BUTTERWORTH HEINEMANN

Feed-forward neural networks for secondary structure prediction

T.W. Barlow

Physical Chemistry Laboratory, Oxford, England

A feed-forward neural network has been employed for protein secondary structure prediction. Attempts were made to improve on previous prediction accuracies using a hierarchical mixture of experts (HME). In this method input data are clustered and used to train a series of different networks. Application of an HME to the prediction of protein secondary structure is shown to provide no advantages over a single network. We have also tried various new input representations, chosen to incorporate the effect of residues a long distance away in the one-dimensional amino acid chain. Prediction accuracy using these methods is comparable to that achieved by other neural networks.[1–4]

Keywords: protein secondary structure, neural networks, hierarchical mixtures of experts, pattern recognition

INTRODUCTION

Theoretical methods that could reliably predict the secondary or tertiary structure of any sequence of amino acids would be invaluable. Ever since experiments showed that a protein could be denatured and then reform in its original conformation,[5] it has been assumed that all the information required for a protein to form its tertiary structure is contained within its primary sequence of amino acids. It is thought that if the secondary structure of a protein can be accurately predicted, then the tertiary structure would be much easier to solve. One of the more successful methods of predicting secondary structure has been with artificial neural networks.

AN ARTIFICIAL NEURAL NETWORK

The neural network we use is a simple feed-forward network. It consists of layers of computational units as shown in Figure 1. Connections between units are weighted so that each unit takes as its input the weighted outputs from the preceding layer. Figure 2 shows how each unit in the network processes its input to generate an output. The weighted sum of all inputs to a given unit is termed "the activation." The activation a_j of unit j is a sum of the outputs from the previous layer multiplied by the connection weights between layers.

$$a_j = \sum_k w_{jk} x_k \tag{1}$$

where w_{jk} is the weight of the connection from unit k to unit j, and x_k is the output from unit k.

The output from any unit depends on the activation function. A common activation function is the sigma function:

$$g(a_j) = \frac{1}{1 + e^{-a_j}} \tag{2}$$

as shown in Figure 3. The network uses Eqs. (1) and (2) at every unit to compute output patterns given any input pattern and a set of connection weights.

The weights of such a network can be optimized so that input patterns match to particular output patterns. This is achieved using an algorithm called "back propagation." Starting with randomly assigned weights, the network is trained with a set of corresponding pairs of input and output patterns. For every input pattern in this training set the actual output can be compared to the desired output pattern to produce a cost function:

$$E = \frac{1}{2} \sum_i [\zeta_i - O_i]^2 \tag{3}$$

ζ_i is the target output for unit i, and O_i is the actual output for unit i.

The most common method for minimizing the cost function E is by steepest descent.[6–8] In this case, for every pattern in the training set, weights are updated proportionally to the negative of the gradient of the cost in weight space. That is, according to the following rule,

$$\Delta w_{ij} = -\eta \frac{\partial E}{\partial w_{ij}} \tag{4}$$

Address reprint requests to Dr. Barlow at the Physical Chemistry Laboratory, South Parks Road, Oxford OX1 3QZ England.
Received 22 February 1995; accepted 7 March 1995

655 Avenue of the Americas, New York, NY 10010

0263-7855/95/$10.00
SSDI 0263-7855(95)00016-Y

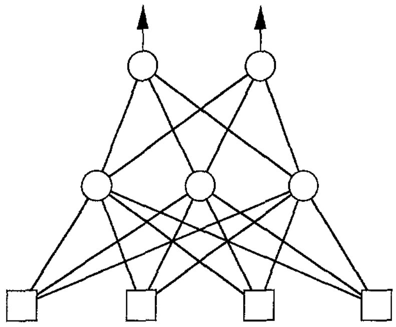

Figure 1. An example of a general feed-forward neural network. Input units are labeled as squares; all other units are shown as circles.

where η is a parameter defining the learning rate and w_{ij} is the weight being updated. The back propagation learning algorithm is summarized as follows.

1. Initialize network weights to small random values.
2. Choose an input pattern and apply to input layer.
3. Propagate pattern through the network according to Eqs. (1) and (2) to calculate final outputs.
4. Use Eq. (4) to update every connection.
5. Return to step 2 and cycle.

Once a network has been "trained" by this procedure with a set of input patterns and corresponding output patterns, novel data can be fed through the network to test its ability to generalize.

A NEURAL NETWORK FOR STRUCTURE PREDICTION

In this study, the Xerion[9] software package has been used to construct a network as shown in Figure 4. The input patterns are sparsely coded representations of amino acid sequence. Each amino acid is represented by a 21-component vector with 1 component equal to 1.0 and the rest to 0.0. (One of the units is a spacer unit.) A window of 13 amino acids, thus coded, forms each input pattern. Thus there are 21 × 13 = 273 input units. There are three output units, one for each class of secondary structure. Thus α helix is coded by (1,0,0), β sheet by (0,1,0), and random coil by (0,0,1). For prediction purposes, the output unit with the largest output value is used to allocate the secondary structure.

