ACTA BIOCHIMICA POLONICA
Vol. 39 1992 No. 4

MICHAŁ VIETH[1,2], ANDRZEJ KOLIŃSKI[1,2], JEFFREY SKOLNICK[1] and ANDRZEJ SIKORSKI[2*]

PREDICTION OF PROTEIN SECONDARY STRUCTURE BY NEURAL NETWORKS: ENCODING SHORT AND LONG RANGE PATTERNS OF AMINO ACID PACKING**

[1]Department of Molecular Biology, The Scripps Research Institute, La Jolla, California 92037, U.S.A.
[2]Department of Chemistry, University of Warsaw, L. Pasteura 1; 02 - 093 Warsaw, Poland

Received 28 October, 1992

A complex, cascaded neural network designed to predict the secondary structure of globular proteins has been developed. Information about the local buried-unburied pattern and the average tendency of the particular types of amino acids to be buried inside the globule were used. Nonspecific information about long distance contact maps was also employed. These modifications result in a noticeable improvement (3 - 9%) of prediction accuracy. The best result for the average success ratio for the testing set of nonhomologous proteins was 68.3% (with corresponding Matthews' coefficients, $C_{\alpha,\beta,\mathrm{coil}}$ equal to 0.60, 0.47, 0.43, respectively).

* To whom correspondence should be addressed

**This research was supported in part by grant number GM-37408 from the Division of General Medical Sciences, United States Public Health Service

For a given set of environmental conditions, the amino acid sequence of a globular protein determines its three dimensional structure [1, 2]. The conformation of the polypeptide chain inside these closely packed globules exhibits several characteristic regular motifs: helices, β-strands, well defined turns, etc. It is convenient to classify these secondary structure elements according to the pattern of the hydrogen bonds between the main chain atoms of the polypeptide backbone [3]. Some elements of the secondary structure can be detected even under denaturing conditions. This means that various sequences can exhibit an intrinsic tendency to adopt a specific secondary structure, which is further modulated by tertiary interactions in the folded state [1, 4]. Therefore, the ability to predict secondary structure may be a very important step towards elucidation of the protein folding problem, which is one of the most challenging tasks of contemporary molecular biology [5 - 7].

Various theoretical concepts have been applied to the problem of prediction of the secondary structure from the amino acid sequence [8 - 13], and provided new insights into the protein folding problem. Use of neural networks provided one of the most accurate methods for the prediction of secondary structure [14 - 16]. First, the neural network has to be "trained" by presentation of the amino acid sequences and their corresponding secondary structures. The trained neural network can then be applied to other sequences, giving predictions for their secondary structure. By dividing the secondary structure motifs into three classes: helix, β-sheet and coil (which means all structures other than various helices or β-sheets) Qian & Sejnowski [14] achieved an average success ratio at the level of 64%. The *a priori* classification of the protein as all-α, all-β, and mixed α/β leads to a considerable improvement in prediction accuracy [16]. An additional increase in the success ratio has been obtained by designing a neural network that incorporates the periodicity of α-helices and β-sheets. This way Kneller *et al.* [16] achieved average success ratios of 79% for all α-proteins, 70% for all β-proteins and 64% for α/β mixed proteins, respectively. The neural network predictions of β-turns in globular proteins give an average success ratio of 26% [17] that is more accurate than the results obtained by other methods. Neural networks can also be used to detect protein homology [18] and to predict surface exposure of the amino acids [19].

In this work, we used a neural network model which was designed to encode some additional information besides that of the amino acid sequence alone. We have investigated whether the use of this supplementary information provides any improvement in the prediction accuracy. The following additional information was used: the actual local buried-unburied pattern of residues, the average buried-unburied pattern and specific residue type information about tertiary interactions experienced by the residue for which the secondary structure is predicted. We have found that, on including tertiary interactions, the accuracy of prediction of secondary structure (classified as helix, β-sheet, and coil) was significantly increased.

MATERIALS AND METHODS

a) Model of a simple neural network

For the sake of clarity, it is convenient to describe first a simple neural network model with three layers. As shown in Fig. 1, several such networks can be subsequently linked to form a more complex model. The top branch in Fig.1 presents a scheme of a cascaded neural network for all proteins, the branches at the bottom show cascaded neural networks trained on successive α, β and mixed α/β classes.

The simple computational model of a neural network consists of three layers of units. The first one is the input layer where the amino acid sequence is encoded. This layer consists of 13 input group with 21 units per group. Each cluster encodes one of the amino acids (plus one dummy unit to represent the chain ends). Each cluster contains, in the position which encodes a particular amino acid, one and only one unit designated by a number (other than zero) which is 1 for the simple network without any additional information, and 1 or 2 for the network containing additional information. A dummy residue is assigned to unit number 0. Alanine is the unit number 1, etc. In this way, 13 clusters represent the sequence of 12 amino acids surrounding the central amino acid (which is placed in cluster number 7). The second layer of the neural network, the so called hidden layer, is linked with each unit of the input layer and with each unit of the output layer. In the present work, this layer contains either 2 or 40 units. The output layer contains three units corresponding to the helix, β-sheet or

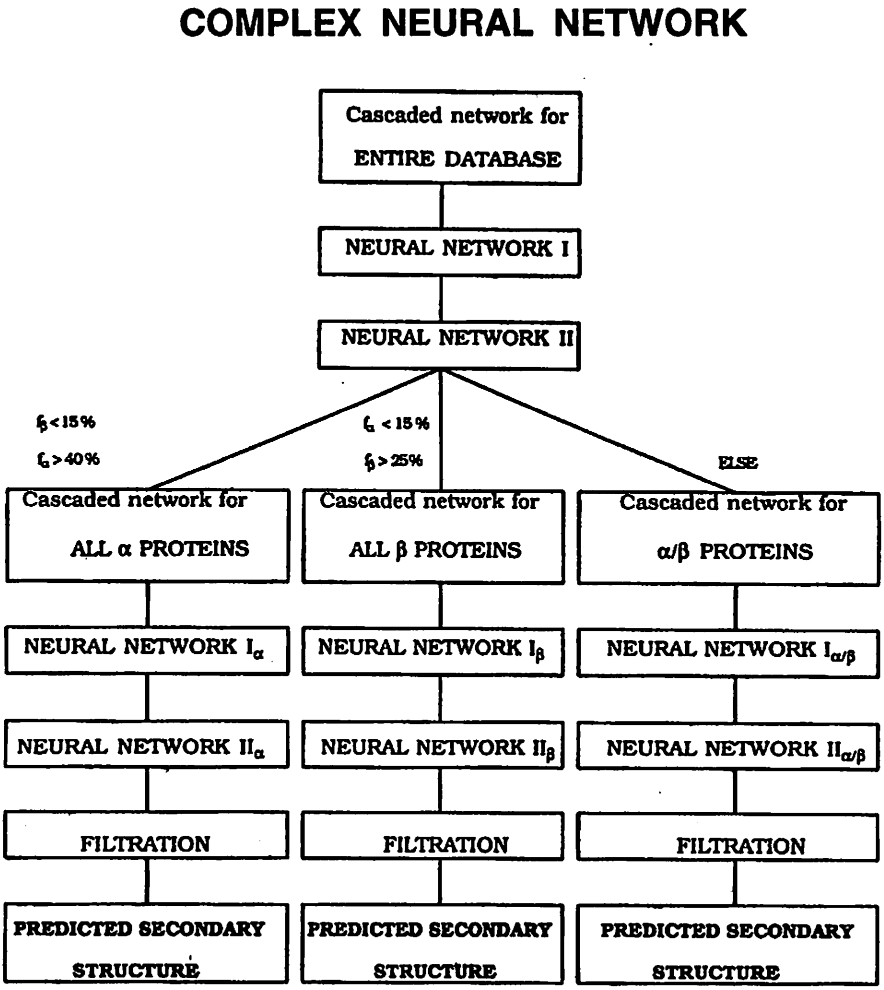
COMPLEX NEURAL NETWORK
Cascaded network for ENTIRE DATABASE
NEURAL NETWORK I
NEURAL NETWORK II
$f_\beta < 15\%$
$f_\alpha > 40\%$
$f_\alpha < 15\%$
$f_\beta > 25\%$
ELSE
Cascaded network for ALL α PROTEINS
Cascaded network for ALL β PROTEINS
Cascaded network for α/β PROTEINS
NEURAL NETWORK I_α
NEURAL NETWORK I_β
NEURAL NETWORK $I_{\alpha/\beta}$
NEURAL NETWORK II_α
NEURAL NETWORK II_β
NEURAL NETWORK $II_{\alpha/\beta}$
FILTRATION
FILTRATION
FILTRATION
PREDICTED SECONDARY STRUCTURE
PREDICTED SECONDARY STRUCTURE
PREDICTED SECONDARY STRUCTURE

Fig. 1. Design of the complex neural network used for secondary structure prediction. Top branch depicts the scheme of a cascaded neural network for all proteins. The three bottom branches depict schemes of cascaded neural networks trained on α, β and mixed α/β proteins

coil state of the central residue of the 13 residue input window and is connected only to the hidden layer.

