Proc. Natl. Acad. Sci. USA
Vol. 86, pp. 152–156, January 1989
Biophysics

Protein secondary structure prediction with a neural network

L. HOWARD HOLLEY AND MARTIN KARPLUS

Department of Chemistry, Harvard University, Cambridge, MA 02138

Contributed by Martin Karplus, October 5, 1988

ABSTRACT A method is presented for protein secondary structure prediction based on a neural network. A training phase was used to teach the network to recognize the relation between secondary structure and amino acid sequences on a sample set of 48 proteins of known structure. On a separate test set of 14 proteins of known structure, the method achieved a maximum overall predictive accuracy of 63% for three states: helix, sheet, and coil. A numerical measure of helix and sheet tendency for each residue was obtained from the calculations. When predictions were filtered to include only the strongest 31% of predictions, the predictive accuracy rose to 79%.

Accurate prediction of protein secondary structure is a step toward the goal of understanding protein folding. A variety of methods have been proposed that make use of the physicochemical characteristics of the amino acids (1), sequence homology (2–4), pattern matching (5), and statistical analyses (6–11) of proteins of known structure. In a recent assessment (12) of three widely used methods (1, 6, 9), accuracy was found to range from 49% to 56% for predictions of three states: helix, sheet, and coil. The limited accuracy of the predictions is believed to be due to the small size of the data base and/or the fact that secondary structure is determined by tertiary interactions not included in the local sequence.

In this paper* we describe a secondary structure prediction method that makes use of neural networks. The neural network technique has its origins in efforts to produce a computer model of the information processing that takes place in the nervous system (13–16). A large number of simple, highly interconnected computational units (neurons) operate in parallel. Each unit integrates its inputs, which may be both excitatory and inhibitory, and according to some threshold generates an output, which is propagated to other units. In many applications, including the present work, the biological relevance of neural networks to nervous system function is unimportant. Rather, a neural network may simply be viewed as a highly parallel computational device. Neural networks have been shown to be useful in a variety of tasks including modeling content-addressable memory (17), solving certain optimization problems (18), and automating pattern recognition (19).

The neural networks used here for secondary structure prediction are of the feed-forward variety. These networks are organized into layers as shown in Fig. 1. Values of the input layer are propagated through one or more hidden layers to an output layer. Specialization of a neural network to a particular problem involves the choice of network topology—that is, the number of layers, the size of each layer, and the pattern of connections—and the assignment of connection strengths to each pair of connected units and of thresholds to each unit. Interest in such networks has been stimulated by the recent development of a learning rule for the automatic assignment of connection strengths and thresholds (20). In a "training" phase, initially random connection strengths (weights) and thresholds (biases) are modified in repeated cycles by use of a data set, in this case known protein structures. In each cycle adjustments are made to the weights and biases to reduce the total difference between desired and observed output. At the end of the training phase, the "knowledge" in the network consists of the connection strengths and thresholds that have been derived from the training data. This may be contrasted to pattern recognition by expert systems (5), in which "knowledge" of the problem domain lies in the rules that are supplied by the "expert."

METHODS

Data. The secondary structure assignment used is based on the work of Kabsch and Sander (21). Their program, DSSP, is used to classify known structures in the Brookhaven Protein Data Bank (22) as helices (H) and sheets (E); residues that are neither H nor E are classified as "coil." To facilitate a comparison with other prediction methods, the 62 proteins listed in table 1 of ref. 12 are used. The first 48 in the list (8315 residues) are taken as the training set and the last 14 (2441 residues) as the test set. The training set has a composition of 26% helix, 20% sheet, and 54% coil; the test set has a composition of 27% helix, 20% sheet, and 53% coil.

Network Formulation and Calculation. The network used for most of the calculations consists of an input layer, a single hidden layer, and an output layer (Fig. 1). The input layer encodes a moving window in the amino acid sequence and prediction is made for the central residue in the window. An initial window size of 17 is used based on evidence of statistical correlation with secondary structure as far as 8 residues on either side of the prediction point (9).

A binary encoding scheme is used for network input. In this scheme each amino acid at each window position is encoded by a group of 21 inputs, one for each possible amino acid type at that position and one to provide a null input used when the moving window overlaps the amino- or carboxyl-terminal end of the protein. In each group of 21 inputs, the input corresponding to the amino acid type at that window position is set to 1 and all other inputs are set to 0. Thus, the input layer consists of 17 groups of 21 inputs each and for any given 17 amino acid window, 17 network inputs are set to 1 and the rest are set to 0.