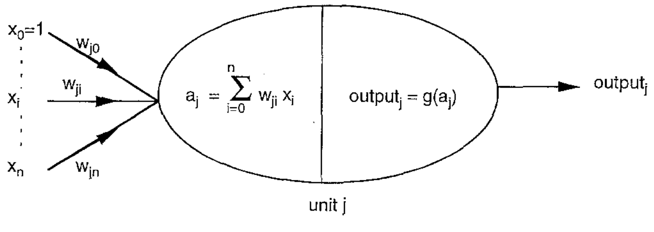

Figure 2. A computational unit from an artificial neural network. Inputs to the unit are labeled x_i. *These could be inputs to the network, or outputs from other units in the network.* w_{ji} *is the weight of the connection between unit* i *and unit* j. a_j *is the activation of unit* j. *One of the inputs,* x_0, *is fixed equal to one. Thus the weight* w_{j0} *acts as a bias on the activation. The output is a function of the activation. For this work,* $g(a_j) = [1 + e^{-a_j}]^{-1}$.

The data used to train the network were derived from 105 proteins obtained from the Brookhaven Protein Data Bank.[10] These proteins are listed in Table 1.[11–112] This set of proteins is similar to that employed by Qian and Sejnowski[1] and by Kneller et al.[3] for their neural networks. It differs from that of the latter group in the case of a few proteins for which updated crystal structures were available.

The secondary structure was assigned using QUANTA.[113] This assignment is based on the method of Kabsch and Sander[114,115] (whose software was used by Kneller et al. in their work). If an amino acid has an undefined position in the crystal structure, no secondary structure is assigned. This is occasionally the case for residues in the chosen database. In those cases the residues were not presented to the neural network because no valid secondary structure allocations could be made. This left a total of 20 462 residues, slightly less than the number in the database of Kneller et al., which included all residues. Of the total number of residues, 5 549 were in α helices, 4 714 were in β sheet, and 10 251 were in random coils.

A network with no hidden units was trained using 91 of the proteins in the data set. The ability to generalize was then tested using the remaining 14 proteins. These 14 proteins were the same as those chosen by Qian and Sejnowski[1] to minimize homologies between the training and testing sets. We assume that the same 14 proteins were also used by Kneller et al. An architecture was chosen with no hidden units. Previous studies[1,2,116] have found such an architecture to give similar results to those with hidden units.

Prediction accuracy for the network is determined as the percentage of correct predictions:

$$Q_3 = \frac{1}{N}(p_\alpha + p_\beta + p_{\text{coil}}) \tag{5}$$

where N is the total number of residues, and p_s is the number of residues of type s predicted correctly.

Matthew's correlation function[117] gives a measure of accuracy for each secondary structure type, taking overprediction into account.

$$c_s = \frac{p_s n_s - u_s o_s}{[(n_s + u_s)(n_s + o_s)(p_s + u_s)(p_s + u_s)]^{1/2}} \tag{6}$$

where p_s is the number of residues of type s predicted correctly, n_s is the number of residues of type s properly rejected, u_s is the number of underpredicted residues of type s, o_s is the number of overpredicted residues of type s.

Our results were as follows. After training for 100 epochs (i.e., 100 passes of the training data through the network) the network was able to predict the secondary structure of the training set with the following accuracy: $Q_3 = 0.64$; $c_1 = 0.42$; $c_2 = 0.37$; $c_3 = 0.42$. When the test set was subsequently presented to the network the prediction accuracy was $Q_3 = 0.61$; $c_1 = 0.38$; $c_2 = 0.27$; $c_3 = 0.41$. This result is slightly poorer than previous results we were attempting to reproduce. (For example, Kneller et al.[3] achieved $Q_3 = 0.63$; $c_1 = 0.35$; $c_2 = 0.30$; $c_3 = 0.41$ for

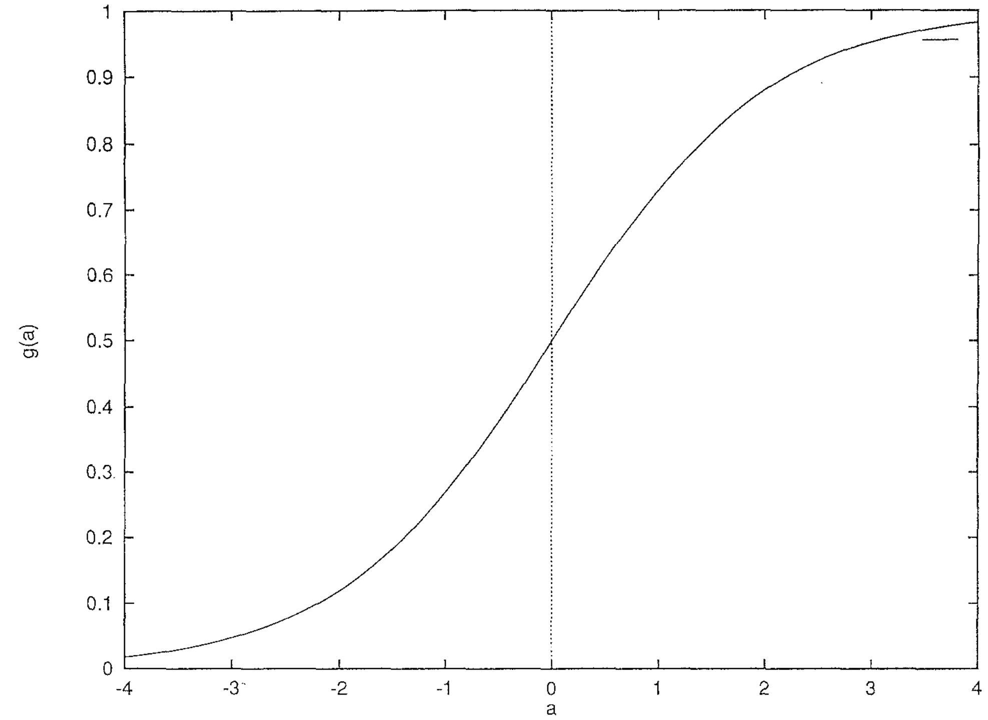

Figure 3. The activation function (sigma function) used to calculate the final output from a unit in the neural network. $g(a_j) = [1 + e^{-a_j}]^{-1}$.

the same test set.) We attribute the slight difference in results to the differences in the input data, as discussed above.

