if sufficient computational resources were available, networks with much larger hidden layers could be used with the large database before any over-training effects would be observed. As seen in Table I, prediction set accuracy (Q_3) on the smaller database improves by about 1% as the training set Q_3 increases from 75% to 82%; however, this small improvement requires over a 5-fold increase in the hidden layer size and network training time.

The increase in accuracy as the database size increases might be attributed to two effects: the accidental addition of more easily predicted sequences to the database, or better predictive patterns learned by the networks trained on more sequences. To distinguish between the two effects, sequences present in both databases (with at least 70% sequence identity) were identified. For these 205 sequences, prediction set accuracy measured during cross validation tests using networks with a hidden layer size of 40 on the 258 protein database was 72.70%, with Matthews correlation coefficients of 0.66, 0.52, and 0.50 for helix, strand, and coil predictions. In cross-validation tests using networks with a hidden layer size of 30 on the 681 protein database, prediction set accuracy on these 205 sequences was 74.47%, with Matthews correlation coefficients of 0.69, 0.55, and 0.53 for helix, strand, and coil. Because the improvement in accuracy is greater for the 205 common sequences than for the database as a whole, the increased accuracy can be entirely attributed to the more accurate networks: the additional sequences in the larger database are slightly *more* difficult to predict than the average sequence in the smaller database. Therefore, exposing the networks to more sequences during training enables them to make more accurate predictions on unrelated proteins.

For the 205 common sequences, raw H and E network outputs correlated very well (R = 0.92) between tests conducted on the small and large databases. The increase in accuracy observed when networks were trained on larger training sets was due mainly to improvements in poorly predicted sequences. In the 33% of the sequences with the lowest prediction accuracy, average accuracy improved by 3.9%, from 65.4% to 69.3%. For the middle third, average accuracy improved by only 0.9%, from 72.9% to 73.8%. For the sequences predicted most accurately, average accuracy decreased slightly, from 79.9% to 79.8%.

Second level network

Rost and Sander[5] found that a second level network can improve prediction accuracy by about 1%. We tested networks with a window of 19 residues amd a hidden layer ranging in size from 15 to 20 units. In a previous study on a database of 318 proteins, the window width of 19 residues was found to be optimal (Chandonia and Karplus, unpublished results). Output from the primary network (after 1,000 steps of training) was applied to the second level networks, which were trained for an additional 1,000 steps. Combined results of 15-fold cross validation are shown in Table II.

TABLE II. Second Level Network Results†

HLS	Training Set Q_3	Prediction Set Q_3	C_H	C_E	C_C
16	76.99%	74.23%	0.683	0.548	0.534
17	76.96%	74.25%	0.684	0.548	0.534
18	76.98%	74.31%	0.685	0.549	0.535
19	76.98%	74.28%	0.684	0.548	0.535
20	76.98%	74.29%	0.683	0.550	0.535

†Second level secondary structure prediction networks were tested on the database of 681 proteins. Networks with several hidden layer sizes (HLS) were tested. Results from the first level network (with prediction set accuracy of 73.34% and Matthews correlation coefficients of 0.66, 0.52, and 0.51 for helix, strand, and coil) were used as input for all networks. Networks were trained for 1,000 steps using the Scaled Conjugate Gradient method. Combined results of 15-fold cross validation trials are shown.

TABLE III. Structural Class Prediction Using 4-Output Networks†

HLS	Training Q_4	Prediction Q_4	C_α	C_β	$C_{\alpha/\beta}$	C_O
4	87.4	82.4	0.72	0.71	0.81	0.71
5	88.0	83.1	0.72	0.73	0.82	0.72
6	88.7	82.7	0.72	0.73	0.82	0.71
7	87.5	83.3	0.73	0.73	0.83	0.72
8	88.4	82.9	0.74	0.72	0.81	0.72
9	87.9	82.8	0.73	0.72	0.81	0.72
10	88.2	82.6	0.72	0.72	0.81	0.71

†Secondary structure predictions were produced using second level networks with a HLS of 18 units. Network training times were optimized for each hidden layer size using the optimal stopping point procedure. Correlation coefficients are shown for the prediction set. Combined and averaged results of five separate 15-fold cross validation trials are shown.

Results showed little variation with hidden layer size, although the networks with a hidden layer of 18 units were slightly more accurate. Accuracy on the prediction set increased from 73.34% to 74.31%, while Matthews correlation coefficients for helix (0.66 to 0.67), strand (0.52 to 0.54) and coil (0.51 to 0.52) all increased.

Class Prediction

Networks with hidden layers of several sizes were tested. Results for four-output networks are shown in Table III, and results for single-output networks are shown in Table IV. As a result of the optimal stopping point procedure and the larger database, accuracy improved significantly, from 80% to about 83%. Because the optimal stopping point procedure prevented over-training, the results show little variation with hidden layer size, as was observed in previous studies.[6] Tests were repeated 5 times to measure variation in the results. When network training and testing are repeated with different initial random weights, Q_4 for class prediction varies by ±0.7% (estimated standard deviation), and correlation coefficients vary by ±0.01. Therefore, none of the results vary significantly with hidden layer size. The four-output version of the network is more accurate than the single-output networks, in agreement with earlier work.[6,12] As with smaller data-

TABLE IV. Structural Class Prediction Using Single-Output Networks[†]

HLS	C_α	C_β	$C_{\alpha/\beta}$	C_O
4	0.69	0.67	0.78	0.72
5	0.70	0.69	0.77	0.70
6	0.70	0.69	0.81	0.71
7	0.69	0.68	0.78	0.72
8	0.71	0.68	0.78	0.70
9	0.70	0.69	0.79	0.71
10	0.71	0.69	0.79	0.71

[†] Network training times were optimized for each hidden layer size using the optimal stopping point procedure. Correlation coefficients are shown for the prediction set. Combined and averaged results of five 15-fold cross validation trials are shown.