Information is propagated from the input layer to the hidden layer in the following way:

$$O_k^H = \frac{1}{1+\exp(-O_k^h)} \qquad (1)$$

$$\text{with} \quad O_k^h = \sum_{j=1}^{N} W_{kj} I_j + b_k^h \qquad (2)$$

where:

w_{kj} are the values of the connection weights between the j-th input unit and the k-th hidden unit ($j = 0, 2, ..., 20*13 = 273 = N$, and $k = 1, 2 .... N_{hidden}$), I_j are the values of the input units, O_k^H are the resulting values of the hidden layer units, b_k^h are the values of the biases (see our previous work [20] for further details).

In a similar way, the information from the hidden layer is propagated to the output layer. The output unit with the largest value determines the structural assignment for the central amino acid of the input window.

b) Training procedure for the simple neural network

The training procedure starts from the random distribution of connecting weights (uniformly distributed in the range of – 0.5; 0.5) and small positive values of the biases. The learning rates for both biases and weights were chosen to be 0.07 [14].

The input patterns from the sequences of the training set of proteins are sequentially presented to the network input. Modifications of the connection weights are made according to the delta learning rule [21] based on comparison of the expected secondary structure with that obtained by the network. Further details of the training procedure can be found in the literature [14, 21]. The large number of hidden units and small values of the learning rates assumed in the training procedure prevent weight oscillations. On the other hand, the number of units in the hidden layer cannot be too large, in order to prevent direct memorization of the data. To examine the above possibility, either 2 or 40 hidden units were used. After several hundred iterations over the entire training set (about 300 for 40 hidden units and 800 for 2 hidden units) the values of weights and biases approached a constant level.

c) Database

Appendix A lists 66 proteins (11 049 residues) for which high resolution X-ray structures can be found in the Protein Brookhaven Data Bank [22, 23]. The first 56 proteins (R. Schneider, personal communication) (9 370 residues) were used as the training set, the remaining 10 (1 679 residues) as the small testing set. Appendix B presents another training set containing 137 nonhomologous proteins (28 502 residues) [22]. Twenty nine proteins from Appendix A (5 147 residues) nonhomologous to those from Appendix B were used as the second testing set. Secondary structure assignments for these proteins were generated using the Kabsch & Sander DSSPS program [3]. The various kinds of helices (α, π, 3.10) were grouped together as "helix". Proteins from Appendices A and B were classified as all α, all β and mixed α/β according to the following criterion: a protein was assumed to be in the class when the fraction of the amino acids with β-strand assignment according to the method of Kabsch & Sander [3] was less than 11.3%. All proteins contained less than 12.1% of the amino acids with various helical assignments. All other proteins were assumed to be mixed α/β.

For all proteins, the average number of contacts per residue type was generated from the X-ray positions of all heavy atoms; neighboring residues in the sequence were omitted. Two residues were considered to be in contact (often at multiple points) when the distance between any two heavy atoms belonging to these residues was smaller than 1.1 times the sum of their Van der Waals radii. For C_α and C atoms of the side groups, the Van der Waals radius was assumed to be 2 Å, in accordance with the size of the methyl group, the $-CH_2-$ or the $=CH-$ groups in hydrocarbons. Because of the limited accuracy of X-ray structures, this somewhat arbitrary choice seems to be acceptable. The mean number of long range contacts for each amino acid used for the training set (137 proteins) are compiled in Table 1.

d) Neural network with buried-unburied pattern

Based on the average number of contacts per residue obtained from the set of 137 proteins, the buried-unburied pattern for all these proteins was generated as follows: For each type of amino acid, the distribution of the

Table 1

Average distribution of contact points for each amino acid of the training set of 137 proteins listed in Appendix B

	CODE	N_{cont} [1]	N_{am} [2]	C_{av} [3]	C_{pr} [4]	% [5]	B-U_{av} [6]
1	ALA	8004	2447	3.3	3	63	B
2	ARG	6125	1125	5.4	3	42	U
3	ASN	4926	1280	3.8	2	43	U
4	ASP	6529	1663	3.9	2	43	U
5	CYS	3163	536	5.9	8	77	B
6	GLN	4588	1034	4.4	3	54	U
7	GLU	6552	1615	4.1	2	44	U
8	GLY	5879	2352	2.5	0	0	U
9	HIS	4086	635	6.4	6	60	B
10	ILE	8654	1500	5.8	4	59	B
11	LEU	12887	2365	5.4	5	61	B
12	LYS	7372	1730	4.3	2	42	U
13	MET	3235	537	6.0	6	64	B
14	PHE	10627	1152	9.2	9	59	B
15	PRO	4950	1391	3.6	3	49	U
16	SER	5990	1906	3.1	2	53	U
17	THR	6607	1800	3.7	2	46	U
18	TRP	5168	413	12.5	14	66	B
19	TYR	9526	1001	9.5	9	57	B
20	VAL	9547	2039	4.7	4	59	B

[1] N_{cont} indicates total number of contact points for all residues of this type,
[2] N_{am} – total number of residues of this type,
[3] C_{av} – average number of contact points for a given type of residue,
[4] C_{pr} – the most probable number of contact points,
[5] % – percentage of cases in which the given type of residue was considered to be in the buried state,
[6] B-U_{av} – average buried/unburied pattern: B - buried on average, U - unburied on average.

number of contact points was obtained and histograms showing this distribution were plotted. For most of the amino acids, these plots show single well defined maxima. An example of such a histogram for glutamine is presented in Fig. 2. For a given residue, when the number of contact points

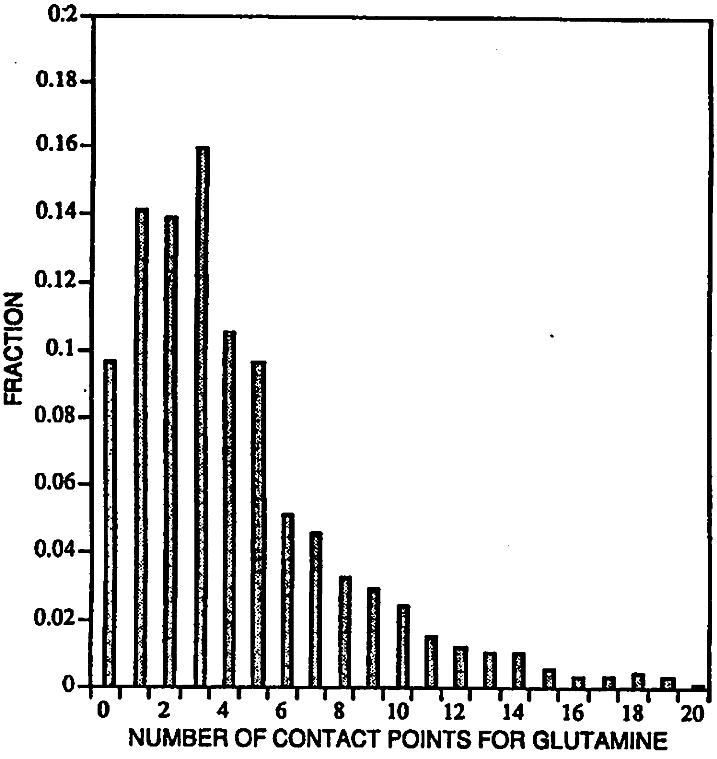

Fig. 2. Distribution of contact points for glutamine

exceeds a specified threshold, that residue is considered to be buried. It should be noted that this pattern is correlated with the surface exposure parameters described in the literature at the level of 67% (the Matthews' coefficient equals to 0.44) [24]. It is also possible to predict this pattern from the sequence of the amino acids [19] at the level of 72% (the Matthews' coefficient is 0.52). For testing the prediction accuracy, we used either a pattern obtained from the real contacts map or that predicted for the sequence by the neural network method.

There are many ways of supplying additional information to the network [14, 20]. In order to represent the buried-unburied pattern, instead of setting the units as equal to 1 in the 13 cluster amino acid window, we set each unit as equal to 1 when the residue was unburied and as 2 when it was buried (Fig. 3). In this way, in addition to the sequence itself, the local buried-unburied pattern was propagated through the network. A schematic representation is shown in Fig. 3.