HIERARCHICAL MIXTURES OF EXPERTS

It can be useful in training a neural network to cluster the input or output spaces. One method of doing this is by using a hierarchical mixture of experts (HME). Such a system is summarized in Figure 5.

The process is as follows.

1. Train a net on the complete data.
2. Split the training set into correct and incorrect sets (i.e., data that the first net learned correctly and that which it failed to learn correctly.
3. Train a second network on the incorrect set.
4. Train a third network to decide which network to use for new data.

We have tried to apply this method to improve the prediction accuracies of our original network. The 36% of the training set that the first network failed to learn, was used to train a second network. The second network learned these data with an accuracy of $Q_3 = 0.55$; $c_1 = 0.32$; $c_2 = 0.27$; $c_3 = 0.38$. However, this network proved unable to generalize. On the test data its accuracy was much poorer: $Q_3 = 0.25$. This result is worse than one would get by randomly guessing secondary structure.

A less sophisticated variation of the HME is simply to train a network to distinguish data that the first network learned correctly from that which it learned incorrectly. This corresponds to stage d above. At least then, one would know which predictions were correct. The remaining predictions could be randomly reassigned one of the other possible secondary structures. There would now be a 50% chance that these previously incorrect predictions were correct.

Consider the case in which

P_1 = proportion of correctly assigned data from net I, or prediction accuracy of net I
P_2 = proportion of incorrect predictions that would become correct by random reassignment (0.50)
P_3 = proportion of data from net I distinguished correctly as correct or incorrect by net III (net from stage d), i.e., prediction accuracy of net III

After random reassignment of those predictions identified as incorrect by net III the overall secondary structure prediction accuracy becomes

$$Q_3 = P_1P_3 + P_2P_3\{1 - P_1\} \qquad (7)$$

In our case, $P_1 \simeq 0.6$ and $P_2 = 0.5$. In order that $Q_3 > P_1$, it is necessary that we train a net to decide which data

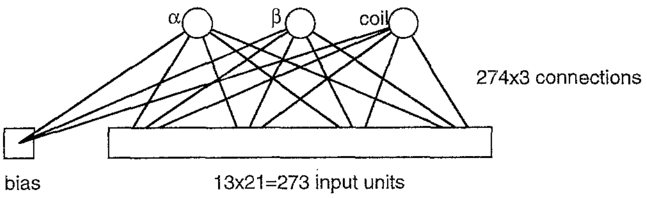

Figure 4. The network architecture used to predict secondary protein structure. The input pattern corresponds to a window of amino acids centred about a central amino acid. Each amino acid is encoded over 21 units. Every input unit is connected to the three output units. Each output units represents a different class of secondary structure.

Table 1. Protein database for secondary structure prediction[a]

Code	Protein name	Unit	Residues	Helix	Sheet	Coil	Ref.
1ABE*	L-Arabinose-binding protein		305	137	62	106	11
1ACX*	Actinoxanthin		108	0	49	59	12
1AZU	Azurin		126	9	35	82	13
1BP2	Phospholipase A_2		123	56	11	56	14
1CA2	Carbonic anhydrase II		256	16	85	155	15
1CC5	Cytochrome c_5(oxidized)		83	41	2	40	16
1CCR	Cytochrome *c*		111	41	5	65	17
1CRN	Crambin		46	19	9	18	18
1CTX	Alpha cobratoxin		71	4	16	51	19
1CYC	Ferrocytochrome *c*		103	31	0	72	20
1ECD	Hemoglobin (erythocruorin, deoxy)		136	99	0	37	21
1EST	Tosyl-elastase		240	17	93	130	22
1FC2	Immunoglobulin Fc	D	207	16	93	98	23
1FDH	Hemoglobin (deoxy human fetal F_{II})	A	141	100	0	41	24
		G	146	101	0	45	24
1FDX	Ferredoxin		54	5	11	38	25
1FX1	Flavodoxin		147	55	33	59	26
1GCN	Glucagon		29	12	0	17	27
1GF1	Insulin-like growth factor I		70	20	2	48	28
1GF2	Insulin-like growth factor II		67	20	6	41	28
1GP1	Glutathione peroxidase	B	184	45	30	109	29
1HDS	Hemoglobin (deer sickle cell)	A	141	79	0	62	30
		B	145	80	0	65	30
1HIP	Oxidized high-potential iron protein		85	10	20	55	31
1LZ1	Lysozyme		130	39	15	76	32
1LZT	Lysozyme (triclinic crystal form)		129	38	15	79	33
1MBD	Myoglobin (sperm whale, deoxy)		153	112	0	41	34
1MBS	Myoglobin (seal, met)		153	111	0	42	35
1NXB*	Neurotoxin b		62	0	27	35	36
1PFC	pFe(prime) fragment (IgG1)		111	4	35	72	37
1PPD*	2-Hydoxyethylthiopapain-crystal form D		212	50	45	117	38
1PPT	Avian pancreatic polypeptide		36	19	0	17	39
1PYP*	Inorganic pyrophosphatase		281	31	47	203	40
1REI	Bence–Jones immunoglobulin REI var. portion	A	107	0	60	47	41
1RHD	Rhodanese		293	85	37	171	42
1TGS	Trypsinogen/pancreatic secretory inhibitor	Z	225	20	87	118	43
1TIM	Triose phosphate isomerase	A	247	109	41	97	44
2ACT*	Actinidin (sulfhydryl proteinase)		218	55	51	112	45
2ALP*	Alpha-lytic protease		198	7	114	77	46
2APR	Acid proteinase (rhizopuspepsin)		325	30	143	152	47
2AZA	Azurin (oxidized)	A	129	17	48	64	48
2CAB	Carbonic anhydrase form B		256	18	85	153	49
2CCY	Cytochrome *c* (prime)	A	127	86	2	39	50
2CDV*	Cytochrome c_3		107	23	12	72	51
2CY3	Cytochrome c_3		118	31	6	81	52
2CYP	Cytochrome *c* peroxidase (ferrocytochrome *c*)		293	134	18	141	53
2DHB	Hemoglobin (horse, deoxy)	A	141	98	0	43	54
		B	146	96	0	50	54
2GCH	Gamma chymotrypsin A		237	21	85	131	55
2GN5	Gene 5 DNA-binding protein		87	0	12	131	56
2HMQ*	Hemeruthrin (met)	A	113	77	0	36	57
2IG2	Immunoglobulin G1	L	256	16	104	136	58
		H	239	16	114	109	58
2KAI	Kallikrein A (bovine, pancreatic trypsin)	A	80	0	37	43	59
		B	152	12	45	95	59
2LDX	Apo-lactate dehydrogenase isoenzyme C_4		331	120	59	152	60
2LH1	Leghemoglobin (acetate, met)		153	108	0	45	61