The hidden layer consists of two units. The output layer also consists of two units. Secondary structure is encoded in these output units as follows: (1,0) = helix, (0,1) = sheet, and (0,0) = coil. Actual computed output values are in the range 0.0–1.0 and are converted to predictions with the use of a threshold value, t. Helix is assigned to any group of four or more contiguous residues, the minimum helix in Kabsch and Sander classifications, having helix output values greater than sheet outputs and greater than t. β-Strand is assigned to any group of two or more contiguous residues, the minimum β-strand, having sheet output values greater than helix outputs and greater than t. Residues not assigned to helices

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

*A preliminary report of this work was presented at the Second Symposium of the Protein Society, August 13, 1988, San Diego, CA, Abstr. WK1008.

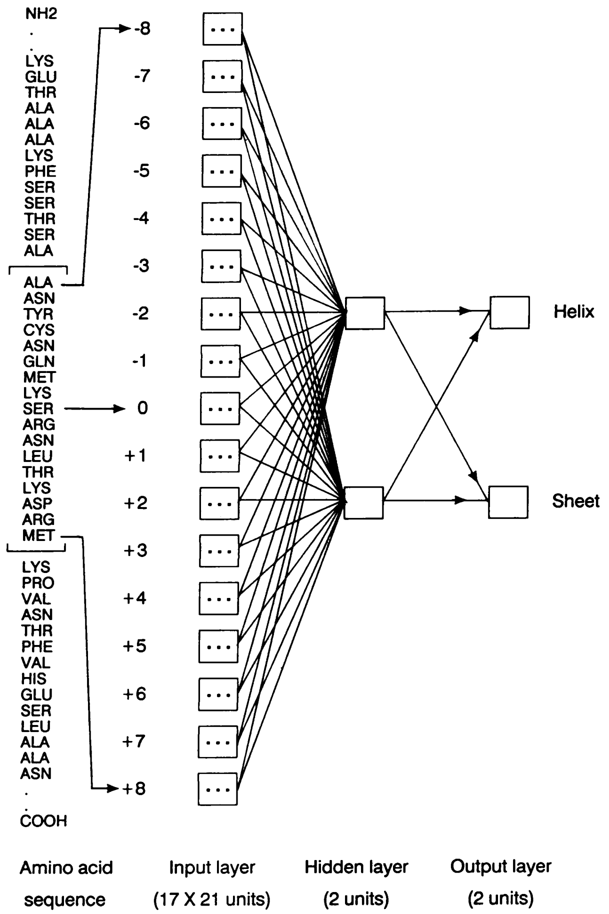

FIG. 1. Neural network topology. Each of the 17 blocks shown in the input layer represents a group of network inputs used to encode the amino acid at the corresponding window position. Each group consists of 21 inputs, one for each possible amino acid at that position plus a null input used when the moving window overlaps the end of the amino acid sequence. Thus, for a given window in the amino acid sequence, 17 of the 357 network inputs are set to 1 and the remainder are set to 0. A block in the hidden layer or in the output layer represents a single unit. Prediction is made for the central residue in the input window.

or β-strands are considered coil. The value of the threshold parameter, t, is adjusted by maximizing the accuracy of secondary structure assignment for the training set.

For both training and prediction, inputs are propagated forward through the network as illustrated in Fig. 2. (See ref. 20 for additional discussion.) The rule of Fig. 2 together with the network topology, the input encoding and output decoding described above, and the set of weights and biases produced by network training constitute a complete description of the prediction scheme.

The training procedure used is that described by Rumelhart *et al.* (19, 20). Network weights are initially assigned random values in the range −0.1 to 0.1. In each cycle, all of the training proteins are presented to the network and the input window moves through the amino acid sequences one residue at a time. At the end of the cycle, weights are adjusted and the procedure is repeated. The adjustments in network weights result in a gradient descent in the total output error defined by

$$E = \sum_{c} \sum_{j} (O_{j,c} - D_{j,c})^2, \qquad [1]$$

where $O_{j,c}$ is the observed output on unit j for training case c and $D_{j,c}$ is the desired output. Training is halted when the reduction in E becomes asymptotic (in practice when the fractional change in E per cycle is less than 2×-10^{-4}).