TABLE V. Importance of Class Network Inputs[†]

Inputs[a]	Training Q_4	Prediction Q_4	C_α	C_β	$C_{\alpha/\beta}$	C_O
All (HLS 9)	87.9	82.8	0.73	0.72	0.81	0.72
AA, L (HLS 9)	76.9	73.6	0.46	0.51	0.71	0.66
AA, 2ary, L (HLS 9)	87.4	82.7	0.72	0.73	0.81	0.73
2ary, L (HLS 9)	85.7	82.5	0.74	0.73	0.80	0.71
2ary, Strn, L (HLS 9)	86.6	83.1	0.75	0.73	0.81	0.72
All (HLS 4)	87.4	82.4	0.72	0.71	0.81	0.71
2ary, L (HLS 4)	86.1	82.5	0.72	0.72	0.81	0.72

[†] Secondary structure predictions were produced using second level networks with a HLS of 18 units. Structural class prediction was done using 4-output networks with some groups of inputs eliminated; those inputs that were present are shown in the first column. Network training times were optimized for each hidden layer size using the optimal stopping point procedure. Correlation coefficients are shown for the prediction set. Combined and averaged results of five separate 15-fold cross validation trials are shown.

[a] Key: AA: 20 units encoding amino acid content, 2ary: 3 units; predicted helix and strand content and alternations, Strn: 2 units; "strong" predictions of helix and strand, L: 1 unit; sequence length.

bases, the networks demonstrated the ability to eliminate classes very accurately in all tests. No protein in the All-α class was misclassified as All-β, and only one protein in the All-β class was predicted to be All-α. This protein, wheat germ lectin (9wgaA), contains only 9% helix and 9% strand, but is placed in the All-β class by the Kneller[20] automatic classification method.

Importance of class network inputs

To understand which information is important for class prediction, we tested networks in which some groups of inputs were eliminated. Results for four-output networks are shown in Table V. Results for single-output networks were similar, although slightly lower (data not shown).

Information on predicted secondary structure content is necessary to predict the class of a protein with optimum accuracy. Although networks given only information on amino acid content and sequence length produced more accurate results on the current database than on previous, smaller databases (Q_4 increased from 62% to 73%, relative to results on 69-chain database used in previous work[12]), information on amino acid content is redundant if a reliable secondary structure prediction has been made. It is interesting to note that networks lacking inputs for amino acid composition were able to correctly classify the 9wgaA protein, due to an accurate (81%) secondary structure prediction; for this protein, an amino acid composition similar to proteins in the All-α class leads to an error if the composition inputs are present. Information on "strong" helix and strand predictions also appears redundant; results are not significantly different if the information is excluded. All inputs are important if the class prediction networks are trained on smaller databases,[12] possibly because the secondary structure predictions are significantly less reliable.

Networks without an input for sequence length were not tested, because this information is always available and known to be important in the definitions of the structural classes. Networks were observed to misclassify several sequences that were smaller than the minimum length for proteins of the predicted class. This frequently occurred because no shorter protein belonging to the class was present in the training database, preventing the networks from learning the exact length cutoff. If corrections are made based only on the known sequence length, the prediction accuracy increases from 82.7% (averaged among all four-output networks except for the ones lacking information on predicted secondary structure) to 84.8%. All Matthews correlation coefficients also increase: the coefficient for prediction of All-α proteins increases from 0.727 to 0.767; All-β from 0.726 to 0.740; α/β from 0.811 to 0.828; and Other from 0.719 to 0.768.

The information on predicted secondary structure and length can also be used to identify the structural class directly. Secondary structure predictions made on the large database using second level networks with a hidden layer of 18 units and 15-fold cross validation were used to estimate the helix and strand content of each protein in the database. This was done by dividing the number of residues predicted to be helix and strand by the total number of residues in the protein. The average estimate of the helix content was off by 5.8% ± 5.7%. The average estimate of strand content was off by 6.3% ± 6.3%. Pearson correlation coefficients for predicted vs. actual content are 0.92 for helix, and 0.81 for strand. These estimates compare favorably to predictions made using the method of Rost and Sander,[5] which can estimate helical content to within 8.5% error (0.87 correlation coefficient) and strand content to within 7.5% error (0.74 correlation coefficient). Estimates of structural content are comparable to a statistical method which combines data from vibrational circular dichroism, electronic circular dichroism and Fourier transform infrared spectroscopic techniques.[23] For 19 proteins not used in parameterizing the algorithm, this method predicts both helix and strand content with approximately 5% average error. The predicted helix and strand percentages can be used in the Kneller[20] class definitions to produce class predictions without using a specialized neural network. On average, these predictions are more accurate than predictions made by the class prediction

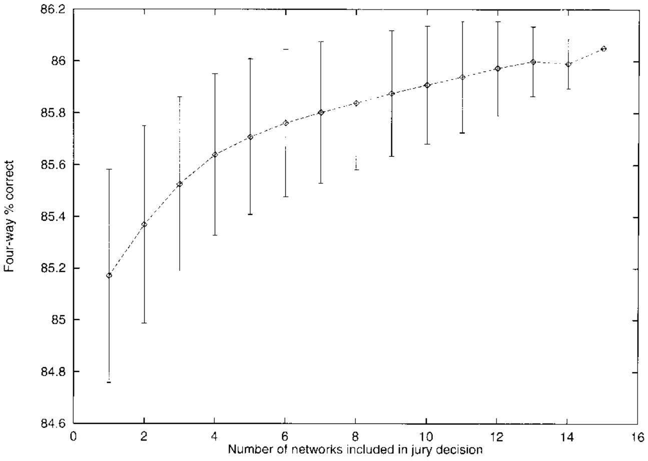

Fig. 3. Jury decisions with class prediction networks. Fifteen class prediction networks containing all input units (five each with hidden layer sizes of 5, 7, and 9) were used to predict the class of each protein independently, using 15-fold cross validation. Jury decisions were made by averaging the raw outputs from several of the networks. All 2^{15}-1 possible combinations of networks were tested. Results were sorted into sets according the number of networks averaged in the jury decision. The mean and standard deviation in Q_4 for each set are shown.

network, with a Q_4 of 85.9% and Matthews correlation coefficients of 0.759, 0.786, 0.837, and 0.788 for All-α, All-β, α/β, and Other proteins.