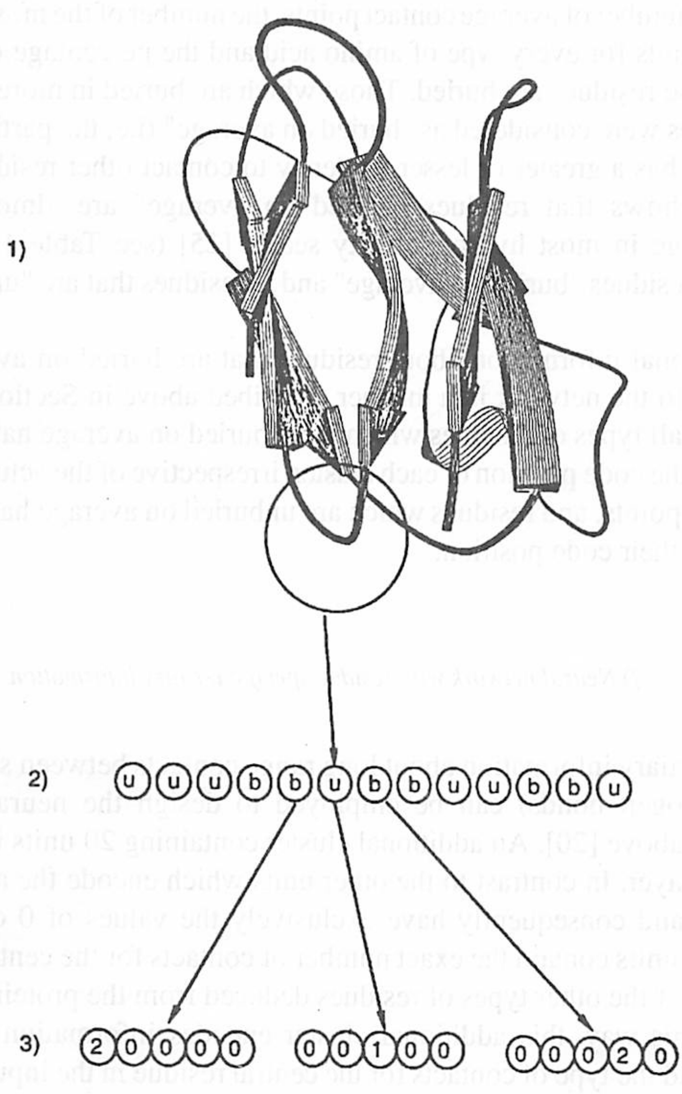

Fig. 3. Method of presenting the buried-unburied pattern to the network. 1) Polypeptide chain of plastocyanin (Priestle, J. P., personal communication) with the 13-residue window circled. 2) Buried-unburied pattern for these 13 residues. Some of the residues are buried (b), some are unburied (u). 3) Coding this pattern by numbers 2 and 1, respectively, in the input layer

e) Neural network with burial status statistically defined

The 20 types of amino acids were classified as buried or unburied "on average" based on the percentage of cases in which they were buried. Table

1 lists the number of average contact points, the number of the most probable contact points for every type of amino acid and the percentage of cases in which these residues are buried. Those which are buried in more than 55% of the cases were considered as "buried on average" (i.e, the particular type of residue has a greater or lesser tendency to contact other residues). This division shows that residues "buried on average" are almost always hydrophobic in most hydrophobicity scales [25] (see Table 1, where B indicates residues "buried on average" and U residues that are "unburied on average").

Additional information about residues that are buried on average was presented to the network in a manner described above in Section *d*. However, here all types of residues which were buried on average have the unit set as 2 in the code position of each cluster, irrespective of the actual number of contact points, and residues which are unburied on average have the unit set as 1 in their code position.

f) Neural network with residue specific tertiary information

The tertiary information about long range contacts between side groups (and hydrogen bonds) can be employed to design the neural network described above [20]. An additional cluster containing 20 units is added to the input layer. In contrast to the other units which encode the amino acid sequence and consequently have exclusively the values of 0 or 1, these additional units contain the exact number of contacts for the central residue with each of the other types of residues deduced from the protein's contact map. In this way, this additional cluster encodes information about the number and the type of contacts for the central residue in the input window.

RESULTS AND DISCUSSION

In this section, the results of supplying various kinds of additional information to the network are presented. Modifications of the network design which lead to the successive improvements of the prediction accuracy are also discussed. The standard measures of neural network performance were applied. The overall success ratio is given as:

$$Q3 = \frac{1}{N} (p\alpha + p\beta = p_{coil}) \tag{3}$$

where $p\alpha$, $p\beta$ and p_{coil} are the number of helix, β-sheet and coil assignment, predicted correctly for the total number of residues N. Matthews' coefficients [26] $C\alpha$, $C\beta$ and C_{coil} provide a more adequate measure of performance of the algorithm, where

$$C\alpha = \frac{(p\alpha n\alpha - u\alpha o\alpha)}{\sqrt{(n\alpha + o\alpha)\,(n\alpha + u\alpha)\,(p\alpha + u\alpha)\,(p\alpha + o\alpha)}} \tag{4}$$

where $p_{j\alpha}$, $n_{j\alpha}$, u_j and $o_{j\alpha}$ are the numbers of $j = \alpha$, β or coil assignment properly predicted, properly rejected as non j type, underpredicted (not predicted) and overpredicted (predicted in non j type positions), respectively. These measures allow comparison between various methods used in this work and with other works. However, since different databases have been employed, the comparison is only qualitative.

a) Cascaded neural networks and filtration of the data

It was shown by Qian & Sejnowski [14] that a cascaded neural network improves the performance of neural network models. The simple neural network often produces an artificial sequence of structural assignments, for example, a separated helical or β-strand assignment. These artifacts can be partially removed through the use of a second neural network with the values of output units from the first network used as the input data to the second network [14]. Thus, the input layer of the second network consists of a window of 13 clusters, each containing 3 units with values corresponding to the output values from the first network (helix, β-sheet, coil). This network is separately trained on the same training set of proteins. Briefly speaking, the second network uses the secondary structure predicted by the first network as the input and gives improved secondary structure assignments as the output. The testing procedure uses a cascade of both networks.

The above procedure, while considerably improving the network's performance, still leaves some artificial sequences of structural assignments. Therefore, we introduced a filter which replaces some sequences of structural assignments by other, presumably more physical, ones. The complete procedure is schematically depicted in Table 2. Figure 4 shows an

Table 2

Filtration of the predicted secondary structure assignments

Predicted assignment	Accepted assignment
HH+H	HHHH
H+HH	HHHH
–E–	–N–
+H+	+N+
+HH+	+??+
–EE–	–??–
Residues at the beginning and at the end of the chain are always coil	

+, means an assignment other than helical

–, means an assignment other than β-sheet

?, means assignment generated in random fashion (53% chance for coil, 28% chance for H (helix), 19% chance for E (β-sheet))

N, means coil assignment

example of the effect of application of the second cascaded network, and the filtration of predicted secondary structure assignments.

b) Complex neural network and prediction of protein classes

The accuracy of prediction of secondary structures of globular proteins significantly increases when the prediction is performed within a given class of protein (α, β or mixed α/β). Therefore, it seems appropriate to design a neural network to make the classification assignment. The entire complex network is constructed according to the idea given in Figure 1.

The first two networks which work as a cascade described in Section *a* are used only for the assignment of the protein into a structural class. These networks are trained on the whole training set of 137 proteins. If a given protein is predicted to contain less than 15% β-sheet and more than 40% helix, it is considered to be an all α-protein. When the prediction gives less than 15% helix and more than 25% β-sheet, the protein is considered as all β. The remaining proteins are considered to be α/β. The applicability of this method for the prediction of structural classes for the large testing set was also examined, giving a 79% accuracy. It should be noted that in the

```
i)    MKIVYWSGTGNTEKMAELIAKGIIESGKDVNTINVSDVNIDELLNEDILILGCSAMGDEVLEESEFEPF

ii)   _EEEEE____HHHHHHHHHHHHHHH_____EEEE_____________EEEEEE____________HHHH

iii)  _EEEEEE____HHHHHEHHHHH_EE____HEHEEE__E___H_HHHH_EE___HEHH_HHHE___H__H

iv)   _EEEEEE___HHHHHHHHHHHHHHH_____EEEEE______HHHHHHHHE___HHHHHHHH_______H

v)    _EEEEEE___HHHHHHHHHHHHHHH_____EEEEE______HHHHHHHHH___HHHHHHHH_______H


i)    IEEISTKISGKKVALFGSYGWGDGKWMRDFEERMNGYGCVVVETPLIVQNEPDEAEQDCIEFGKKIANI

ii)   HHHH_______EEEEEEEE_____HHHHHHHHHHHH___EE____EEEE___HHHHHHHHHHHHHHH__

iii)  HEHHHH_H__HEEEEE_H_______EHHEHHE____H__EEE___HEH_______E_E_HHH__HHEEE

iv)   HHHHHHHH___EEEEEE________EHHHHH_______EEEE___HH__________EHHHHHHHH_EE

v)    HHHHHHHH___EEEEEE_________HHHHH_______EEEE________________HHHHHHHH___
```

Fig. 4. Prediction of secondary structure of flavodoxin 1fdx (138 residues) by the neural network. The percentages of correct predictions are shown below in brackets. i) Amino acid sequence in one-letter code; ii) real structure; iii) structure predicted by standard neural network (55%); iv) structure predicted by cascaded neural network (62%); v) structure predicted by cascaded neural network with filtration (63%). H, indicates helical assignment of a given amino acid, E, –β-sheet assignment and "_", – coil assignment

remaining 21% of the cases, one α and four β class proteins were classified as mixed α/β and there was no case in which the network classified a β protein as all α or *vice versa*. This means that, in the worst case, the prediction accuracy of a complex network will be on the same level as for the simple neural network.