Evaluation of Results. To obtain suitable quality indices (23) for the comparison of secondary structure prediction algorithms, we consider the percent of total correct predictions and the correlation coefficients for each state: helix, sheet, and coil (24). The latter avoid the exaggeration of results due to overprediction when a single figure of merit is used. In addition, we report for each state S the percent of the residues that are correctly predicted to be in state S, $PC(S)$. This quality index (12) is a direct measure of the probability that a given prediction is correct.

RESULTS

The neural network shown in Fig. 1 was trained for 100 cycles over the 48 proteins. Training and prediction results are summarized in Table 1. A threshold value of $t = 0.37$ was found to produce a maximum percent correct for the training set (data not shown); this threshold value was used throughout. In Table 2 the results for the 14 proteins are compared with predictions from the methods of Chou and Fasman (6), Garnier *et al.* (9), and Lim (1) as reported by Kabsch and Sander (12).

A number of tests were made to determine the dependence of the accuracy of predictions on the network parameters. Table 3 shows the effect of varying window size on prediction accuracy. The optimum window size depends on the quality index chosen. Thus, although the output error reaches a minimum for a window size of 15, the correlation coefficients are somewhat better for a window size of 17. In general, prediction accuracy is a slowly increasing function of window size, confirming the importance of nearest-neighbor interactions.

Table 4 shows the effect of varying hidden-layer size.

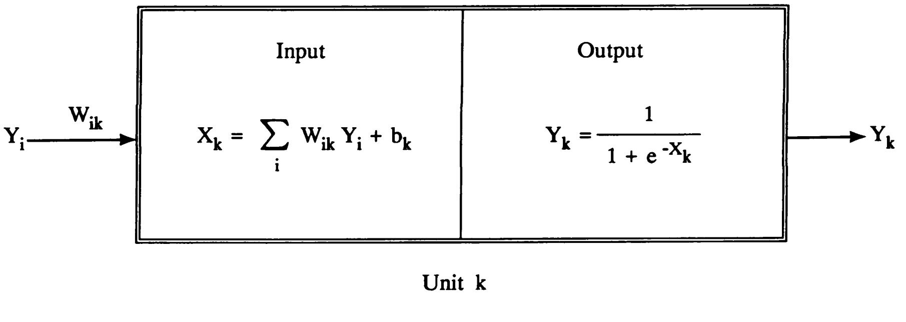
Input
Output
Y_i
W_{ik}
$X_k = \sum_i W_{ik} Y_i + b_k$
$Y_k = \frac{1}{1 + e^{-X_k}}$
Y_k
Unit k

FIG. 2. Network computation. During foward propagation through each layer of the neural network, the computation illustrated above takes place at each hidden unit and at each output unit. The products of outputs from the preceding layer, Y_i, with the connection strengths, W_{ik}, are summed over all inputs to the unit. The resulting sum is adjusted by the bias for the unit, b_k. The output of unit k is then generated according to the given formula and propagated to the next layer of the network. Unit outputs are in the range 0–1. Connection strengths may be either positive or negative.

Table 1. Summary of neural network training and prediction

Quality index	Training set (48 proteins)	Prediction set (14 proteins)
Percent correct (three-state total)	68.5	63.2
Correlation coefficients		
C_α	0.49	0.41
C_β	0.47	0.32
C_{coil}	0.43	0.36
Percent correct of predicted		
$PC(\alpha)$	65.3	59.2
$PC(\beta)$	63.4	53.3
PC(coil)	71.1	66.9

Substantially longer training was required for the larger networks; networks with 5, 10, or 20 hidden units were trained for 500 cycles. After training, the network with 20 hidden units achieved an accuracy on the training set of about 91%, but prediction accuracy fell to 59%. A 20-hidden-unit network involves over 7000 weights and biases. Since there are only 8315 residues in the training set, the free variables in the network are sufficient for it to learn to reproduce most of the "idiosyncratic" aspects of the training set. In the process, however, the network loses some of its ability to generalize, and prediction accuracy goes down. This potential for a loss of correlation between training accuracy and predictive accuracy is a fundamental limitation of neural networks.