Table 1. *Continued*

Code	Protein name	Unit	Residues	Helix	Sheet	Coil	Ref.
2LHB*	Hemoglobin V (cyano, met)		149	98	0	51	62
2LZM	Lysozyme (bacteriophage T4)		164	105	10	49	63
2MCP	Immunoglobulin McPC603 with Fab-phosphocholine	L	220	9	106	105	64
		H	222	6	117	99	64
2MLT	Melittin	A	26	23	0	3	65
2PAB	Prealbumin (human plasma)	A	114	7	61	46	66
2RHE	Bence–Jones protein (lambda variable domain)		114	6	58	50	67
2SBT*	Subtilisin novo		275	59	27	189	68
2SGA	Proteinase A		181	11	106	64	69
2SN3	Scorpion neurotoxin (variant 3)		65	7	19	39	70
2SNS	Staphylococcal nuclease complex		141	34	27	80	71
2SOD	Cu, Zn superoxide dismutase	O	151	0	70	81	72
2SSI	*Streptomyces* subtilisin inhibitor		107	18	26	63	73
2STV	Satellite tobacco necrosis virus		184	10	91	83	74
2TAA	Taka-amylase A		478	90	74	314	75
2TBV	Tomato bushy stunt virus	A	321	4	125	192	76
3ADK	Adenylate kinase		194	110	23	61	77
3APP	Acid proteinase (penicillopepsin)		323	25	146	152	69
3B5C	Cytochrome b_5 (oxidized)		86	27	19	40	78
3C2C	Cytochrome c_2 (reduced)		112	48	5	59	79
3CNA	Concanavalin A		237	0	93	144	80
3FXC	Ferredoxin		98	9	19	70	81
3GPD*	D-Glyceraldehyde-3-phosphate dehydrogenase	R	334	85	77	172	82
3GRS*	Glutathione reductase		461	136	116	209	83
3HHB	Hemoglobin (human, deoxy)	A	141	100	0	41	84
		B	146	107	0	39	84
3ICB	Calcium-binding protein		75	39	2	34	85
3PCY	Plastocyanin (Hg^{2+} substituted)		99	6	38	55	86
3PGK	Phosphoglycerate kinase complex with ATP		415	136	43	236	87
3PGM	Phosphoglycerate mutase (dephospho enzyme)		230	64	17	149	88
3RN3	Ribonuclease A		124	22	46	56	89
3RP2	Rat mast cell protease II	A	224	11	90	123	90
2SGB	Proteinase B from streptomyces griseus	E	185	11	106	68	91
451C	Cytochrome c_{551} (reduced)		82	38	2	42	92
4APE	Acid proteinase endothiapepsin		330	24	144	162	93
4CTS	Citrate synthase oxaloacetate complex	A	437	250	15	172	94
4DFR	Dihydrofolate reductase complex	A	159	34	53	72	95
4FXN	Flavodoxin (semiquinone form)		138	49	30	59	96
4GCR	Gamma-B crystallin		174	5	69	100	97
4INS	Insulin	A	21	11	2	8	48
		B	30	11	4	15	48
4MDH	Cytoplasmic malate dehydrogenase	A	333	130	65	138	98
4MT2	Metallothionein isoform II		61	0	2	59	99
4P2P	Phospholipase A_2 (porcine)		124	56	9	59	100
4SBV	Southern bean mosaic virus coat protein	A	222	26	76	120	101
5AT1	Aspartate carbamoyltransferase	A	310	111	45	154	102
		B	146	17	53	76	102
5CPA	Carboxypeptidase A_α(Cox)		307	115	56	136	103
5CPV	Calcium-binding parvalbumin B		108	56	2	50	104
5FD1	Ferredoxin (oxidized)		106	15	13	78	105
5LDH	Lactate dehydrogenase H_4 with S-Ic-NAD$^+$		333	117	35	181	106
5PTI	Trypsin inhibitor (crystal form II)		58	8	13	37	107
5RXN	Rubredoxin (oxidized, Fe(III))		54	0	10	37	108
6ADH	Holo-liver dehydrogenase complex	A	374	85	82	207	109
8API*	Modified α_1-antitrypsin	A	340	105	120	115	110
		B	36	0	16	20	
8CAT	Catalase	A	498	142	79	277	111