Juries of class prediction networks

Juries of several class prediction networks were tested, in which multiple networks were used to predict the class of each protein. For each protein, the raw outputs of several networks were averaged before translating to a prediction. Fifteen class prediction networks (five each with hidden layer sizes of 5, 7, and 9) were used. Juries composed of each of 2^{15}-1 possible subsets of these 15 networks were tested, and results were sorted by the number of networks involved. Prediction accuracy for each jury size is shown in Figure 3.

The average accuracy of the jury increases as the number of networks in the jury increases. Although there is some variation caused by particular combinations of networks, the juries with more members were more accurate on average than juries with fewer networks. Random variations in accuracy, resulting from testing networks trained using different initial weights, are greatest for single networks. As the number of networks in the jury increases, the random variation in prediction accuracy caused by particular combinations of individual networks decreases, because juries containing a large number of networks sample both the most and least accurate of the individual networks. Although only one of the 15 networks was more accurate than the non-network method described above, a jury containing all 15 networks is slightly more accurate, with a Q_4 of 86.0% and Matthews correlation coefficients of 0.781, 0.748, 0.851, and 0.787 for All-α, All-β, α/β, and Other proteins.

TABLE VI. Reduced Training Sets†

Test Set	Full Set Q_3	Reduced Set Q_3	C_H	C_E	C_C
All-α proteins	76.49%	76.45%	0.607	0.251	0.571
All-β proteins	71.77%	72.16%	0.334	0.503	0.480
α/β proteins	74.93%	74.92%	0.684	0.566	0.544
"Other" proteins	72.73%	72.75%	0.649	0.505	0.499
All Proteins	74.31%	74.37%	0.685	0.551	0.536

†Secondary structure prediction using reduced training sets, from which All-α and All-β proteins were potentially eliminated. Results are summarized by the actual class of the proteins tested (the test set). Tests were done using first level networks with a HLS of 30 units, and second level networks with a HLS of 18 units, each of which was trained for 1,000 steps. Combined results of 15-fold cross validation are shown.

Reduced training sets

In previous studies, the reduced training set algorithm improved results on proteins in the All-α and All-β classes.[6] We applied the training set reduction method to proteins which were predicted as belonging to the All-α or All-β classes. First level networks with a hidden layer size of 30 units were trained for 1,000 steps on the reduced training sets. Predictions made at the end of training were presented to second level networks with a hidden layer size of 18 units, which were trained for an additional 1,000 steps. Results are shown in Table VI.

For all proteins, including the ones for which the algorithm could not produce a specialized training set, average accuracy increased by a small but significant amount, from 74.31% to 74.37%. This increase is almost entirely due to improved prediction of beta strand in proteins in the All-β class. While accuracy on proteins of other classes increased or decreased only slightly, accuracy on proteins in the All-β class increased by 0.4% due to more accurate predictions of both beta strand and coil. The reduced training set method was significantly less effective on this data set than on a smaller set containing only single sequence information;[6] for the latter data set, accuracy for proteins in both the All-α and All-β classes improved by 1%. The decreased effectiveness is probably due to the smaller proportions of All-α and All-β proteins in the current database than in previous databases; elimination of either of these groups from the training sets does not change the composition of the training set as much as in previous studies.

Estimated Accuracies

The training set estimated accuracy method produces an estimate of the probabilities of finding each type of secondary structure at all residues in the sequence, while producing a new prediction corresponding to the highest of the three probabilities. The new predictions improve the Q_3 relative to the results after training set reduction (from 74.37% to 74.58%) while increasing coil prediction slightly at the expense of helix and strand (C_H decreases from 0.685 to 0.683, C_E decreases from 0.551 to 0.549, C_C increases from 0.536 to 0.546).

To test the validity of the predictions, all residues were clustered by predicted helix probability into bins of 5% width. The real helix frequency was calculated for each bin, and compared to the median expected probability of the bin. A similar procedure was done for strand and coil probabilities. The method was found to systematically overestimate the real frequency when the estimated probability was above 40–45%. For coil, the overestimate is approximately 2%; for helix and strand, the overestimate is approximately 5%. At estimated probabilities below 40%, the method underestimates the frequencies to a similar degree. This is a result of using statistics derived from testing neural networks on their own training set; as shown in Table II, the training set results are slightly (2.7%) more accurate due to some memorization of the sequences.

Correcting for the systematic errors in the training set estimated accuracy method would result in more accurate frequency estimates. However, the type of secondary structure predicted (and thus, accuracy measures such as Q_3 or correlation coefficients) would only change if two of the predicted frequencies were similar, and corrected by different amounts. In the region in which the two highest frequencies might be similar (33%–50%) the systematic errors are small. Furthermore, such a corrective procedure might be biased by using information on the prediction set. Therefore, rather than correcting systematic errors in the training set estimated accuracy method, the prediction set estimated accuracy method was developed to replace it.

The prediction set estimated accuracy method, while slightly more difficult to implement, produces more accurate probability estimates. Like the training set estimated accuracy method, it increases the Q_3 from 74.37% to 74.58%. This method also increases coil prediction slightly (C_C increases from 0.536 to 0.547) at the expense of helix (C_H decreases from 0.685 to 0.680) and strand (C_E decreases from 0.551 to 0.543). The estimated accuracy method produces estimated probabilities which correlate extremely well ($R = 0.99$) with the actual probabilities in each bin; no systematic overestimates or underestimates of accuracy were observed for any of the bins.