The results with 0 hidden units are especially interesting. In this case the neural network has only two layers with inputs directly connected to outputs. It can be shown that there are pattern recognition problems that require at least one layer of hidden units (25). These are problems in which similarity clustering of inputs is not necessarily reflected in similarity of outputs. In our case, however, a neural network without hidden units is able to achieve a predictive accuracy that is close to optimal. The weights and biases for this 0-hidden-unit network are shown in Table 5. In this simple network the effect of any input sequence may be easily computed. From the formulae of Fig. 2 and the biases given in the footnote to Table 5, a helix output of 0.40, near the threshold value, requires a sum of weights from part A of Table 5 for the sequence to be about 0.46. To produce a helix output greater than 0.70, the sum from part A of Table 5 must be at least 1.72. To produce a sheet output of 0.40, the sum of weights from part B of Table 5 for the sequence must be about 1.26; a sheet output of 0.70 requires a sum of 2.52.

Table 2. Neural network predictive accuracy compared to other methods

Brookhaven identification	No. of residues	Percent correct predictions			
		Chou	Robson	Lim	Neural net
1GPD	333	47	55	58	66
4ADH	374	39	44	52	53
2GRS	461	45	49	48	65
2SOD	151	56	72	64	69
1LH1	153	52	69	50	71
1CRN	46	37	33	44	54
1OVO	56	48	54	77	65
2SSI	107	51	63	54	65
1CTX	71	68	65	69	72
1MLT	26	42	42	46	27
1NXB	62	50	61	60	71
2ADK	194	52	73	65	69
1RHD	293	55	54	54	65
2PAB	114	46	42	40	44
Total	2441	48	55	54	63

Comparative data are taken from table 1b of ref. 12. Totals are computed by weighting each protein by the number of residues. All coordinate sets are from the Brookhaven Protein Data Bank (22). Coordinate set 1LH1 is substituted for the obsolete 1HBL in the earlier work.

Table 3. Effect of input window size on prediction accuracy

Window size	Output error (E)	Percent correct	Correlation coefficients		
			C_α	C_β	C_{coil}
3	392.7	60.0	0.34	0.21	0.29
5	377.5	60.6	0.32	0.29	0.33
7	371.9	59.6	0.31	0.27	0.32
9	365.8	62.3	0.37	0.33	0.35
11	362.2	61.6	0.38	0.31	0.33
13	360.6	62.7	0.38	0.33	0.37
15	359.4	62.9	0.41	0.32	0.35
17	366.8	63.2	0.41	0.32	0.36
19	371.8	62.6	0.39	0.33	0.35
21	376.0	62.9	0.39	0.31	0.35

Neural networks are as shown in Fig. 1 except that input window sizes are those given in column 1. The total output error (E) is defined in Eq. 1.

The weights in Table 5 show, as expected, that glutamic acid, alanine, and leucine are favorable for helix prediction, whereas proline and glycine are unfavorable. Proline reaches an unfavorable peak at position +1 downstream from the prediction site and thus influences most strongly the preceding residue. Charged residues show an asymmetrical distribution with a helical preference for positive charge on the carboxyl side (arginine, lysine, and histidine) and negative charge (glutamic and aspartic acids) on the amino side of the prediction point. This is consistent with the hypothesis of stabilizing interactions between charged residues and the helix dipole (26). For sheet prediction it can be seen in part B of Table 5 that most hydrophobic residues are favorable around the center of the window. As one moves outward from the center of the window in part B of Table 5, the weights for many residues switch signs. This may reflect the fact that the majority of β-strands are shorter than 5 residues, in the Kabsch and Sander classifications (21). Tryptophan and methionine are interesting in that they have significantly larger weights away from the center of the window for both helix and sheet. Weights for the null input used to pad sequences at the chain termini (dashes in "Amino acid" column in Table 5) show that proximity to either the amino or the carboxyl terminus is unfavorable for both helix and sheet.