Table 1. ***Continued***

Code	Protein name	Unit	Residues	Helix	Sheet	Coil	Ref.
8TLN	Thermolysin		316	125	56	135	112

[a] Protein coordinates were taken from the Brookhaven Protein Data Bank.[10] Shown here are the number of residues assigned a secondary structure for each protein.

are correctly and which incorrectly predicted to an accuracy of $P_3 > 0.75$. The best result we managed to obtain trying different architectures and inputs representations was $Q_3 = 0.63$.

LOCAL AND GLOBAL EFFECTS

Previous attempts to predict secondary structure using neural networks[1–4,118] have used similar representations for amino acid sequences. They have tended to do as we have done above, coding for each amino acid with a local representation. We tried various alternatives using more compact, distributed representations, scaled according to various criteria such as the consensus hydrophobicity scale of Eisenberg et al.[119] None of these alternatives improved prediction accuracy.

In addition, all previous studies have considered only amino acids in a local window along the protein chain. Qian and Sejnowski[1] tested various window sizes from 1 to 21 amino acids wide, finding 13 to be the optimal size. This is essentially a local approach to the problem, which neglects possible global effects. It has been suggested that the apparent limit on current secondary structure prediction methods at 65% accuracy may be a result of this assumption. It is impossible to include every amino acid in a protein as the input for a network, as many proteins are too large. How, then, should some sort of global effect be incorporated into the input patterns? A naive argument might suggest we need 35% of units devoted to this purpose.

The hydrophobic moment[119,120] gives a measure of periodicity in the hydrophobicity of a sequence of amino acids. It is defined by Eq. (8):

$$\mu_i = \frac{1}{N}(\mu_{ix}\mu_{ix} + \mu_{iy}\mu_{iy})^{1/2} \tag{8}$$

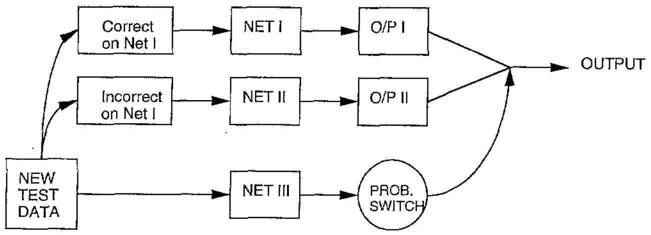

Figure 5. A hierarchical mixture of experts. Net I is trained on the complete data set. Those that net I fails to learn correctly are used to train net II. Net III is then trained on the complete data to learn which of the first two nets of new data should be put through.

μ_i is the hydrophobic moment of a sequence containing N residues about residue i. $\mu_{ix} = \Sigma_j\phi_j\cos[(j - i)\omega]$ and $\mu_{iy} = \Sigma_j\phi_j\sin[(j - i)\omega]$, where ϕ_i is the hydrophobicity of residue j and ω is a characteristic frequency. For the helical hydrophobic moment $\omega = 2\pi/3.6$. Kneller et al.[3] tried including an additional unit in their network giving the helical hydrophobic moment of the local window. They found that this improved the training accuracy (but not the prediction accuracy) of their network by 1%.

We tried using as inputs various combinations of hydrophobic moments calculated over different chain lengths. It was our hope that hydrophobic moments calculated for residues outside the local window could provide a means of including global information about the protein. We succeeded in improving the training accuracy to 67% (an improvement of 3%). However, such extra inputs had no significant effect on the prediction accuracy.

CONCLUSIONS

Neural networks have been applied to the prediction of protein secondary structure. Several new ideas have been tried. However, none have produced predictions of greater accuracy than any earlier attempts. In particular, ideas that might have incorporated global effects into network input patterns were shown to be ineffectual. It is significant that with a slightly different database (containing updated structures and not including amino acids with inaccurately specified positions) we achieved prediction accuracies marginally lower than those of Kneller et al.[3] While predictions remain at this level (60–65%) the method has little practical use. We are currently exploring an alternative approach to the problem.

ACKNOWLEDGMENTS

The author is grateful for the support he receives from Kobe Steel, Ltd. and from the ORS award scheme through CVCP. He is thankful for helpful discussions with Professor Ah-Chung Tsoi of the Electrical Engineering Department at the University of Queensland and Simon Blackburn in the Electrical Engineering Department at Cambridge University.