Estimated accuracies were also computed for every sequence. These are compared with the actual accuracies in Figure 4. While the predicted accuracies have fair correlation ($R = 0.52$) with the actual accuracy of the prediction, significant variance and some overestimates of accuracy also occur. The degree of variation at the level of individual proteins is much higher than in the previous (bin) test, because each bin contained tens of thousands of predictions, several orders of magnitude more than the length of a typical protein.

Juries of secondary structure prediction networks

Juries of multiple secondary structure prediction networks were also tested. For each residue, the estimated probabilities of helix, strand, and coil were computed using several methods; the probabilities were averaged and translated to a prediction. One prediction was made using the reduced training set method described above. Seven additional predictions were made using the second level secondary structure prediction networks shown in Table II; two second level networks with hidden layer size of 19 units, and single second level networks with hidden layer sizes of 15, 16, 17, 18, and 20 units were used. All predictions were made using the prediction set accuracy method and validated with 15-fold cross validation. Juries were tested containing each of the 2^8-1 possible subsets of these 8 predictions; results were sorted by the number of predictions in each subset. Results are shown in Figure 5.

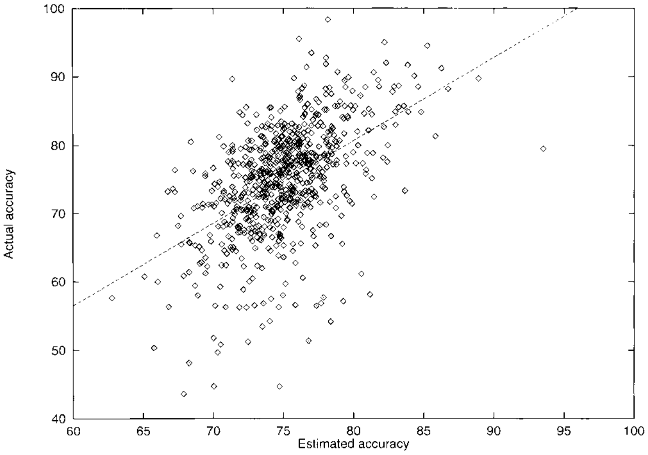

Fig. 4. Prediction set estimated accuracy method, by protein. Secondary structure was predicted using the reduced training set procedure. The prediction set estimated accuracy method was used to calculate the probabilies of finding helix, strand, and coil at each postion (results obtained using the cutoff method to make predictions are shown in Table VI). Estimated accuracy for each protein was calculated by averaging the highest of the three secondary structure probabilities at each position in the sequence. The estimated accuracy for each sequence is compared to the actual accuracy of the prediction.

As was the case with class prediction networks, juries of secondary structure networks provide significantly better accuracy than single networks. Accuracy varies significantly with the particular choice of networks in the jury, but increasing the number of networks increases the average accuracy. The jury using all 8 predictions scored 74.76% accuracy, with Matthews correlation coefficients of 0.684, 0.544, and 0.550 for helix, strand, and coil. The most accurate jury of the 255 tested achieved 74.81% accuracy using four of the networks. However, there appears to be no a priori method of picking those particular four predictions; when tested independently, two of the networks had lower than average accuracy. It is interesting to note that results obtained using the reduced training set method were included in every jury (10 of the 255) scoring above 74.8%. This suggests that more diverse results in the jury may lead to higher accuracy. Rost and Sander[5] observed over a 1% increase in accuracy in a single jury test using 12 networks with more varied accuracy. However, as seen in Table II, results on our database using two levels of networks do not vary as significantly with hidden layer size as in previous studies on smaller databases.[6]

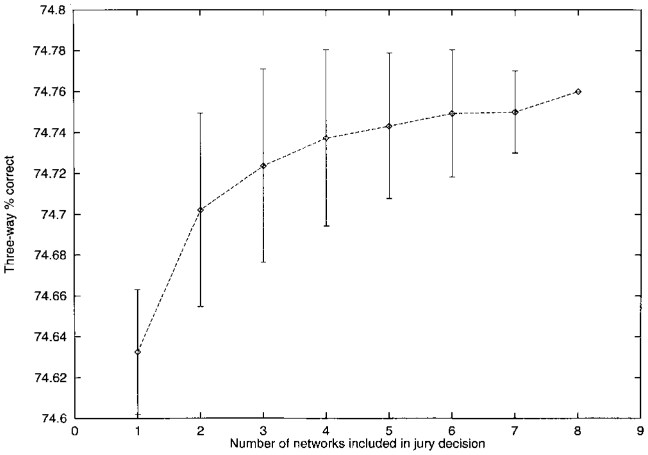

Fig. 5. Jury decisions with secondary structure prediction networks. Eight combinations of networks were used to predict the secondary structure of each protein independently, using 15-fold cross validation. One prediction was made using the reduced training set method (results are shown in Table VI); the other predictions were made using the second level networks shown in Table II. For all 8 predictions, the prediction set estimated accuracy method was used to produce estimated probablilities of finding helix, strand, and coil at each residue. Jury decisions were made by averaging the predicted probabilities from several of the networks. All 2^8-1 possible combinations of networks were tested. Results were sorted into sets according the number of networks averaged in the jury decision. The mean and standard deviation in Q_3 for each set are shown.