Correlates of Prediction Accuracy. The predictive accuracy for the first 20 residues from the amino terminus is 71% with $C_\alpha = 0.43$, $C_\beta = 0.56$, and $C_{\text{coil}} = 0.54$, in accord with earlier observations (27). Prediction accuracy drops to approximately average values when the first 50 residues are considered. No such trend is observed at the carboxyl terminus. Interior residues of helices and sheets (that is, excluding the first and last residue) are predicted more accurately. For interior helical residues $C_\alpha = 0.44$ and for helical termini $C_\alpha = 0.14$; for interior β-strand residues $C_\beta = 0.34$ and for strand termini $C_\beta = 0.19$. We have calculated solvent-accessible surface areas for the 14-protein test set to see whether core helical or β-strand residues are predicted more accurately than other residues. β-Strand residues that are less than 5%

Table 4. Effect of hidden-layer size

Hidden-layer size	Percent correct	
	Training	Prediction
0	68.4	62.3
2	68.5	63.2
5	81.5	60.9
10	89.9	59.5
20	90.9	59.3

Table 5. Weights and biases* on a trained neural network with no hidden units

Amino acid	Weight at window position																
	−8	−7	−6	−5	−4	−3	−2	−1	0	1	2	3	4	5	6	7	8
Part A. Input layer to helix output																	
-	0.10	0.20	-0.33	0.02	-0.27	-1.42	-1.56	-1.22	0.04	-1.39	-0.81	-0.45	-0.80	-0.39	0.10	-0.19	0.82
ALA	0.02	0.09	0.10	0.16	0.28	0.46	0.58	0.56	0.65	0.63	0.43	0.42	0.41	0.28	0.12	0.16	0.14
ARG	0.11	-0.05	-0.06	-0.04	-0.10	0.01	0.45	0.43	0.45	0.45	0.46	0.52	0.17	0.03	-0.17	0.05	-0.21
ASN	-0.11	-0.15	-0.01	-0.05	-0.17	-0.14	-0.30	-0.25	-0.55	-0.23	-0.29	-0.28	-0.17	-0.13	-0.18	-0.06	-0.09
ASP	-0.04	-0.05	-0.05	-0.07	0.04	0.06	-0.04	-0.06	-0.14	-0.18	-0.50	-0.45	-0.24	0.02	-0.02	-0.03	0.08
CYS	0.08	-0.15	-0.01	-0.19	-0.20	-0.04	-0.13	-0.54	-0.24	-0.02	-0.43	-0.60	-0.33	-0.43	-0.38	-0.48	-0.64
GLN	0.21	-0.03	-0.09	-0.26	-0.25	-0.18	-0.34	0.08	0.29	0.49	0.37	0.42	0.23	0.04	-0.27	-0.38	-0.39
GLU	0.45	0.41	0.48	0.51	0.53	0.62	0.59	0.68	0.76	0.69	0.25	-0.15	0.02	-0.24	-0.33	-0.29	-0.26
GLY	0.15	0.12	0.07	-0.08	-0.19	-0.26	-0.68	-1.00	-1.26	-0.80	-0.60	-0.62	-0.51	-0.58	-0.49	-0.46	-0.50
HIS	0.05	0.06	0.11	0.16	0.19	0.12	-0.03	0.22	0.25	0.54	0.43	0.61	0.59	0.58	0.51	0.62	0.63
ILE	-0.07	-0.13	-0.20	0.02	0.11	0.25	0.18	0.24	0.26	0.24	0.10	0.20	0.04	0.22	0.29	0.33	0.19
LEU	-0.23	-0.18	-0.04	0.13	-0.02	0.38	0.29	0.36	0.43	0.45	0.41	0.64	0.36	0.30	0.24	0.23	0.19
LYS	-0.22	-0.24	-0.05	-0.07	-0.14	-0.00	0.07	0.10	0.09	0.27	0.38	0.42	0.45	0.35	0.24	0.30	0.41
MET	-0.17	-0.21	0.04	-0.16	0.30	-0.05	0.33	0.34	0.27	0.58	0.64	0.66	0.45	0.40	0.21	0.43	0.23
PHE	-0.26	-0.17	-0.13	-0.05	-0.02	0.05	0.33	0.40	0.32	0.35	0.30	0.37	0.07	0.19	0.26	0.03	0.17