REFERENCES

1 Qian, N. and Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. *J. Mol. Biol.* 1988, **202,** 865

2 Holley, H. and Karplus, M. Protein secondary struc-

ture prediction with a neural network. *Proc. Natl. Acad. Sci. U.S.A.* 1989, **86,** 152

3 Kneller, D.G., Cohen, F.E., and Langridge, R. Improvements to protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 1990, **214,** 171

4 Stolorz, P., Lapedes, A., and Xia, Y. Predicting protein secondary structure using neural net and statistical methods. *J. Mol. Biol.* 1992, **225,** 363

5 Anfinsen, C.B., Haber, E., Sela, M., and White, F.H., Jr. *Proc. Nat. Acad. Sci. U.S.A.* 1961, **47,** 1309

6 Hertz, J., Krogh, A., and Palmer, R.G. *Introduction to the Theory of Neural Computing*. Addison-Wesley, Reading, Massachusetts, 1991

7 Lippmann, R.P. An introduction to computing with neural nets. *IEEE, ASSP Mag.* 1987, April: 4–22

8 Wassermann, P.D. and Schwartz, T. Neural networks: What are they and why is everybody so interested in them? *IEEE Expert* 1987, Winter: 10–14

9 van Camp, D. *XERION 3.1*. Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, 1993

10 Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Jr., Rodgers, J.R., Kennard, O., Shimanouchi, T., and Tasumi, M. *J. Mol. Biol.* 1977, **112,** 535

11 Vyas, N.K. and Quiocho, F.A., *Nature* (*London*) 1984, **310,** 381

12 Pletnev, V.Z., Kuzin, A.P., and Malinina, L.V. *Bioorg. Khim.* 1982, **8,** 1637

13 Adman, E.T. and Jensen, L.H. *Isr. J. Chem.* 1981, **21,** 8

14 Dijkstra, B.W., Kalk, K.H., Hol, W.G.J., and Drenth, J. *J. Mol. Biol.* 1981, **147,** 97

15 Eriksson, A.E., Jones, T.A., and Liljas, A. *Proteins Struct. Funct.* 1988, **4,** 274

16 Carter, D.C., Melis, K.A., O'Donnell, S.E., Burgess, B.K., Furey, W.F., Jr., Wang, B.-C., and Stout, C.D. *J. Mol. Biol.* 1985, **184,** 279

17 Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S., and Morita, Y. *J. Mol. Biol.* 1983, **166,** 407

18 Hendrickson, W.A. and Teeter, M.M. *Nature* (*London*) 1981, **290,** 107

19 Walkinshaw, M.D., Saenger, W., and Maelick, A. *Proc. Natl. Acad. Sci. U.S.A.* 1980, **70,** 2400

20 Tanaka, N., Yamane T., Tsukihara, T., Ashida, T., and Kakudo, M. *J. Biochem.* (*Tokyo*) 1975, **1975,** 147

21 Steigemann, W. and Weber, E. *J. Mol. Biol.* 1979, **127,** 309

22 Sawyer, L., Shotton, D.M., Campbell, J.W., Wendell, P.L., Muirhead, H., Watson, H.C., Diamond, R., and Ladner, R.C. *J. Mol. Biol.* 1978, **118,** 137

23 Deisenhofer, J. *Biochemistry* 1981, **20,** 2361

24 Frier, J.A. and Perutz, M.F. *J. Mol. Biol.* 1977, **112,** 97

25 Adman, E.T., Sieker, L.C., and Jensen, L.H. *J. Biol. Chem.* 1976, **251,** 3801

26 Watenpaugh, K.D., Sieker, L.C., and Maelicke, A. *Proc. Natl. Acad. Sci. U.S.A.* 1973, **70,** 3857

27 Blundell, T.L., Sasaki, K., Dockerill, S., and Tickle, I.J. *Nature* (*London*) 1975, **257,** 751

28 Blundell, T.L., Debarkar, S., and Humbel, R.E. *Fed. Proc. Fed. Am. Soc. Exp.* 1983, **42,** 2592

29 Epp, O., Ladenstein, R., and Wendel, A. *Eur. J. Biochem.* 1983, **133,** 51

30 Girling, R.L., Houston, T.E., Junior, W.C.S., and Amma, E.L. *Acta Crystallogr. Sect. A* 1980, **146,** 341

31 Carter, C.W., Jr., Kraut, J., Freer, S.T., Xuong, N., Alden, R.A., and Bartsch, R.G. *J. Biol. Chem.* 1974, **249,** 4212

32 Artymiuk, P.J. and Blake, C.C.F. *J. Mol. Biol.* 1981, **152,** 737

33 Kurachi, K., Sieker, L.C., and Jensen, L.H. *J. Mol. Biol.* 1976, **101,** 11

34 Phillips, S.E.V. *J. Mol. Biol.* 1981, **42,** 531

35 Scouloudi, H. and Baker, E.N. *J. Mol. Biol.* 1978, **126,** 637

36 Tsernoglou, D., Petsko, D.A., and Hudson, R.A. *Mol. Pharmacol.* 1978, **14,** 710

37 Bryant, S.H., Amzel, L.M., Phizackerley, R.P., and Poljak, R.J. *Acta Crystallogr. Sect. B* 1985, **41,** 362

38 Priestle, J.P., Ford, G.C., Glor, M., Mehler, E.L., Smit, J.D.G., Thaller, C., and Jansonius, J.N. *Acta Crystallogr. Sect. A* 1984, **40,** 17

39 Blundell, T.L., Pitts, J.E., Tickle, I.J., Wood, S.P., and Wu, C.-W. *Proc. Natl. Acad. Sci. U.S.A.* 1981, **78,** 4175

40 Arutyunyan, E.G., Terzyan, S.S., Voronova, A.A., Kuranova, I.P., Smirnova, E.A., Vainshtein, B.K., Hoehe, W.E., and Hansen, G. *Dokl. Biochem.* (*Eng. Transl.*) 1981, **258,** 189