The second part of the complex neural network is very similar; however, each branch of the cascade is trained on the subsets of the training set (α, β, mixed classes of proteins). The results for these cases are compared in Table 3. It can be seen that the use of the complex network improves the prediction accuracy by 3% on average, giving better predictions especially for helices as compared to the cascaded network for all proteins.

c) Effect of the number of hidden units

For all the tested cases the effect of using either 2 or 40 hidden units was examined. Better results were obtained for 2 hidden units. This can be rationalized by the argument that using 2 hidden units (less elements) one obtains a more general local sequence – secondary structure dependence. The poorer results of the network with 40 hidden units indicate that the network tends to memorize some sequential patterns present in the training set.

d) Effect of additional information

The division of proteins into structural classes improves the prediction accuracy by 2 - 4% for the case of 2 hidden units. For 40 hidden units, this effect is much smaller. The use of the real or predicted buried-unburied pattern improves the prediction accuracy of a complex network by 1% on average. Both the real buried-unburied pattern and the division of the proteins into structural classes gives a 4 - 5% increase in prediction accuracy compared to the network with only sequence information [14, 16]. The best result, 68.3%, was obtained from the complex network with two hidden units and the predicted buried-unburied pattern. This is by 4.5% better than the value of 63.8% obtained from the complex network using no additional information and by 8% better than with the cascaded network for all proteins proposed by Qian & Sejnowski [14] for these particular training and testing sets. In general, these effects allow for better prediction of helices. Using 40 hidden units, the success ratio for the complex network with additional information was no better than for the simple complex network. This can be rationalized by the conjuncture that information about the buried-unburied pattern is partially coded by the sequence and does not provide any important additional information to the network.

e) Complex neural network with residue specific tertiary interactions

Suppose that for a protein from the testing set some information about the number and type of contact points for the central residue is given. The network is of a design similar to that shown in Fig. 1, however, an additional (20 units) input cluster as described in Materials and Methods Section *d* is used. As shown in Table 3, this network gives the best prediction of 72.6%. The values of the Matthews' coefficients (greater than 0.5) indicate very high correlation to real structures. We also examined the prediction accuracy of a network with information about specific residue type contact maps for all 13 residues in the amino acid window (20*13 additional units), with about the same average success ratio. This indicates that either the accuracy of the neural network method is limited to 72% for secondary structure prediction or there is some other, more important physical information for secondary structure formation that is not encoded in the model.

Finally, let us note that the prediction accuracy for secondary structure of homologous proteins by the above described neural network method approached 90 - 100%, when the network was trained on the set of homologous proteins, or, at least, on sets containing a substantial fraction of homologous sequences.

This work demonstrates that neural network can be successfully used to divide proteins into structural classes, and that a complex neural network which gives a prediction accuracy for secondary structure at the level of 68% can be developed. The overall effect of all modifications used in tis work is at the level of a 4% improvement. Surprisingly, no kind of local buried/unburied pattern information did improve noticeably the success of secondary structure prediction. Another finding is that inclusion of residue specific tertiary information leads to a greater improvement in the accuracy of secondary structure prediction. This makes sense, from the physical point of view, since it is rather unlikely that local sequence of the amino

Table 3

Comparison of the performance of various neural networks

For details see text

TYPE OF THE INPUT INFORMATION	TYPE OF CASCADED NETWORK	PREDICTION ACCURACY							
		2 HIDDEN UNITS				40 HIDDEN UNITS			
		Q_3*	C_α**	C_β	C_{coil}	Q_3*	C_α	C_β	C_{coil}
[1]SEQUENCE + "BURIED ON AVERAGE" PATTERN	ALL[2]	60.9	0.43	0.36	0.41	64.3	0.49	0.36	0.43
	α [3]	76.4	0.50	0.0	0.48	75.5	0.46	-.01	0.46
	β [4]	73.4	0.0	0.53	0.49	70.5	0.0	0.45	0.44
	MIXED [5]	62.2	0.45	0.35	0.41	61.1	0.48	0.36	0.36
	COMPLEX[6]	68.0	0.57	0.45	0.45	67.7	0.56	0.41	0.47
[7]SEQUENCE ONLY	ALL [2]	59.0	0.44	0.37	0.42	63.0	0.50	0.33	0.38
	α [3]	72.6	0.42	0.0	0.42	75.9	0.48	0.21	0.46
	β [4]	74.0	0.0	0.54	0.48	75.7	0.0	0.59	0.51
	MIXED [5]	62.3	0.43	0.36	0.41	59.5	0.42	0.27	0.38
	COMPLEX[6]	64.4	0.53	0.42	0.37	64.9	0.51	0.46	0.39
SEQUENCE AND REAL "B-U" PATTERN[8]	ALL [2]	63.5	0.46	0.42	0.42	63.0	0.52	0.32	0.37
	α [3]	73.4	0.38	0.0	0.36	72.2	0.44	0.0	0.40
	β [4]	73.4	0.0	0.51	0.50	72.8	0.0	0.51	0.44
	MIXED [5]	60.1	0.45	0.29	0.30	58.8	0.41	0.29	0.36
	COMPLEX [6]	68.3	0.57	0.50	0.43	67.0	0.59	0.40	0.41
SEQUENCE AND B-U PATTERN PREDICTED BY A NETWORK[9]	ALL [2]	61.0	0.46	0.42	0.42	59.7	0.43	0.28	0.37
	α [3]	77.0	0.48	0.0	0.46	77.3	0.53	0.0	0.50
	β [4]	74.0	0.0	0.53	0.52	75.1	0.0	0.52	0.53
	MIXED [5]	62.2	0.50	0.34	0.37	57.8	0.41	0.27	0.33
	COMPLEX [6]	68.3	0.60	0.47	0.43	64.7	0.53	0.39	0.39
SEQUENCE AND REAL RESIDUE SPECIFIC CONTACT MAP[10]	α [3]	81.6	0.6	0.0	0.56	80.8	0.59	0.0	0.54
	β [4]	70.5	0.0	0.49	0.4	74.0	0.0	0.55	0.49
	MIXED [5]	61.5	0.39	0.39	0.41	64.5	0.53	0.38	0.43
	COMPLEX [6]	68.7	0.56	0.49	0.46	72.6	0.66	0.52	0.51

[1] Results for the network with input information about residues "buried on average".

[2] Cascaded network trained and tested on all types of proteins with filtration.

[3] Cascaded network trained and tested on α proteins with filtration.

[4] Cascaded network trained and tested on β proteins with filtration .

[5] Cascaded network trained and tested on mixed α/β proteins with filtration.

[6] Complex neural network with internal division of proteins into classes.

[7] Results for the network with only sequence information but with protein classes. assigned from the known structures.

acids alone determines the secondary structure of the native conformation. While there are some (sometimes rather well defined) intrinsic tendencies to form secondary motifs, the actual secondary structure is also dictated by packing and long range interactions within the globular state [1, 2, 4, 27]. However, the fact that information about residue specific contacts doesn't on average give an accuracy for secondary structure prediction greater than 72% may indicate that this is the upper precision limit of this method. Thus, alternative schemes may be required to improve the level accuracy of secondary structure prediction. One such scheme that uses a supersecondary structure predictor is currently under development.

We wish to thank Drs. Adam Godzik and Mariusz Milik for stimulating discussions.

Appendix A

List of proteins used as the smaller training set and testing set. The last 10 proteins were used as the small testing set. Proteins marked with [a] were used as the large testing set:
1ALC[a1] 1BP2[2] 1CCR[a3] 1ECO[4] 1FD2[5] 1HMQ[6] 1L01[7] 1MBD[8]
1R69[9] 1UTG[10] 2CDV[a11] 2CPP[12] 2CYP[13] 2LH2[14] 2LHB[a15] 2WRP[16]
3C2C[a17] 5TNC[18] 1HNE[a19] 1NXB[20] 2CNA[a21] 2FB4[a22] 2PKA[23]
2SOD[24] 3RP2[a25] 5I1B[26] 2GCR[27] 2SGA[a28] 2RHE[a29] 1CSE[a30] 1CTF[31]
1FDX[32]1FX1[33] 1GD1[34] 1RDG[35] 1UBQ[36] 2ACT[37] 3APP[a38] 2AZA[39]
3B5C[a40]2CA2[41] 2CI2[a42] 2OVO[43] 2PRK[a44] 3EST[a45] 3RNT[46] 3TLN[47]
5CPA[48]5RXN[49] 6LDH[50] 7RSA[a51] 8DFR[a52] 9PAP[53] 9WGA[54] 1PAZ[55]
3APR[56]3GRS[a57] 4FXN[a58] 1SNS[a59] 1PCY[a60] 1SGT[a61] 1HOE[62a]
4HHB[a63] 2CCY[a64]1MBA[a65] 1LZ1[a66]

References to Appendix A

1. K.R.Acharya, D.I.Stuart, N.P.C.Walker,M.Lewis, D.C.Phillips (1989), *J.Mol.Biol.*, **208**, 99.
2. B.W.Dijkstra, K.H.Kalk, W.G.J.Hol, J.Drenth (1981), *J.Mol.Biol.*, 147, 97.
3. H.Ochi, Y.Hata, N.Tanaka, M.Kakudo, T.Sakurai, S.Aihara, Y.Morita (1983), *J.Mol.Biol.*, **166**, 407.