PRO	-0.53	-0.58	-0.57	-0.61	-0.58	-0.77	-0.66	-0.84	-1.15	-2.07	-1.27	-1.11	-1.02	-0.99	-0.95	-0.61	-0.62
SER	-0.20	-0.14	0.01	0.02	-0.11	-0.05	-0.22	-0.29	-0.63	-0.24	-0.38	-0.37	-0.24	-0.21	-0.34	-0.40	-0.37
THR	-0.20	-0.41	-0.30	-0.38	-0.36	-0.36	-0.28	-0.28	-0.37	-0.33	-0.42	-0.25	-0.19	-0.21	-0.25	-0.21	-0.20
TRP	0.37	0.50	0.28	0.08	-0.00	0.22	0.13	0.05	-0.20	-0.39	-0.55	-0.51	-0.14	-0.02	0.12	0.34	0.34
TYR	-0.00	0.05	0.04	-0.03	0.03	0.07	0.07	-0.22	-0.21	-0.03	0.00	0.12	0.04	0.13	0.14	-0.06	-0.45
VAL	0.08	-0.04	0.01	-0.06	-0.10	-0.01	-0.08	0.09	0.16	0.19	0.20	0.06	0.04	0.12	0.18	0.13	0.14
Part B. Input layer to sheet output																	
-	0.07	-0.32	0.98	0.05	-0.22	-1.41	-1.41	-1.50	0.10	-1.39	-0.42	-0.13	0.32	0.29	0.37	0.16	-0.78
ALA	-0.13	-0.14	-0.25	-0.33	-0.21	-0.18	-0.16	-0.15	-0.40	-0.27	-0.10	0.01	-0.09	-0.11	-0.15	-0.24	-0.25
ARG	-0.45	-0.30	0.03	0.02	-0.01	-0.05	-0.23	-0.27	-0.28	-0.28	-0.19	-0.50	-0.69	-0.60	-0.08	0.36	0.51
ASN	0.20	0.38	0.38	0.24	0.26	0.10	-0.57	-1.09	-0.74	-0.05	-0.08	0.07	0.15	0.10	0.37	0.36	0.21
ASP	-0.12	-0.13	-0.07	-0.01	0.04	-0.19	-0.66	-1.21	-1.31	-0.67	-0.48	-0.25	-0.01	-0.04	-0.03	-0.14	-0.20
CYS	-0.28	0.22	0.00	-0.32	-0.58	-0.81	-0.40	0.07	0.20	0.37	0.09	-0.29	-0.25	-0.06	-0.21	-0.37	-0.37
GLN	0.38	0.36	0.26	0.23	0.08	0.02	-0.02	0.11	0.05	-0.23	-0.18	-0.07	0.14	-0.11	-0.24	0.08	0.24
GLU	-0.36	-0.48	-0.21	-0.17	-0.14	-0.13	-0.55	-0.53	-0.61	-0.36	-0.31	-0.14	-0.30	-0.14	0.00	-0.34	-0.25
GLY	0.15	0.15	0.24	0.47	0.45	0.35	0.24	-0.52	-0.92	-0.24	0.23	0.42	0.51	0.36	0.47	0.12	0.22
HIS	0.03	0.07	-0.33	0.00	0.10	0.23	-0.23	-0.52	-0.34	-0.32	-0.43	0.25	0.17	0.11	0.11	-0.31	-0.11
ILE	-0.43	-0.20	-0.50	-0.16	0.00	0.19	0.52	0.99	1.02	0.86	0.21	-0.22	-0.43	-0.66	-0.51	-0.83	-0.23
LEU	-0.31	-0.35	-0.24	-0.34	-0.43	-0.29	0.27	0.50	0.52	0.50	0.21	-0.08	-0.33	-0.23	-0.20	-0.19	-0.13
LYS	-0.34	-0.39	0.13	0.13	0.08	0.17	-0.21	-0.32	-0.56	-0.72	-0.66	-0.42	-0.44	-0.66	-0.68	-0.39	-0.34
MET	0.01	-0.39	-0.60	-0.70	-0.73	0.07	0.26	0.49	0.67	0.20	-0.69	-1.53	-1.13	-0.35	-0.50	-0.21	-1.18
PHE	-0.23	-0.28	-0.50	-0.33	-0.21	-0.23	0.35	0.69	0.82	0.52	0.00	0.15	0.31	-0.01	-0.39	-0.61	-0.04
PRO	0.35	0.39	0.33	0.61	0.24	0.12	-0.17	-0.81	-2.08	-2.30	-0.59	0.12	0.20	0.24	-0.04	-0.22	-0.21
SER	0.37	0.23	0.18	0.19	0.26	0.09	-0.17	-0.26	-0.29	-0.04	0.25	0.38	0.42	0.48	0.51	0.34	0.27
THR	-0.08	0.07	0.31	0.41	0.29	0.23	0.37	0.60	0.31	0.13	0.26	0.39	0.08	0.22	0.34	0.37	0.36
TRP	-0.07	-0.52	-0.94	-1.12	-0.86	0.10	0.01	0.36	0.49	0.69	0.48	0.04	0.12	-0.22	-0.36	-0.03	0.29
TYR	0.05	0.12	-0.30	-0.31	-0.16	-0.10	0.55	0.60	0.60	0.31	0.19	0.12	-0.17	-0.09	0.27	0.27	0.16
VAL	-0.47	-0.33	-0.24	-0.08	-0.10	0.07	0.57	0.94	1.19	1.21	0.61	0.18	0.15	-0.06	-0.20	-0.14	-0.13