41 Epp, O., Lattman, E.E., Schiffer, M., Huber, R., and Palm, W. *Biochemistry* 1975, **14,** 4943

42 Ploegman, J.H., Drent, G., Kalk, K.H., and Hol, W.G.J. *J. Mol. Biol.* 1978, **123,** 557

43 Bolognesi, M., Gatti, G., Menegatti, E., Guareri, M., Marquart, M., Papmokos, E., and Huber, R. *J. Mol. Biol.* 1982, **162,** 839

44 Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., and Wilson, I.A. *Biochem. Biophys. Res. Commun.* 1976, **72,** 146

45 Baker, E.N. and Dodson, E.J. *Acta Crystallogr. Sect. A* 1980, **36,** 559

46 Fujinaga, M., Delbaere, L.T.J., Brayer, G.D., and James, M.N.G. *J. Mol. Biol.* 1985, **184,** 479

47 Suguna, K., Bott, R.R., Padlan, E.A., Subramanian, E., Sheriff, S., Cohen, G.H., and Davies, D.R. *J. Mol. Biol.* 1987, **196,** 877

48 Baker, E.N. *J. Mol. Biol.* 1988, **203,** 1071

49 Kannan, K.K., Ramanadham, M., and Jones, T.A. *Ann. N.Y. Acad. Sci.* 1984, **429,** 49

50 Finzel, B.C., Weber, P.C., Hardman, K.D., and Salemme, F.R. *J. Mol. Biol.* 1985, **186,** 627

51 Higuchi, Y., Kusunoki, M., Matsuura, Y., Yasuoka, N., and Kakudo, M. *J. Mol. Biol.* 1984, **172,** 109

52 Czjzek, M., Payan, F., Guerlesquin, F., Bruschi, M., and Haber, R. 1995 (to be published)

53 Finzel, B.C., Poulos, T.L., and Kraut, J. *J. Biol. Chem.* 1984, **259,** 13027

54 Bolton, W. and Perutz, M.F. *Nature* (*London*) 1970, **228,** 551

55 Cohen, G.H., Davies, D.R., and Silverton, E.W. *J. Mol. Biol.* 1981, **148,** 449

56 Brayer, G.D. and McPherson, A. *J. Mol. Biol.* 1983, **169,** 565

57 Holmes, M.A. and Stenkamp, R.E. *J. Mol. Biol.* 1991, **220,** 723

58 Marquart, M., Deisenhofer, J., Huber, R., and Palm, W. *J. Mol. Biol.* 1985, **141,** 369

59 Bode, W. and Chen, Z. *J. Mol. Biol.* 1983, **164,** 283

60 Hogrefe, H.H., Griffith, J.P., Rossmann, M.G., and Goldberg, E. *J. Biol. Chem.* 1987, **262,** 13155

61 Arutyunyan, E.G., Kuranova, I.P., Vainshtein, B.K., and Steigemann, W. *Kristallografiya* 1980, **25,** 80

62 Honzatko, R.B., Hendrickson, W.A., and Love, W.E. *J. Mol. Biol.* 1985, **184,** 147

63 Weaver, L.H. and Matthews, B.W. *J. Mol. Biol.* 1987, **193,** 189

64 Padlan, E.A., Cohen, G.H., and Davies, D.R. 1995 (to be published)

65 Gribskov, M., Wesson, L., and Eisenberg, D. 1995 (to be published)

66 Blake C.C.F., Geisow, M.J., Oatley, S.J., Rerat, B., and Rerat, C. *J. Mol. Biol.* 1978, **167,** 339

67 Furey, W., Jr., Wang, B.C., Yoo, C.S., and Sax, M. *J. Mol. Biol.* 1983, **167,** 661

68 Drenth, J., Hol, W.G.J., Jansonius, J.N., and Koekoek, R. *Cold Spring Harbor Symp. Quant. Biol.* 1972, **36,** 107

69 James, M.N.G. and Sielecki, A.R. *J. Mol. Biol.* 1983, **163,** 299

70 Zhad, B., Carson, M., Ealick, S.E., and Bugg, C.E. *J. Mol. Biol.* 1992, **227,** 239

71 Legg, M.J. Ph.D. Thesis. Texas Agricultural and Mechanical University, , Texas, 1977

72 Tainer, J.A., Getzoff, E.D., Beem, K.M., and Richardson, J.S. *J. Mol. Biol.* 1982, **160,** 181

73 Mitsui, Y., Satow, Y., Watanabe, Y., and Iitaka, Y. *J. Biochem.* (*Tokyo*) 1980, **88,** 1739

74 Jones, T.A. and Liljas, L. *J. Mol. Biol.* 1984, **177,** 735

75 Matsuura, Y., Kusunoki, M., Harada, W., and Kakudo, M. *J. Biochem.* (*Tokyo*) 1984, **95,** 697

76 Hopper, P., Harrison, S.C., and Sauer, R.T. *J. Mol. Biol.* 1984, **177,** 701

77 Dreusicke, D., Karplus, P.A., and Schulz, G.E. *J. Mol. Biol.* 1988, **199,** 359

78 Mathews, F.S., Argos, P., and Levine, M. *Cold Spring Harbor Symp. Quant. Biol.* 1972, **36,** 387

79 Bhatia, G. Ph.D. Thesis, University of California San Diego, San Diego, California, 1981

80 Hardman, K.D. and Ainsworth, C.F. *Biochemistry* 1972, **11,** 4910

81 Tsukihara, T., Fukuyama, K., Nakamura, M., Katsube, Y., Tanaka, N., Kakudo, M., Wada, K., Hase, T., and Matsubara, H. *J. Biochem.* (*Tokyo*) 1981, **90,** 1763