Legend to Table 3 cont.

[8] Results for the network with real local buried-unburied pattern.
[9] Results for the network with the buried-unburied pattern predicted by the network.
[10] Results for the network with information about type and number of contact points for the central residue.
*Percentage of correct prediction.
** Matthews' coefficients.

4. W.Steigemann, E.Weber (1979), *J.Mol.Biol.*, **127**, 309.

5. A.E.Martin, B.K.Burgess, C.D.Stout, V.L.Cash, D.R.Dean (1990), *Proc.Natl.Acad.Sci.USA*, **87**, 598.

6. R.E.Stenkamp, L.C.Sieker, L.H.Jensen (1983), *Acta Crystallogr., Sect.B*, **39**, 697.

7. M.G.Gruetter, T.M.Gray, L.H.Weaver, T.Alber, K.Wilson, B.W.Matthews (1987), *J.Mol.Biol.*, **197**, 315.

8. S.E.V.Phillips, B.P.Schoenborn (1981), *Nature*, **292**, 81.

9. A.Mondragon, S.Subbiah, S.C.Almo, M.Drottar, S.C.Harrison (1989), *J.Mol.Biol.*, **205**, 189.

10. I.Morize, E.Surcouf, M.C.Vaney, Y.Epelboin, M.Buehner, F.Fridlansky, E.Milgrom, J.P.Mornon (1987), *J.Mol.Biol.*, **194**, 725.

11. Y.Higuchi, M.Kusunoki, Y.Matsuura, N.Yasuoka, M.Kakudo (1984), *J.Mol.Biol.*, **172**, 109.

12. T.L.Poulos, B.C.Finzel, A.J.Howard (1987), *J.Mol.Biol.*, **195**, 687.

13. B.C.Finzel, T.L.Poulos, J.Kraut (1984), *J.Biol.Chem.*, **259**, 13027.

14. E.G.Arutyunyan, I.P.Kuranova, B.K.Vainshtein, W.Steigemann (1980), *Kristallografiya*, **25**, 80.

15. R.B.Honzatko, W.A.Hendrickson, W.E.Love (1985), *J.Mol.Biol.*, **184**, 147.

16. C.L.Lawson, R.-G.Zhang, R.W.Schevitz, Z.Otwinowski, A.Joachimiak, P.B.Sigler, to be published.

17. G.E.Bhatia, Thesis, University of California, San Diego.

18. O.Herzberg, M.N.G.James (1988), *J.Mol.Biol.*, **203**, 761.

19. M.A.Navia, B.M.McKeever, J.P.Springer, T.-Y.Lin, H.R.Williams, E.M.Fluder, C.P.Dorn, K.Hoogsteen (1989), *Proc.Natl.Acad.Sci.USA*, **86**, 7.

20. D.Tsernoglou, G.A.Petsko, R.A.Hudson (1978), *Mol.Pharmacol.*, **14**, 710.

21. G.N.Reeke, Jr., J.W.Becker, G.M.Edelman (1975), *J.Biol.Chem.*, **250**, 1525.

22. H.D.Kratzin, W.Palm, M.Stangel, W.E.Schmidt, J.Friedrich, N.Hilschmann (1989), *Biol. Chem.Hoppe-Seyler*, **370**, 263.

23. W.Bode, Z.Chen, K.Bartels, C.Kutzbach, G.Schmidt-Kastner, H.Bartunik (1983), *J.Mol.Biol.*, **164**, 237.

24. J.A.Tainer, E.D.Getzoff, K.M.Beem, J.S.Richardson, D.C.Richardson (1982), *J.Mol.Biol.*, **160**, 181.

25. S.J.Remington, R.G.Woodbury, R.A.Reynolds, B.W.Matthews, H.Neurath (1988), *Biochemistry*, **27**, 8097.

26. D.H.Ohlendorf, A.C.Treharne, P.C.Weber, J.J.Wendoloski, F.R.Salemme, M.Lischwe, R.C.Newton, to be published.

27. H.E.White, H.P.C.Driessen, C.Slingsby, D.S.Moss, P.F.Lindley (1989), *J.Mol.Biol.*, **207**, 217.

28. J.Moult, F.Sussman, M.N.G.James (1985), *J.Mol.Biol.*, **182**, 555.

29. W.Furey, Jr., B.C.Wang, C.S.Yoo, M.Sax (1983), *J.Mol.Biol.*, **167**, 661.

30. C.A.McPhalen, M.N.G.James (1988), *Biochemistry*, **27**, 6582.

31. M.Leijonmarck, A.Liljas (1987), *J.Mol.Biol.*, **195**, 555.

32. E.T.Adman, L.C.Sieker, L.H.Jensen (1976), *J.Biol.Chem.*, **251**, 3801.

33. K.D.Watenpaugh, L.C.Sieker, L.H.Jensen, in: *Flavins and Flavoproteins* (1976), T.P.Singer (ed.), Elsevier Scientific Publ. Co., Amsterdam.

34. T.Skarzynski, P.C.E.Moody, A.J.Wonacott (1987), *J.Mol.Biol.*, **193**, 171.

35. M.Frey, L.Sieker, F.Payan, R.Haser, M.Bruschi, G.Pepe, J.Le Gall (1987), *J.Mol.Biol.*, **197**, 525.

36. S.Vijay-Kumar, C.E.Bugg, W.J.Cook (1987), *J.Mol.Biol.*, **194**, 531.

37. E.N.Baker, E.J.Dodson (1980), *Acta Crystallogr., Sect.A*, **36**, 559.

38. M.N.G.James, A.R.Sielecki (1983), *J.Mol.Biol.*, **163**, 299.

39. E.N.Baker (1988), *J.Mol.Biol.*, **203**, 1071.

40. F.S.Mathews, P.Argos, M.Levine (1972), *Cold Spring Harbor Symp.*, **36**, 387.

41. A.E.Eriksson, P.M.Kylsten, T.A.Jones, A.Liljas (1988), *Proteins.Struct., Funct.*, **4**, 283.

42. C.A.McPphalen, M.N.G.James (1987), *Biochemistry*, **26**, 261.

43. W.Bode, O.Epp, R.Huber, M.Laskowski, Jr., W.Ardelt (1985), *Eur.J.Biochem.*, **147**, 387.

44. C.Betzel, G.P.Pal, W.Saenger (1988), *Acta Crystallogr., Sect.B*, **44**, 163.
45. E.Meyer, G.Cole, R.Radhakrisbnan, O.Epp (1988), *Acta Crystallogr., Sect.B*, **44**, 26.
46. D.Kostrewa, H.-W.Choe, U.Heinemann, W.Saenger (1989), *Biochemistry*, **28**, 7592.
47. M.A.Holmes, B.W.Matthews (1982), *J.Mol.Biol.*, **160**, 623.
48. D.C.Rees, M.Lewis, W.N.Lipscomb (1983), *J.Mol.Biol.*, **168**, 367.
49. K.D.Watenpaugh, to be published.
50. C.Abad-Zapatero, J.P.Griffith, J.L.Sussman, M.G.Rossmann (1987), *J.Mol.Biol.*, **198**, 445.
51. A.Wlodawer, L.A.Svensson, L.Sjolin, G.L.Gilliland (1988), *Biochemistry*, **27**, 2705.
52. J.F.Davies, D.A.Matthews, S.J.Oatley, B.T.Kaufman, N.-H.Xuong, J.Kraut, to be published.
53. I.G.Kamphuis, K.H.Kalk, M.B.A.Swarte, J.Drenth (1984), *J.Mol.Biol.*, **179**, 233.
54. C.S.Wright (1990), *J.Mol.Biol.*, **215**, 635.
55. K.Petratos, Z.Dauter, K.S.Wilson, to be published.
56. K.Suguna, E.A.Padlan, C.W.Smith, W.D.Carlson, D.R.Davies (1987), *Proc.Natl.Acad.Sci. USA*, **84**, 7009.
57. P.A.Karplus, G.E.Schulz (1987), *J.Mol.Biol.*, **195**, 701.
58. W.W.Smith, R.M.Burnett, G.D.Darling, M.L.Ludwig (1977), *J.Mol.Biol.*, **117**, 195.
59. M.J.Legg, Thesis, Texas Agricultural and Mechanical University.
60. J.M.Guss, H.C.Freeman (1983), *J.Mol.Biol.*, **169**, 521.
61. R.J.Read, M.N.G.James (1988), *J.Mol.Biol.*, **200**, 523.
62. J.W.Pflugrath, G.Wiegand, R.Huber, L.Vertesy (1986), *J.Mol.Biol.*, **189**, 383.
63. G.Fermi, M.F.Perutz, B.Shaanan, R.Fourme (1984), *J.Mol.Biol.*, **175**, 159.
64. B.C.Finzel, P.C.Weber, K.D.Hardman, F.R.Salemme (1985), *J.Mol.Biol.*, **186**, 627.
65. M.Bolognesi, S.Onesti, G.Gatti, A.Coda, P.Ascenzi, M.Brunori (1989), *J.Mol.Biol.*, **205**, 529.
66. P.J.Artymiuk, C.C.F.Blake (1981), *J.Mol.Biol.*, **152**, 737.