*Biases in the output layer are −0.87 for helix and −1.67 for sheet.

exposed, 45% of strand residues, are predicted with C_β = 0.33 versus C_β = 0.23 for more exposed residues.

To test whether the magnitude of the network outputs is any guide to the accuracy of predictions, we filtered the predictions by excluding those whose outputs fall into a range centered near the threshold, 0.37. Only helix or sheet predictions with network outputs greater than the upper bound of the excluded range were considered. Similarly, only coil predictions with both outputs below the lower bound of the excluded region were considered. Table 6 shows that there is a significant improvement in prediction accuracy for strong predictions. Indeed, the method rises to an overall accuracy of 79% for the strongest 31% of the database.

Physicochemical Encoding. An alternative is to encode amino acid sequence according to the physicochemical properties of the side chains. To test this idea a network was trained on the 48-protein training set with each amino acid categorized by hydrophobicity, charge, and backbone flexibility (that is, proline and glycine were treated as special cases). This network achieved a predictive accuracy on the 14-protein test set of 61.1% with correlation coefficients C_α = 0.37, C_β = 0.27, and C_{coil} = 0.37.

DISCUSSION

It can be seen that the neural network method is generally more accurate than prior methods when compared on identical proteins. The neural network method achieves this accuracy despite the fact that the method is naive in the sense that it does not rely on any theory of secondary structure

Table 6. Effect of prediction strength on accuracy

Excluded range	Database percent	Percent correct	Correlation coefficients		
			C_α	C_β	C_{coil}
None	100.0	63.2	0.41	0.32	0.36
0.30–0.50	80.9	67.0	0.45	0.37	0.43
0.20–0.60	59.9	71.5	0.51	0.44	0.49
0.10–0.70	31.4	78.5	0.63	0.56	0.62
0.05–0.75	13.3	84.0	0.75	0.71	0.78

Shown are predictions on the 14-protein test set for the neural network trained as described in results for Table 1. In computing quality measures, coil predictions are included only when network outputs are below the lower bound of the excluded range. Helix and sheet predictions are included only if the corresponding network output is above the upper limit of the excluded range. Shown in column 2 is the fraction of predictions included. Thus, the 31.4% strongest predictions are 78.5% accurate.

formation [e.g., the method of Lim (1)], nor does it require a sophisticated treatment of the data from which it is derived [e.g., advanced statistical methods (10, 11)].

An important observation from this method is that prediction strength is correlated with prediction accuracy. This observation was confirmed in predictions on an additional 20 proteins drawn from a database screened to include only structures with resolution better than 2.8 Å, crystallographic R factor less than or equal to 0.25, and sequence homology less than 50% (11). For this database the overall predictive accuracy is 63% and the 34% strongest predictions are 78% accurate (detailed data not shown). These strong predictions are distributed throughout the set of proteins. Thus, in applications of the method it may be useful to screen predictions using the magnitudes of network output to establish the presence of helix, β-strand, and coil with higher confidence without knowing the precise boundaries of these regions. Such predictions may be sufficient to determine that a protein belongs to a particular known class of chain fold (28, 29).

It is striking to observe the number of secondary structure methods that, although very different, are able to achieve about 60% accuracy. Although the present method is an incremental improvement, accuracy is still well below what is required for many applications, such as tertiary structure prediction, unless information concerning long-range interactions is available (30). It is tempting to suggest that these methods are approaching a limit dictated by the physics of protein structure, though database limitations may be involved as well. If the local sequence contributes only an approximately 60% tendency toward the formation of secondary structural units, dramatic improvement in predictions will require the introduction of information related to tertiary structure.