Accuracy of second-best predictions

In addition to making a secondary structure prediction at every position in a sequence, the predicted probabilities of finding helix, strand, or coil can be used to rank the secondary structures in order of expected likelihood of occurrence. The accuracy of the second-highest predictions made by the jury of 8 networks was measured and compared to the expected accuracy that could be obtained by chance. In cases where the highest prediction was incorrect (40,161 of 158,428 residues, or approximately 25% of the database), accuracy of the second-best prediction was 79.24%, with Matthews correlation coefficients of 0.59, 0.63, and 0.82 for helix, strand, and coil. Although this accuracy is greater than that for the highest prediction, this is expected because the number of possible secondary structure candidates has been reduced from three to two. Randomly choosing one of the two remaining candidates in accordance with the observed database frequencies (30% helix, 22% strand, and 48% coil) produces an accuracy of only 52.98%, with correlation coefficients of 0.18, 0.16, and 0.57 for helix, strand, and coil. The second-best predictions produced by the jury of neural networks contain significantly more accurate information; such information may be useful in situations where experimental evidence indicates that the best network prediction is incorrect, or for other cases in which a user of the method wishes to make manual corrections to the prediction.

If the highest and second-highest predictions are combined, by scoring a prediction as correct in cases where either of the top two predicted frequencies corresponds to the actual secondary structure, accuracy reaches 94.74%, with correlation coefficients of 0.90, 0.87, and 0.96 for helix, sheet, and coil. Secondary structure types corresponding to the lowest of the three predicted frequencies only occur at 5.26% of the residues in the database, and therefore can be eliminated with high accuracy.

TABLE VII. Prediction Accuracy Varies With Helix/Strand Length†

Helix/strand length	Helix % correct	Strand % correct
2	N/A	31.6
3	N/A	50.5
4	32.7	62.9
5	40.8	71.5
6	52.3	71.5
7	67.3	67.3
8	71.2	65.5
9	75.8	62.3
10	78.2	54.5
11+	83.9	51.3

†Secondary structure predictions for residues in helices and strands are sorted by the correct length of the helix or strand being predicted. The percent of correctly predicted residues are shown for each length. The minimum length of helices (as defined by DSSP) is 4 residues; the minimum length for a strand of beta strand is 2.

Prediction accuracy varies with helix/strand length

Secondary structure predictions made by the jury of 8 networks were sorted by the actual length of the helix or strand element being predicted; the percent of correctly predicted residues in these elements are shown in Table VII.

Helix prediction accuracy increases with the length of the helix; strand prediction accuracy peaks with beta strands of length 5 or 6. Similar results have been observed previously[28] when prediction accuracy is measured on a per-segment, rather than per-residue basis. The drop in strand prediction accuracy for longer strands may be due to the hydrogen bond partner for each residue being outside the window of residues presented to the network. For strands of length 5 or 6, residues in other strands of the same beta sheet are sometimes present in the window. Although increasing the size of the input window might produce more accurate predictions for longer strands, previous studies have shown that the Matthews correlation coefficient for overall strand prediction does not increase with a larger window.[6,12]

The shortest elements of secondary structure defined by DSSP (helices of length 4, and strands of length 2) are the most difficult to predict. If they are redefined as coil,[7] the apparent prediction accuracy increases. When the jury of 8 networks described above was tested on the database using a minimum helix length of 5 and a minimum strand length of 3, prediction accuracy increased from 74.76% (using the DSSP length cutoffs) to 75.67%. All networks used were trained on a database using the standard DSSP definitions of secondary structure; networks trained on databases using the longer length cutoffs might produce more accurate results, in accord with the results of Frishman and Argos.[7]

Predictions made by the jury of 8 networks also reproduce the length distributions of helices and strands observed in the correct structures. For lengths of 6 residues or greater, the number of predicted helices of a given length varies from the correct number by an average of only 12%. Both the distributions of correct and predicted helix lengths have maxima at a helix length of 10 residues. The distribution of strand lengths is also very similar to the correct distribution for strands of length 3 to 7 residues; for longer strands (which are relatively rare) the prediction underestimates the number of strands by a factor of two to five. For short helices (4 or 5 residues) and strands (2 residues), both types of secondary structure are under-represented in the jury prediction. Because the predictions made by the jury are not filtered to remove short segments of helix and strand, elements of secondary structure as short as a single residue are sometimes predicted; DSSP classifies these as coil. Although previous studies[6,12] filtered these short elements of secondary structure out of the prediction by treating them as coil, it is unclear whether such a filter should be applied to the probabilities produced by the estimated accuracy method.

Analysis of accuracy

The secondary structure predictions made by the jury of 8 networks on each of the 681 sequences were analyzed to determine factors that affect the accuracy of predictions on individual sequences. Predictions were obtained using the prediction set accuracy method, and validated by 15-fold cross validation. Several subsets of the database were analyzed, and the average accuracy (without weighting by sequence length) and variance in accuracy for each subset were calculated. For all 681 sequences, the average accuracy is 74.89%, with a standard deviation of 8.17%. Results for several subsets of the database are shown in Table VIII.

Secondary structure predictions are slightly more accurate for medium length sequences. For the shortest 1/3 of the database, containing sequences of 35 to 130 residues with an average length of 89 residues, average accuracy was 74.46%. For the medium length sequences (131 to 280 residues, with an average length of 196), average accuracy was 75.49%. For longer sequences (283 to 905 residues, with an average length of 413), accuracy was only 74.71%. The higher variation in accuracy on shorter sequences is a consequence of the fact that each residue predicted makes more of a difference in the overall accuracy of the sequences.

As shown in both Tables VI and VIII, predictions are most accurate for proteins in the All-α class, but are also good on mixed α/β proteins. The higher accuracy on All-α proteins and α/β proteins is a result of the higher alpha helix content of these proteins; as shown by the Matthews correlation coefficients, helix prediction is more accurate than prediction of strand or coil. The standard deviations in accuracy on the four classes reflect the average sequence length of the proteins; α/β proteins are longer on average, and proteins in the Other class are shorter.