Appendix B

List of 137 proteins used as the second (large) training set and testing set

1Prc[1] 1abp[2] 1cc5[3] 1coh[4] 1crn[5] 1csc[6] 1cy3[7] 1eca[8] 1fc2[9] 3fxb[10] 1gcn[11] 1hip[12] 1hmz[13] 1l12[14] 1lh1[15] 11[16] pRc[17] 1ppt[18] 1prc[19] 1pyp[20] 1rhd[21] 1rns[22] 1wrp[23] 2aat[24] 2atc2[5] 2bp2[26] 2cpp[27] 2cyp[28] 2ins[29] 2mlt[30] 2or1[31] 2tmv[32] 2utg[33] 3gap[34] 3pgk[35] 3pgm[36] 4ins[37] 4ts1[38] 5cpv[39] 5ldh[40] 5mb[n41] 5tnc[42] 5xia[43] 9wga[44] 1F19[45] 1acx[46] 1bds[47] 1bmV[48] 1bmv[49] 1cbh[50] 1cd4[51] 1f19[52] 1fc1[53] 1gcr[54] 1p09[55] 1rmu[56] 1tnf[57] 2gn5[58] 2h1A[59] 2il1b[60] 2kai[61] 2ltN[62] 2ltn[63] 2mEv[64] 2meV[65] 2mev[66] 2mrt[67] 2pab[68] 2oka[69] 2r06[70] 2rsp[71] 2sod[72] 2st[v73] 2tbv[74] 3hmg[75] 4cpa[76] 4er4[77] 4sgb[78] 5ebx[79] 6hir[80] 8a[pI81] 1cho[82] 1cla[83] 1cms[84] 1ctf[85] 1etu[86] 1fcb[87] 1fx[88] 11gp18[9] 1il8[90] 1paz[91] 1pfk[92] 1phh[93] 1prC[94] 1r08[95] 1rbb[96] 1rnt[97] 1sn3[98] 1tec[99] 1tgs[100] 1ubq[101] 1wsY[102] 1wsy[103] 2act[104] 2at1[105] 2aza[106] 2dhf[107] 2fd2[108] 2gbp[109] 2gd1[110] 2gls[111] 2hla[112] 2liv[113] 2sbt[114] 2ssi[115] 2taa[116] 2tpi[117] 2ypi[118] 3adk[119] 3b5c[120] 3bcl[121] 3b1m[122] 3ca2[123] 3dfr[124] 3fxc[125] 3hmG[126] 3icd[127] 4dfr[128] 4mdh[129] 4sbv[130] 4tmn[131] 5cpa[132] 5rxn[133] 6acn[134] 7cat[135] 8adh[136] 8api[137]

References to Appendix B

1. J.Deisenhofer, O.Epp, I.Sinning, H.Michel, to be published.
2. G.L.Gilliland, F.A.Quiocho (1981), *J.Mol.Biol.*, **146**, 341.
3. D.C.Carter, K.A.Melis, S.E.O'Donnell, B.K.Burgess, F.Furey Jr., B.-C. Wang, C.D.Stout (1985), *J.Mol.Biol.*, **184**, 279.
4. B.Luisi, N.Shibayama (1989), *J.Mol.Biol.*, **206**, 723.
5. M.M.Teeter (1984), *Proc.Natl.Acad.Sci.USA*, **81**, 6014.
6. M.Karpusas, D.Holland, S.J.Remington, to be published.
7. M.Pierrot, R.Haser, M.Frey, F.Payan, J.-P.Astier (1982), *J.Biol.Chem.*, **257**, 14341.
8. W.Steigemann, E.Weber (1979), *J.Mol.Biol.*, **127**, 309.
9. J.Deisenhofer (1981), *Biochemistry*, **20**, 2361.
10. K.Fukuyama, H.Matsubara, T.Tsukihara, Y.Katsube (1989), *J.Mol.Biol.*, **210**, 383.
11. K.Sasaki, S.Dockerill, D.A.Adamiak, I.J.Tickle, T.Blundell (1975), *Nature*, **257**, 751.
12. C.W.Carter, Jr., J.Kraut, S.T.Freer, N.-H.Xuong, R.A.Alden, R.G.Bartsch (1974), *J.Biol.Chem.*, **249**, 4212.
13. R.E.Stenkamp, L.C.Sieker, L.H.Jensen (1983), *Acta Crystallogr., Sect. B*, **39**, 697.
14. T.Alber, S.Dao-Pin, K.Wilson, J.A.Wozniak, S.P.Cook, B.W.Matthews (1987), *Nature*, **330**, 41.
15. E.G.Arutyunyan, I.P.Kuranova, B.K.Vainshtein, W.Steigemann (1980), *Kristallografiya*, **25**, 80.
16. S.R.Jordan, C.O.Pabo (1988), *Science*, **242**, 893.
17. see ref.1.
18. T.L.Blundell, J.E.Pitts, I.J.Tickle, S.P.Wood, C.-W.Wu (1981), *Proc.Natl.Acad.Sci.USA*, **78**, 4175.
19. see ref 1.
20. E.G.Arutiunian, S.S.Terzian, A.A.Voronova, I.P.Kuranova, E.A.Smirnova, B.K.Vainstein, W.E.Hohne, G.Hansen (1981), *Dokl.Akad.Nauk SSSR*, **258**, 1481.
21. J.H.Ploegman, G.Drent, K.H.Kalk, W.G.J.Hol (1978), *J.Mol.Biol.*, **123**, 557.
22. R.J.Fletterick, H.W.Wyckoff (1975), *Acta Crystallogr., Sect. A*, **31**, 698.
23. C.L.Lawson, R.-G. Zhang, R.W.Schevitz, Z.Otwinowski, A.Joachimiak, P.B.Sigler, to be published.
24. D.L.Smith, S.C.Almo, M.D.Toney, D.Ringe (1989), *Biochemistry*, **28**, 8161.
25. R.B.Honzatko, J.L.Crawford, H.L.Monaco, J.E.Ladner, B.F.P.Edwards, D.R.Evans, S.G.Warren, D.C.Wiley, R.C.Ladner, W.N.Lipscomb (1982), *J.Mol.Biol.*, **160**, 219.
26. B.W.Dijkstra, G.J.H.Van Nes, K.H.Kalk, N.P.Brandenburg, W.G.J.Hol, J.Drenth (1982), *Acta Crystallogr., Sect. B*, **38**, 793.
27. T.L.Poulos, B.C.Finzel, A.J.Howard (1987), *J.Mol.Biol.*, **195**, 687.
28. B.C.Finzel, T.L.Poulos, J.Kraut (1984), *J.Biol.Chem.*, **259**, 13027.
29. G.D.Smith, W.L.Duax, E.J.Dodson, G.G.Dodson, R.A.G.De Graaf, C.D.Reynolds (1982), *Acta Crystallogr., Sect. B*, **38**, 3028.
30. M.Gribskov, L.Wesson, D.Eisenberg, to be published.
31. K.Namba, R.Pattanayek, G.Stubbs (1989), *J.Mol.Biol.*, **208**, 307.
32. R.Bally, J.Delettre (1989), *J.Mol.Biol.*, **206**, 153.
33. I.T.Weber, T.A.Steitz (1987), *J.Mol.Biol.*, **198**, 311.
35. T.N.Bryant, P.J.Shaw, N.P.Walker, P.L.Wendell, H.C.Watson, to be published.
36. S.I.Winn, J.Warwicker, H.C.Watson, to be published.
37. E.N.Baker, T.L.Blundell, J.F.Cutfield, S.M.Cutfield, E.J.Dodson, G.G.Dodson, D.M.Crowfoot Hodgkin, R.E.Hubbard, N.W.Isaacs, C.D.Reynolds, K.Sakabe, N.Sakabe, N.M.Vijayan (1988), *Philos.Trans.R.Soc.London*, **319**, 369.
38. P.Brick, D.M.Blow (1987), *J.Mol.Biol.*, **194**, 287.
39. A.L.Swain, R.H.Kretsinger, E.L.Amma (1989), *J.Biol.Chem.*, **264**, 16620.