This work shows that neural networks can be applied to secondary structure prediction in proteins. It is likely that such networks will be useful in recognizing other structural motifs in protein and DNA sequences where only homology searching is currently employed.

Note. After the work reported here was completed, an independent study of the use of neural networks in secondary structure prediction was published (31). Although there are some significant differences in detail, the methodology and results are very similar.

We wish to thank Gregory A. Petsko for a discussion of neural networks and Eric Lander, Steve Harrison, and Don Wiley for reading an early version of this manuscript. The work was supported by the National Institutes of Health and the Department of Energy.

1. Lim, V. (1974) *J. Mol. Biol.* **88,** 873–894.
2. Levin, J. M., Robson, B. & Garnier, J. (1986) *FEBS Lett.* **205,** 303–308.
3. Nishikawa, K. & Ooi, T. (1986) *Biochim. Biophys. Acta* **871,** 45–54.
4. Zvelebil, M., Barton, G., Taylor, W. & Sternberg, M. (1986) *J. Mol. Biol.* **195,** 957–961.
5. Cohen, F., Abarbanel, R., Kuntz, I. & Fletterick, R. (1986) *Biochemistry* **25,** 266–275.
6. Chou, P. & Fasman, G. (1974) *Biochemistry* **13,** 222–245.
7. Wu, T. T. & Kabat, E. A. (1973) *J. Mol. Biol.* **75,** 13–31.
8. Nagano, K. (1977) *J. Mol. Biol.* **109,** 251–274.
9. Garnier, J., Osguthorpe, D. & Robson, B. (1978) *J. Mol. Biol.* **120,** 97–120.
10. Maxfield, F. & Scheraga, H. (1979) *Biochemistry* **18,** 697–704.
11. Gibrat, J., Garnier, J. & Robson, B. (1987) *J. Mol. Biol.* **198,** 425–443.
12. Kabsch, W. & Sander, C. (1983) *FEBS Lett.* **155,** 179–182.
13. McCulloch, W. & Pitts, W. (1943) *Bull. Math. Biophys.* **5,** 115–123.
14. Heeb, D. (1949) *The Organization of Behavior* (Wiley, New York).
15. Minsky, M. (1954) Unpublished Doctoral Dissertation (Princeton Univ., Princeton, NJ).
16. Rosenblatt, F. (1962) *Principles of Neurodynamics* (Spartan, New York).
17. Hopfield, J. (1982) *Proc. Natl. Acad. Sci. USA* **79,** 2554–2558.
18. Hopfield, J. & Tank, D. (1986) *Science* **233,** 625–633.
19. Rumelhart, D., Hinton, G. & Williams, R. (1986) *Nature (London)* **323,** 533–536.
20. Rumelhart, D., Hinton, G. & Williams, R. (1986) in *Parallel Distributed Processing*, eds. Rumelhart, D. E. & McClelland, J. E. (MIT Press, Cambridge, MA), Vol. 1, pp. 318–362.
21. Kabsch, W. & Sander, C. (1983) *Biopolymers* **22,** 2577–2637.
22. Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) *J. Mol. Biol.* **112,** 535–542.
23. Schulz, G. & Schirmer, R. (1978) *Principles of Protein Structure* (Springer, New York), pp. 123–125.
24. Matthews, B. (1975) *Biochim. Biophys. Acta* **405,** 442–451.
25. Minsky, M. & Papert, S. (1969) *Perceptrons* (MIT Press, Cambridge, MA).
26. Shoemaker, K., Kim, P., York, E., Stewart, J. & Baldwin, R. (1987) *Nature (London)* **326,** 563–567.
27. Argos, P., Schwarz, J. & Schwarz, J. (1976) *Biochim. Biophys. Acta* **439,** 261–273.
28. Schulz, G. E. (1988) *Annu. Rev. Biophys. Biophys. Chem.* **17,** 1–21.
29. Crawford, I. P., Niermann, T. & Kirschner, K. (1987) *Proteins Struct. Func. Genet.* **2,** 118–129.
30. Burgess, A. & Scheraga, H. (1975) *Proc. Natl. Acad. Sci. USA* **72,** 1221–1225.
31. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202,** 865–884.