TABLE VIII. Analysis of Accuracy[†]

Subset	Number of Proteins	Mean Q_3 (%)	Standard deviation of Q_3 (%)
All proteins	681	74.89	8.17
Shortest third of sequences	227	74.46	10.48
Middle third of sequences	226	75.49	7.55
Longest third of sequences	228	74.71	5.76
All-α proteins	102	77.78	6.98
All-β proteins	104	72.60	7.00
α/β proteins	274	75.84	5.77
"Other" proteins	201	73.30	10.99
Many charged residues	100	76.46	9.68
Few charged residues	100	73.19	8.68
Many hydrophobic residues	100	74.88	8.93
Few hydrophobic residues	100	73.77	9.45
1 sequence in HSSP profile	32	69.65	7.46
2 sequences	25	70.94	6.61
3 sequences	37	71.39	9.15
4 sequences	28	74.63	6.11
5 sequences	21	74.86	7.46
4 or more sequences	587	75.56	8.01
Integral membrane proteins	11	67.01	7.77
Viral coat proteins	20	68.54	8.67
Proteins containing heme	41	77.88	6.93
Structures determined by NMR	27	72.90	13.01

[†] Secondary structure predictions made by the jury of 8 networks on each of the 681 sequences are divided into several subsets according to characteristics of each protein. The average and standard deviation of the prediction accuracy (Q_3) are given for each subset. Results are not weighted by sequence length.

The accuracy of predictions varies somewhat for proteins with extreme proportions of charged (H, Q, E, K, and R) and nonpolar (A, V, L, I, P, F, W, M, and C) residues. Accuracy on the 100 proteins with the highest fraction of charged residues (from 29% to 49%) is over 3% higher than accuracy on the 100 proteins with the lowest fraction of charged residues (9% to 19%). Proteins with many hydrophobic residues (47% to 61%) are similar to the average for all proteins, while proteins with few hydrophobic residues (26% to 38%) score slightly below average. The differences can be explained in terms of the class of the proteins. Almost one third of the proteins (32 of 100) with many charged residues belong to the All-α class, which is predicted with the highest accuracy of the four classes, while only a few of the sequences (3 of 100) belong to the All-β class, which is predicted with the lowest accuracy. Proteins with few charged residues are more likely (38 of 100 sequences) to be members of the All-β class. Half of the proteins with few hydrophobic residues are members of the Other class, which is also predicted with below average accuracy.

The results also confirm the importance of multiple sequence data. For the 32 proteins for which only one sequence was available in the HSSP profile, average accuracy is 5% lower than the average for the entire database. Prediction accuracy increases steadily as the number of sequences in the profile increases. For the 587 proteins for which at least four sequences are aligned in the HSSP profile, average accuracy is almost 1% higher than the average for the entire database. These results are similar to those of Rost and Sander,[11] who found that networks trained on multiple sequence profiles performed with 7% lower accuracy when tested on single sequences.

The method also performs poorly for integral membrane proteins, including porins and parts of the photosynthetic reaction center. As the networks are trained on data sets containing mostly soluble proteins, they are presumably unable to learn predictive patterns which can be applied to integral membrane proteins, due to the differing native environments of the proteins. Neural network methods specialized at predicting transmembrane helices and topology of membrane proteins have been more successful.[24,25] Another set of poorly-predicted proteins are viral coat proteins. The relatively low accuracy (68.54%) is possibly due to a greater influence of tertiary contacts on the secondary structure of these proteins, although most of the proteins also belong to the All-β or Other classes, for which the average accuracy is slightly below 73%. Both the viral coat proteins and integral membrane proteins contain many sequences in their HSSP profiles, indicating that the lower accuracy is not caused by a lack of multiple sequence information.

The presence of prosthetic groups does not significantly affect accuracy, possibly because these contacts have no more effect on secondary structure than other tertiary contacts. Proteins containing heme groups average 77.88% accuracy, similar to the accuracy of the method for other proteins containing large amounts of alpha helix. Structures determined by NMR are also predicted with similar accuracy to proteins of similar size and secondary structure content for which the structure was determined by X-ray crystallography.

For 552 proteins in the database which are not integral membrane proteins or viral coat proteins, and contain at least four aligned sequences in the profile, secondary structure prediction accuracies form a normal distribution with a mean of 75.93% and a standard deviation of 7.91%.

Prediction of Rost and Sander test set

A database of 126 non-homologous protein chains with multiple sequence information assembled by Rost and Sander[5] has frequently been used for comparison of secondary structure prediction methods.[7,27] To eliminate bias due to homology with proteins in our training sets, we first calculated the sequence identity of the 126 proteins with all proteins in our database. Eleven proteins were found to have no significant sequence homology (as defined by Sander and Schneider;[3] i.e., 25% sequence identity over lengths of 80 residues or more) with any proteins in our database. The secondary structure of these proteins was predicted using a jury of all networks described above. One hundred and twelve (112) proteins were homologous to a single sequence in one of our 15 prediction sets used for cross validation. These proteins were each tested with the jury of 8 networks used for cross validation tests on the corresponding prediction set. Three additional proteins (1fxi_A, 1r09_2, and 2ltn_B) were homologous to sequences in two of our 15 sets. Because our cross validation procedure involved removing one set at a time from the database, no networks were trained on sets from which both homologous proteins were simultaneously removed. Training additional juries of networks to test these three proteins would be computationally expensive. Therefore, each of the three proteins was tested along with the prediction set including the sequence with the higher degree of homology. However, because one other protein with significant homology to the one being tested (28% identity over >80 residues for both 1fxi_A and 1r09_2, 40% identity over 47 residues for 2ltn_B) could not be eliminated from each of the training sets, statistics on these three proteins were compiled separately.