40. U.M.Grau, W.E.Trommer, M.G.Rossmann (1981), *J.Mol.Biol.*, **151**, 289.

41. T.Takano, in: *Methods and Applications in Crystallographic Computing* (1984), S.Hall, T.Ashida (eds.), Oxford University Press, Oxford, England.

42. see ref.18 in Appendix A.

43. K.Henrick, C.A.Collyer, D.M.Blow (1989), *J.Mol.Biol.*, **208**, 129.

44. see ref.54 in Appendix A.

45. M.-B.Lascombe, P.M.Alzari, G.Boulot, P.Saludjian, P.Tougard, C.Berek, S.Haba, E.M.Rosen, A.Nisonoff, R.J.Poljak (1989), *Proc.Natl.Acad.Sci.USA*, **86**, 607.

46. V.Z.Pletnev, A.P.Kuzin, L.V.Malinina (1982), *Bioorg.Khim.*, **8**, 1637.

47. P.C.Driscoll, A.M.Gronenborn, L.Beress, G.M.Clore (1989), *Biochemistry*, **28**, 2188.

48. Z.Chen, C.Stauffacher, Y.Li, T.Schmidt, W.Bomu, G.Kamer, M.Shanks, G.Lomonossoff, J.E.Johnson (1989), *Science*, **245**, 154.

49. see ref.48.

50. P.J.Kraulis, G.M.Clore, M.Nilges, T.A.Jones, G.Pettersson, J.Knowles, A.M.Gronenborn (1989), *Biochemistry*, **28**, 7241.

51. S.-E.Ryu, P.D.Kwong, A.Truneh, T.G.Porter, J.Arthos, M.Rosenberg, X.Dai, N.-H.Xuong, R.Axel, R.W.Sweet, W.A.Hendrickson (1990), *Nature*, **348**, 419.

52. see ref.45

53. J.Deisenhofer (1981), *Biochemistry*, **20**, 2361.

54. L.Summers, G.Wistow, M.Narebor, D.Moss, P.Lindley, C.Slingsby, T.Blundell, H.Bartunik, K.Bartels (1984), *Pept.Protein Rev.*, **3**, 147.

55. R.Bone, J.L.Silen, D.A.Agard,to be published.

56. J.Badger, S.Krishnaswamy, M.J.Kremer, M.A.Oliveira, M.G.Rossmann, B.A.Heinz, R.R.Rueckert, F.J.Dutko, M.A.McKinlay (1989), *J.Mol.Biol.*, **207**, 163.

57. M.J.Eck, S.R.Sprang (1989), *J.Biol.Chem.*, **264**, 17595.

58. G.D.Brayer, A.McPherson (1983), *J.Mol.Biol.*, **169**, 565.

59. P.J.Bjorkman, M.A.Saper, B.Samraoui, W.S.Bennett, J.L.Strominger, D.C.Wiley (1987), *Nature*, **329**, 506.

60. J.P.Priestle, H.-P.Schaer, M.G.Gruetter (1989), *Proc.Natl.Acad.Sci.USA*, **86**, 9667.

61. Z.Chen, W.Bode (1983), *J.Mol.Biol.*, **164**, 283.

62. T.Prasthofer, S.R.Phillips, F.L.Suddath, J.A.Engler (1989), *J.Biol.Chem.*, **264**, 6793.

63. see ref.62.

64. S.Krishnaswamy, M.G.Rossmann, to be published.

65. see ref.64.

66. see ref.64

67. P.Schultze, E.Woergoetter, W.Braun, G.Wagner, M.Vasak, J.H.R.Kaegi, K.Wuthrich (1988), *J.Mol.Biol.*, **203**, 251.

68. C.C.F.Blake, M.J.Geisow, S.J.Oatley, B.Rerat, C.Rerat (1990), *J.Mol.Biol.*, **121**, 339.

69. see ref.23 in Appendix A.

70. J.Badger, I.Minor, M.A.Oliveira, T.J.Smith, M.G.Rossmann, to be published.

71. M.Jaskolski, M.Miller, J.K.M. Rao, J.Leis, A.Wlodawer, to be published.

72. see ref.24 in Appendix A.

73. L.Liljas, B.Strandberg, (1984) in: *Biological Macromolecules.* (F.A.Jurnak, A.McPherson,eds.), Vol.1, John Wiley and Sons, New York.

74. P.Hopper, S.C.Harrison, R.T.Sauer (1984), *J.Mol.Biol.*, **177**, 701.

75. W.I.Weis, A.T.Bruenger, J.J.Skehel, D.C.Wiley (1990), *J.Mol.Biol.*, **212**, 737.

76. D.C.Rees, W.N.Lipscomb (1982), *J.Mol.Biol.*, **160**, 475.

77. S.I.Foundling, J.B.Cooper, F.E.Watson, A.Cleasby, L.H.Pearl, B.L.Sibanda, A.Hemmings, S.P.Wood, T.L.Blundell, M.J.Valler, C.G.Norey, J.Kay, J.Boger, B.M.Dunn, B.J.Leckie, D.M.Jones, B.Atrash, A.Hallet, M.Szelke (1987), *Nature*, **327**, 87.

78. H.M.Greenblatt, C.A.Ryan, M.N.G.James (1989), *J.Mol.Biol.*, **205**, 201.

79. P.W.R.Corfield, T.-J.Lee, B.W.Low (1989), *J.Biol.Chem.*, **264**, 9239.

80. P.J.M.Folkers, G.M.Clore, P.C.Driscoll, J.Dodt, S.Koehler, A.M.Gronenborn (1989), *Biochemistry*, **28**, 2601.

81. R.Engh, H.Loebermann, M.Schneider, G.Wiegand, R.Huber, C.-B.Laurell (1989), *Protein Eng.*, **2**, 407.

82. M.Fujinaga, A.R.Sielecki, R.J.Read, W.Ardelt, M.Laskowski Jr, M.N.G.James (1987), *J.Mol.Biol.*, **195**, 397.

83. A.Lewendon, I.A. Murray, W.V.Shaw, M.R.Gibbs, A.G.W.Leslie (1990), *Biochemistry*, **29**, 2075.

84. G.L.Gilliland, E.L.Winborne, J.Nachman, A.Wlodawer, to be published.

85. M.Leijonmarck, A.Liljas (1987), *J.Mol.Biol.*, **195**, 555.

86. T.F.M.La Cour, J.Nyborg, S.Thirup, B.F.C.Clark (1985), *EMBO J.*, **4**, 2385.

87. Z.Xia, F.S.Mathews (1990), *J.Mol.Biol.*, **212**, 837.

88. see ref.33 in Appendix A.

89. O.Epp, R.Ladenstein, A.Wendel (1983), *Eur.J.Biochem.*, **133**, 51.

90. G.M.Clore, E.Appella, M.Yamada, K.Matsushima, A.M.Gronenborn (1990), *Biochemistry*, **29**, 1689.

91. K.Petratos, D.W.Banner, T.Beppu, K.S.Wilson, D.Tsernoglou (1987), *FEBS Lett.*, **218**, 209.

92. Y.Shirakihara, P.R.Evans (1988), *J.Mol.Biol.*, **204**, 973.

93. H.A.Schreuder, J.M.Van Der Laan, W.G.J.Hol, J.Drenth (1988), *J.Mol.Biol.*, **199**, 637.

94. see ref.1.

95. J.Badger, I.Minor, M.A.Oliveira, T.J.Smith, M.G.Rossmann, to be published.

96. R.L.Williams, S.M.Greene, A.McPherson (1987), *J.Biol.Chem.*, **262**, 16020.

97. R.Arni, U.Heinemann, M.Maslowska, R.Tokuoka, W.Saenger (1987), *Acta Crystallogr., Sect. B*, **43**, 549.

98. R.J.Almassy, J.C.Fontecilla-Camps, F.L.Suddath, C.E.Bugg (1983), *J.Mol.Biol.*, **170**, 497.

99. P.Gros, M.Fujinaga, B.W.Dijkstra, K.H.Kalk, W.G.J.Hol, to be published.

100. M.Bolognesi, G.Gatti, E.Menegatti, M.Guarneri, M.Marquart, E.Papamokos, R.Huber (1982), *J.Mol.Biol.*, **162**, 839.

101. S.Vijay-Kumar, C.E.Bugg, W.J.Cook (1987), *J.Mol.Biol.*, 194, 531.

102. C.C.Hyde, E.W.Miles (1990), *Bio-Technology*, **8**, 27.

103. see ref.102.