82 Mercer, W.D., Winn, S.I., and Watson, H.C. *J. Mol. Biol.* 1976, **104,** 277

83 Karplus, P.A. and Schulz, G.E. *J. Mol. Biol.* 1987, **195,** 701

84 Fermi, G., Perutz, M.F., Shannan, B., and Fourme, R. *J. Mol. Biol.* 1984, **175,** 159

85 Szebenyi, D.M.E. and Moffat, K. *J. Biol. Chem.* 1986, **261,** 8761

86 Church, W.B., Guss, J.M., Potter, J.J., and Freeman, H.C. *J. Biol. Chem.* 1986, **261,** 234

87 Bryant, T.N., Shaw, P.J., Walker, N.P., and Wendell, P.L. 1995 (to be published)

88 Winn, S.I., Warwicker, J., and Watson, H.C. 1995 (to be published)

89 Howlin, B., Moss, D.S., and Harris, G.W. *Acta Crystallogr. Sect. A* 1989, **45,** 851

90 Remington, S.J., Woodbury, R.G., and Reynolds, R.A. *Biochemistry* 1988, **27,** 8097

91 Read, R.J., Fujinaga, M., Sielecki, A.R., and James, M.N.G. *Biochemistry* 1983, **22,** 4420

92 Matsuura, Y., Takano, T., and Dickerson, R.E. *J. Mol. Biol.* 1982, **156,** 389

93 Pearl, L. and Blundell, T. *FEBS Lett.* 1984, **174,** 96

94 Wiegand, G., Remington, S., Deisenhofer, J., and Huber, R. *J. Mol. Biol.* 1984, **174,** 205

95 Bolin, J.T., Filman, D.J., Matthews, D.A., Hamlin, R.C., and Kraut, J. *J. Biol. Chem.* 1982, **257,** 13650

96 Smith, W.W., Burnett, R.M., Darling, G.D., and Ludwig, M.L. *J. Mol. Biol.* 1977, **117,** 195

97 Najmudin, S., Nalini, V., Driessen, H.P.C., Slingsby, C., Blundell, T.L., Moss, D.S., and Lindley, P.F. 1995 (to be published)

98 Birktoft, J.J., Rhodes, G., and Banaszak, L.J. *Biochemistry* 1989, **28,** 6065

99 Robbins, A.H., McRee, D.E., Williamson, M., Collett, S.A., Xoung, N.H., Furey, W.F., Wang, B.C., and Stout, C.D. *J. Mol. Biol.* 1991, **221,** 1269

100 Finzel, B.C., Ohlendorf, D.H., Weber, P.C., and Salemme, F.R. *Acta Crystallogr Sect B* 1991, **47,** 588

101 Silva, A.M. and Rossman, M.G. *Acta Crystallogr. Sect. B* 1985, **41,** 147

102 Stevens, R.C., Goulaux, J.E., and Lipscomb, W.N. *Biochemistry* 1990, **29,** 7691

103 Rees, D.C., Lewis, M., and Lipscomb, W.N. *J. Mol. Biol.* 1983, **168,** 367

104 Swain, A.L., Kretsinger, R.H., and Amma, E.L. *J. Biol. Chem.* 1989, **264,** 16620

105 Stout, C.D. 1995 (to be published)

106 Grau, U.M., Trommer, W.E., and Rossmann, M.G. *J. Mol. Biol.* 1981, **151,** 289

107 Wlodawer, A., Walter, J., Huber, R., and Sjolin, L. *J. Mol. Biol.* 1984, **180,** 301

108 Watenpaugh, K.D. 1995 (to be published)

109 Eklund, H., Samama, J.-P., Wallen, L., Branden, C.-I., Akeson, A., and Jones, T.A. *J. Mol. Biol.* 1981, **146,** 561

110 Engh, R., Loebermann, H., Schneider, M., Wiegand, G., and Huber, R. *Protein Eng.* 1989, **2,** 407

111 Fita, I. and Rossmann, M.G. *Proc. Natl. Acad. Sci. U.S.A.* 1985, **82,** 1604

112 Holland, D.R., Tronrud, D.E., Pleyk, H.W., and Flaherty, M. *Biochemistry* 1992, **31,** 11310

113 Molecular Simulations, Inc. *QUANTA 4.0?* 200 Fifth Avenue, Waltham, Massachusetts, 1994

114 Kabsch, W. and Sander, C. *Biopolymers* 1983, **22,** 2577

115 Kabsch, W. and Sander, C. *Proc. Natl. Acad. Sci. U.S.A.* 1984, **81,** 1075

116 Holley, H. and Karplus, M. Neural networks for protein structure prediction. *Methods Enzymol.* 1991, **202,** 204

117 Watthews, B.W. *Biochim. Biophys. Acta* 1975, **405,** 442

118 Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Lautrup, B., Norskov, L., Olsen, O.H., and Petersen, S.B. Protein structure and homology by neural networks. *FEBS Lett.* 1988, **241,** 223

119 Eisenberg, D., Weiss, R.M., Terwilliger, T.C., and Wilcox, W. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* 1982, **17,** 109

120 Eisenberg, D., Weiss, R.M., and Terwilliger, T.C. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proc. Natl. Acad. Sci. U.S.A.* 1984, **81,** 140