For the remaining 123 protein chains, the overall three-state accuracy (Q_3) of the jury prediction was 76.6%, with Matthews correlation coefficients of 0.71, 0.57, and 0.57 for helix, strand and coil. If the additional three proteins are included, Q_3 drops to 76.5%; correlation coefficients are unchanged. For comparison, Rost and Sander[5] obtained 71.6% accuracy on the same set, with correlation coefficients of 0.61 and 0.52 for helix and strand. Salamov and Solovyev[27] obtained 73.5% accuracy, with correlation coefficients of 0.65 and 0.53 for helix and strand. Frishman and Argos[7] reported 74.6% accuracy, with correlation coefficients of 0.61, 0.45, and 0.44 for helix, strand, and coil. The increase in Q_3, without an improvement in correlation coefficients, was obtained by treating short helices and strands as coil.[7] In addition, the two latter groups[7,27] excluded two chains of hemagglutinin (3hmg) from the data set. Accuracy of the jury prediction on these two chains is 65.6%, a performance typical for membrane proteins (Table VIII). If these chains are excluded from our statistics, Q_3 increases from 76.5% to 76.7%.

Prediction of CASP2 target β-cryptogein

The target β-cryptogein (1beo) was identified as a particularly difficult target in the CASP2 study,[14] with only 53% of the secondary structure correctly classified by the PHD server.[5,11] We predicted the secondary structure of 1beo from the sequence profile in the HSSP database[3] using a jury of all networks described above. This prediction was correct at 61 of 98 (62%) residues. The improvement in accuracy over the PHD prediction largely resulted from a completely correct prediction of the first helix (residues 5–19), which was predicted as a helix followed by a strand by PHD. As in the PHD prediction, the second (residues 22–31) and last (residues 84–96) helices were completely missed. All residues in the second helix were predicted to be coil, with a maximum helical probability of 24%. In the last helix, the first three residues were predicted to be an extension of the strand at residues 81–82, while all other residues were predicted to be coil; the maximum helical probability in this region was 26%. The jury also predicted a short strand at residues 35–38 which is not observed in the structure; PHD predicted a strand at a similar location, from residues 33–39.

As described above, the level of error observed for β-cryptogein is more common for short sequences. Of the proteins in the shortest third of our database, 31 of 227 (14%) were predicted with Q_3 accuracies lower than that for β-cryptogein. In the case of β-cryptogein, the estimated accuracy of the prediction (66%) may be helpful in indicating to the user of the algorithm that this prediction is somewhat less accurate than average.

DISCUSSION

The increase in accuracy observed as the database size is increased from 258 to 681 proteins, without changes in the algorithm, indicates that neural networks are able to learn more useful predictive patterns from larger databases even without additional hidden units. Accuracy on sequences common to both databases improved by 1.8% due to this effect alone. Results on training set accuracy suggest that accuracy can be improved further with the addition of more hidden units, before significant memorization of the training set occurs. Expected improvements in computer speed should make larger networks practical in the near future. Additional increases in accuracy were made by applying new algorithms such as second level networks, juries, and the estimated accuracy method. Together, these improved accuracy by an additional 1.4%.

The estimated accuracy method not only produces slightly more accurate results, but also allows reliably predicted regions of the sequence to be identified. Although reliably predicted regions of proteins have been identified previously,[5,13] expression of secondary structure predictions as probabilities facilitates combination of this prediction algorithm with other methods. Reliably predicted regions of secondary structure should be useful for tertiary structure prediction algorithms which rely on an accurate knowledge of secondary structure.[5,13] Expression of the secondary structure predictions as probabilities also allows an accurate second prediction to be made in cases where experimental evidence suggests that the first prediction is incorrect.

Increased accuracy resulting from jury decisions has been observed before[11] but never systematically investigated. Averaging several predictions of similar accuracy is almost certain to result in a prediction higher than any of the individual predictions, and the average accuracy increases as the number of predictions added to the jury increases. If all networks made similar predictions, the average accuracy would remain the same if several were averaged in a jury decision, although variation in the results would decrease. However, if the networks vary in which predictions are incorrect, the majority of the networks will override incorrect predictions made by an individual network. The increase in the average accuracy as networks are added to a jury indicates that networks with similar accuracies tend to vary somewhat in their correct and incorrect predictions. This is supported by an examination of the weights of similarly sized networks after training. When networks are trained starting with different initial random weights, the resulting weights between the input layer and any of the hidden units rarely correlate well between networks, even for networks of the same size. Patterns learned by the hidden units appear to be arbitrary, and chosen randomly from a large pool of predictive patterns. Although networks achieve similar accuracy after training, the particular patterns learned by each one are different, allowing a jury to attain higher accuracy than any of the individual networks. This suggests that different prediction methods which achieve similar accuracy, yet make correct and incorrect predictions in different places, would be extremely useful in a jury-based method.

The rather low increase in accuracy resulting from the reduced training set algorithm relative to previous applications[6,12] is probably a result of the relative scarcity of All-α and All-β sequences in the larger database. Due to increased accuracy in prediction of the helix and strand content of proteins, and the increase in the size of the database, it may be possible to create reduced training sets which are similar in helix and strand content to a predicted protein, yet are still large enough to include proteins with some range of helix and strand content in case the predicted content is wrong. It may also be useful to treat each domain of large α/β proteins individually, in cases where single domains could be identified and classified as All-α or All-β.

Increases in the accuracy of secondary structure prediction have made the current class prediction network nearly obsolete. Class superfamilies can be directly identified from the predicted secondary structure content and the sequence length without training specialized networks. Specialized networks may still be useful for distinguishing between particular folds with similar secondary structure content.