104. E.N.Baker, E.J.Dodson (1980), *Acta Crystallogr., Sect.A*, **36**, 559.

105. J.E.Gouaux, W.N.Lipscomb (1990), *Biochemistry*, **29**, 389.

106. E.N.Baker (1988), *J.Mol.Biol.*, **203**, 1071.

107. J.F.Davies, T.J.Delcamp, N.J.Prendergast, V.A.Ashford, J.H.Freisheim, J.Kraut, to be published.

108. J.Soman, S.Iismaa, C.D.Stout, to be published.

109. N.K.Vyas, M.N.Vyas, F.A.Quiocho (1988), *Science*, **242**, 1290.

110. T.Skarzynski, A.J.Wonacott (1988), *J.Mol.Biol.*, **203**, 1097.

111. M.M.Yamashita, R.J.Almassy, C.A.Janson, D.Cascio, D.Eisenberg, to be published.

112. T.P.J.Garrett, M.A.Saper, P.J.Bjorkman, J.L.Strominger, D.C.Wiley (1989), *Nature*, **342**, 692.

113. J.S.Sack, M.A.Saper, F.A.Quiocho (1989), *J.Mol.Biol.*, **206**, 171.

114. J.Drenth, W.G.J.Hol, J.N.Jansonius, R.Koekoek (1972), *Cold Spring Harbor Symp.*, **36**, 107.

115. Y.Satow, Y.Watanabe, Y.Mitsui (1980), *J.Biochem.(Tokyo)*, **88**, 1739.

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

117. J.Walter, W.Steigemann, T.P. Singh, H.Bartunik, W.Bode, R.Huber (1982), *Acta Crystallogr., Sect. B*, **38**, 1462.

118. E.Lolis, G.A.Petsko (1990), *Biochemistry*, **29**, 6619.

119. D.Dreusicke, P.A.Karplus, G.E.Schulz (1988), *J.Mol.Biol.*, **199**, 359.

120. F.S.Mathews, P.Argos, M.Levine (1972), *Cold Spring Harbor Symp.*, **36**, 387.

121. D.E.Tronrud, M.F.Schmid, B.W.Matthews (1986), *J.Mol.Biol.*, **188**, 443.

122. O.Herzberg, to be published.

123. A.E.Eriksson, P.M.Kylsten, T.A.Jones, A.Liljas (1988), *Proteins.Struct., Funct.*, **4**, 283.

124. J.T.Bolin, D.J.Filman, D.A.Matthews, R.C.Hamlin, J.Kraut (1982), *J.Biol.Chem.*, **257**, 13650.

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

126. see ref.75.

127. J.H.Hurley, P.E.Thorsness, V.Ramalingam, N.H.Helmers, D.E.Koshland, Jr., R.M.Stroud (1989), *Proc.Natl.Acad.Sci.USA*, **86**, 8635.

128. J.T.Bolin, D.J.Filman, D.A.Matthews, R.C.Hamlin, J.Kraut (1982), *J.Biol.Chem.*, **257**, 13650.

129. J.J.Birktoft, G.Rhodes, L.J.Banaszak (1989), *Biochemistry*, **28**, 6065.

130. A.M.Silva, M.G.Rossmann (1985), *Acta Crystallogr., Sect. B*, **41** 147.

131. H.M.Holden, D.E.Tronrud, A.F.Monzingo, L.H.Weaver, B.W.Matthews (1987), *Biochemistry*, **26**, 8542.

132. D.C.Rees, M.Lewis, W.N.Lipscomb (1983), *J.Mol.Biol.*, **168**, 367.

133. K.D.Watenpaugh, L.C.Sieker, L.H.Jensen (1980), *J.Mol.Biol.*, **138**, 615.

134. A.H.Robbins, C.D.Stout (1989), *Proc.Natl.Acad.Sci.USA*, **86**, 3639.

135. I.Fita, M.G.Rossmann (1985), *Proc.Natl.Acad.Sci.USA*, **82**, 1604.

136. F.Colonna-Cesari, D.Perahia, M.Karplus, H.Eklund, C.I.Branden, O.Tapia (1986), *J.Biol. Chem.*, **261**, 15273.

137. R.Engh, H.Loebermann, M.Schneider, G.Wiegand,R.Huber, C.-B.Laurell (1989), *Protein Eng.*, **2**, 407.

REFERENCES

1. Wetlaufer, D. (ed.) (1984) *The Protein Folding Problem.* Westview, Boulder, Co.

2. Ghelis, C. & Yon, J. (1980) *Protein Folding.* Academic Press, New York.

3. Kabsch, W. & Sander, C. (1983) Dictionary of Protein Structure. Pattern Recognition of Hydrogen-Bonded Geometrical Features. *Biopolymers*, **22**, 2577 - 2637.

4. Tanford, C. (1968) Protein Denaturation. *Adv. Protein Chem.*, **23**, 121 - 282.

5. Creighton, T., E. (1990) Protein Folding. *Biochem. J.*, **270**, 131 - 146.

6. Skolnick, J. & Koliński, A. (1989) Computer Simulation of a Globular Protein Folding and Tertiary Structure. *Annu. Rev. Phys. Chem.*, **40**, 207 - 235.

7. Skolnick, J. & Koliński, A. (1990) Simulation of the Folding of Globular Protein. *Science*, **250**, 1121 - 1125.

8. Schultz, G. E. & Shrimer, R. A. (1979) *Principles of Protein Structure.* Springer Verlag, New York.

9. Periti, P. F. (1974) Bayesian Approach to the Recognition of Discrete Patterns with an Application to a Problem of Protein Molecular Structure. *Bull. 'Chem. Farm.*, **113**, 187 - 218.

10. Poland, D. & Sheraga, H. A. (1970) *Theory of Helix-Coil Transition in Biopolymers.* Academic Press, New York.

11. Garnier, J., Ostgusthorpe, D. & Robson, B. (1978) Analysis of the Accuracy and Implications of Simple Methods for Predicting the Secondary Structure of Globular Proteins. *J. Mol. Biol.*, **120**, 97 - 120.

12. Chou, P. & Fasman, G. (1974) Conformational Parameters for Amino Acids in Helical, Beta-Sheet and Random Coil Regions Calculated from Proteins. *Biochemistry*, **13**, 211 - 222.

13. Lim, V. I. (1974) Structural Principles of the Globular Organization of Protein Chains. A Stereochemical Theory of Globular Protein Secondary Structure. *J. Mol. Biol.*, **88**, 857 - 872.

14. Qian, N. & Sejnowski, T. J. (1988) Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J. Mol. Biol.*, **202**, 865 - 884.

15. Holley, L. H. & Karplus, M. (1989) Protein Secondary Structure Prediction with a Neural Network. *Proc. Natl. Acad. Sci., U.S.A.*, **86**, 152 - 156.

16. Kneller, D. G., Cohen, F. E. & Landridge, R. (1990) Improvements in Protein Secondary Structure Prediction by an Enhanced Neural Network. *J. Mol. Biol.*, **214**, 171 - 182.

17. McGregor, M. J., Flores, T. P. & Sternberg, M. J. (1989) Prediction of β-turns in Proteins using Neural Network. *Prot. Eng.*, **2(7)**, 521 - 526.

18. Bohr, H., Bohr, J., Brunak, S., Cotterill, R. M., Lautrup, B., Norskov, L., Olsen, O. H. & Petersen, S. B. (1988) Protein Secondary Structure and Homology by Neural Networks –The α-Helices in Rhodopsin. *FEBS Lett.*, **241**, 223 - 228.

19. Holbrook, S. R., Muskal, S. M. & Kim, S. H. (1990) Predicting surface exposure of amino acids form protein sequence. *Protein Eng.*, **3**, 659 - 665.

20. Vieth, M. & Koliński, A. (1991) Prediction of Protein Secondary Structure by an Enhanced Neural Network. *Acta Biochim. Polon.*, **38**, 335 - 351.

21. Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986) in *Parallel Distributed Processing.* Vol 1, pp. 318 - 362, MIT Press, Cambridge, MA.

22. Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992) Selection of representative protein data sets. *Protein Science*, **1**, 409 - 417.

23. Bernstein, C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977) The Protein Data Bank: A Computer-based Archival File for Macromolecular Structures. *J. Mol. Biol.*, **112**, 535 - 542.

24. Frommel, C. (1984) The Apolar Surface Area of Amino Acids and Its Empirical Correlation with Hydrophobic Free Energy. *J. Mol. Biol.*, **111**, 247 - 260.

25. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A. & DeLisi, C. (1987) Hydrophobicity Scales and Computational Techniques for Amphipathic Structures in Proteins. *J. Mol. Biol.*, **195**, 659 - 685.

26. Matthews, B. W. (1975) Comparison of the Predicted and Observed Secondary Structure of T4 Phage Lysozyme. *Biochim. Biophys. Acta*, **405**, 442 - 451.

27. Gregoret, L. M. & Cohen, F. E. (1990) Novel Method for the Rapid Evaluation of Packing in Protein Structures. *J. Mol. Biol.*, **211**, 959 - 974.