Anticipated improvements in the speed of computers and the size of sequence and structural databases should lead to increases in the accuracy of neural network predictions in the near future. Larger hidden layer sizes, which will be feasible with a several fold increase in computer speed, are expected to produce approximately a 1% increase in the accuracy of the algorithm using the current database. As the number of sequenced genomes continues to increase, we expect that multiple homologous sequences will be discovered for most proteins in the database; proteins with at least four homologous sequences are predicted with approximately 1% greater accuracy than the current average. Neural network-based algorithms also achieve higher accuracy as the number of structures increases, as demonstrated by the improvement in accuracy as the database size was increased from 258 to 681 proteins. This trend should continue as the number of unique structures continues to increase, provided multiple sequences are also identified for each structure. An improved reduced training set algorithm may lead to further increases in accuracy, in addition to providing diverse predictions for use in a jury-based method. Additional diverse predictions can be obtained by varying the hidden layer size of the first level secondary structure prediction network (Chandonia and Karplus, unpublished data). Although this is currently computationally expensive on large data sets, improvements in computer speed should make this procedure feasible in the near future. Rost and Sander[5] have shown that a more diverse jury can lead to improvements in accuracy of over 1%. Feature selection algorithms may also provide more accurate predictions by eliminating irrelevant inputs to the neural networks. Frishman and Argos[7] suggest that more accurate local alignment methods than those used in compiling the HSSP database will lead to additional improvements in accuracy. Together, these anticipated improvements should lead to predictions of over 80% average accuracy within several years, without explicit consideration of the effects of tertiary contacts.

CONCLUSIONS

New techniques for processing and decoding protein sequence data with neural networks, combined with a larger training database, improves the average accuracy of secondary structure prediction by 3.2% (to 74.8%) relative to previous studies. The algorithm produces an estimate of the overall accuracy of the prediction for each protein, and allows reliably predicted regions of the sequence to be identified. Structural class prediction accuracy improves by 6% (to 86%), mostly due to improved accuracy in the secondary structure prediction algorithm. The results suggest directions for future research in secondary structure prediction which may lead to over 80% average accuracy.

AVAILABILITY

Alignments of multiple sequences can be submitted for secondary structure and structural class prediction at the WWW URL http://www.cmpharm.ucsf.edu/~jmc/pred2ary/. The web site features a Java applet for making predictions online, and a Java version of the software which can be downloaded for use in most Java-compatible operating environments. Predictions can be made using a single network, or the juries of networks described in this paper. All predictions are made using the prediction set estimated accuracy method. Results obtained using the software currently available on the web site should therefore be comparable in accuracy to results described in the "Analysis of Accuracy" section of this paper. As the juries or the method are updated in the future, further details will be made available on the web site.

ACKNOWLEDGMENTS

We thank Andrej Šali for helping to compile the P681 protein database. This work was supported in part by grants from the National Science Foundation and the NCSA Supercomputer Center. The calculations were done on HP 9000/735, Intel Pentium Pro, SGI R10000, and DEC Alpha machines.

REFERENCES

1. Barton GJ. Protein secondary structure prediction. Curr Opin Struct Biol 1995;5:372–376.
2. Bohm G. New approaches in molecular structure prediction. Biophys Chem 1996;59:1–32.
3. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
4. Rost B, Sander C. Improved prediction of protein secondary structure by use of sequence profiles and neural networks. Proc Natl Acad Sci USA 1993;90:7558–7562.
5. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;19:55–72.
6. Chandonia JM, Karplus M. The importance of larger data sets for protein secondary structure prediction with neural networks. Protein Sci 1996;5:768–774.
7. Frishman D, Argos P. Seventy-five percent accuracy in protein secondary structure prediction. Proteins 1997;27:329–335.
8. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451.
9. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
10. Zhang X, Mesirov JP, Waltz DL. Hybrid system for protein secondary structure prediction. J Mol Biol 1992;225:1049–1063.
11. Rost B, Sander C. Prediction of protein secondary structure at better than 70% accuracy. J Mol Biol 1993;232:584–599.
12. Chandonia JM, Karplus M. Neural networks for secondary structure and structural class predictions. Protein Sci 1995;4:275–285.
13. Holley H, Karplus M. Neural networks for protein structure prediction. Meth Enzymol 1991;202:204–224.
14. Lesk AM. CASP2: Report on ab initio predictions. Proteins 1997;Suppl 1:151–166.
15. Šali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. Evaluation of comparative protein modeling by MODELLER. Proteins 1995; 23:318–326.
16. Kabsch W, Sander C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 1983;22:2577–2637.
17. Holley H, Karplus M. Protein secondary structure prediction with a neural network. Proc Natl Acad Sci USA 1989;86:52–156.
18. Dayhoff MO, editor. Atlas of protein sequence and structure Vol 5, Suppl 3. Washington, DC: National Biomedical Research Foundation; 1978.
19. Rumelhart DE, Hinton GE, Williams RJ. Learning internal representations by error propagation. In: Rumelhart DE, McClelland JL, editors. Parallel distributed processing. Vol 1. Cambridge, MA: MIT Press; 1986. p. 318–362.
20. Kneller DG, Cohen FE, Langridge R. Improvements in protein secondary structure prediction by an enhanced neural network. J Mol Biol 1990;214:171–182.
21. Dubchak IS, Holbrook KS. Prediction of folding class from amino acid composition. Proteins 1993;16:79–91.
22. Møller M. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 1993;6:525–533.
23. Baumruk V, Pancoska P, Keiderling TA. Predictions of secondary structure using statistical analyses of electronic and vibrational circular dichroism and Fourier transform infrared spectra of proteins in H2O. J Mol Biol 1996;259:774–791.
24. Rost B, Casadio R, Fariselli P, Sander C. Prediction of helical transmembrane segments at 95% accuracy. Protein Sci 1995;4:521–533.
25. Rost B, Fariselli P, Casadio R. Topology prediction for helical transmembrane proteins at 86% accuracy. Protein Sci 1996;5: 1704–1718.
26. Needleman S, Wunsch C. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
27. Solovyev V, Salamov A. Local secondary structure prediction using local alignments. J Mol Biol 1997;263:31–36.
28. Solovyev V, Salamov A. Predicting alpha-helix and beta-strand segments of globular proteins. CABIOS 1994;10:661–669.