Comparison of Two Variations of Neural Network Approach to the Prediction of Protein Folding Pattern

Inna Dubchak *, Stephen R. Holbrook# and Sung-Hou Kim*#

#Structural Biology Division, Lawrence Berkeley Laboratory

*Department of Chemistry, University of California at Berkeley

Berkeley, CA 94720

ILDUBCHAK@LBL.GOV

Abstract

We have designed, trained and tested two types of neural networks for the prediction of protein folding pattern from sequence. Here we describe the differences in the networks and compare their performance on a variety of proteins. Both network representations are generally successful in predicting protein fold and can also be used together to confirm a prediction.

Introduction

The prediction of protein structure from its amino acid sequence is a central, unsolved problem in molecular biology. Despite efforts by a large number of laboratories over many years, we are still far from a solution to this problem. The size and complexity of proteins makes a strictly theoretical approach intractable. A more hopeful direction is to use the rapidly growing number of experimentally determined protein structures and sequences to empirically extract the principles which govern protein folding and then apply these rules to the prediction of new structures. This can be done at many levels, from the prediction of the secondary structure of an individual amino acid (Qian & Sejnowski 1988, Holley & Karplus 1989) to the surface exposure of amino acids (Holbrook, Muskal, & Kim 1990), and even tertiary interactions (Bohr et al. 1990). Unfortunately, as the complexity of the prediction and the number of parameters increase, the data to parameter ratio decreases. Thus, we have chosen a top-down path in which we first attempt to predict the overall fold of the protein using very few parameters. The detailed structure may then be completed using other prediction algorithms and energetic calculations. Here we study two variations of a neural network approach to prediction of the overall protein fold.

It is becoming clear that there are a limited number of protein folding patterns to serve as the core or scaffold around which variations are added to perform specific protein functions. In fact, there are now many examples of functionally similar and dissimilar proteins with the same fold (Finkelstein & Ptitsyn 1987). Most proteins pack their secondary structures into one of a limited number of basic geometries and chain topologies. The definition "folding pattern" (Levitt & Chotia 1976) or "topologies" (Richardson 1977) ignore the variations of length and orientation of α- and β-segments and describe only their mutual positions. Some of the macro folding patterns are: the globin fold, the immunoglobulin fold, the nucleotide binding (Rossmann) fold, 4α-helical bundles, parallel α/β barrels, and antiparallel β barrels (jelly roll fold). On a smaller scale, the Greek key motif, EF-hand, helix-loop-helix, kringle motifs and others form fragments of protein structure.

We extended the approach which Muskal and Kim (Muskal & Kim 1992) originally applied to prediction of secondary structure composition, to the discrimination of protein folding patterns (Dubchak, Holbrook, & Kim 1993). In this method (Representation I) there are 21 real-valued input nodes representing the number of amino acids and amino acid percent composition of the protein as a whole, or that of the domain of interest.

In the second version of folding class prediction (Representation II) we made attempts to reduce the number of inputs and accordingly variables by using reasonable physical subdivision of all amino acids into a few classes. In relating sequence and structure the basic classification of residues is usually in terms of their hydrophilic and hydrophobic character. Both groups of authors (Qian & Sejnowski 1988, Holley & Karplus 1989) used physicochemical properties of the amino acid residues such as hydrophobicity, charge, side chain bulk, and backbone flexibility as alternative ways of representing the inputs to the network for prediction of secondary structure. Qian and Sejnowski also provided the network with global information such as an average hydrophobicity of the protein and the position of the residue in the sequence. All these attempts did not help to improve the prediction reliability over the basic scheme. However Kneller and co-authors (Kneller, Cohen, & Langridge 1990) used the scheme with additional helix hydrophobic moment and strand hydrophobic moment input units which improved the testing statistics slightly in comparison with (Qian & Sejnowski 1988).

Basically there are two opposite viewpoints (with a variety of opinions in between) regarding the distribution of hydrophobic and hydrophilic sites along the sequence of protein. Chothia and Finkelstein (Chothia & Finkelstein 1990) proved that this distribution is unique in the case of the globin sequence. The formation of the native secondary structure of any protein brings together sites of the same character to form hydrophobic and hydrophilic surfaces. Protein interiors are occupied mainly by non-polar residues and occasionally by neutral residues. The polar atoms in neutral residues usually form hydrogen bonds within their own secondary structures so the surfaces between secondary structures are almost entirely hydrophobic in character. The converse is also true: the sites in protein that are highly exposed to the solvent are nearly always occupied by polar or neutral residues. The same is supposedly true for other folds. Since an architecture of a definite fold is specific feature it is reasonable to use this kind of classification for neural net input.

Local interactions inside structural segments can often determine the protein secondary structures despite the presence of much stronger long range interactions between different segments. Some of the hydrophobicity scales (Ponnuswamy 1993) takes into account protein structural class information namely the environment of each amino acid residue when estimating its hydrophobicity. The other scale (optimal matching hydrophobicity) was derived on the assumption that families of proteins that fold in the same way, do so because they have the same pattern of residue hydrophobicities along their amino acid sequences (Eisenberg 1984).

At the same time, White and Jacobs (White & Jacobs 1990) studied the statistical distribution of hydrophobic residues along the length of protein chains using a binary hydrophobicity scale which assigns hydrophobic residues a value of one and non-hydrophobes a value of zero. The resulting binary sequences are tested for randomness using the standard run test. For the majority of the 5,247 proteins examined, the distribution of hydrophobic residues along a sequence cannot be distinguished from that expected for a random distribution. The authors suggest that (a) functional proteins may have originated from random sequences, (b) the folding of proteins into compact structures may be much more permissive with less sequence specificity than previously thought, and (c) the clusters of hydrophobic residues along chains which are revealed by hydrophobicity plots are a natural consequence of a random distribution and can be conveniently described by binomial statistics.

In our calculations we used a less simplistic classification than White and Jacob, and made an attempt not only to group the residues according to their relative hydrophobicity, but to take into consideration their immediate environment.

Several scales for classification of hydrophobic-hydrophilic character of residues have been published. Most of these scales differ only in details and so show high correlation. We used the hydrophobicity groups from (Chothia & Finkelshtein 1990), which are very similar to those of consensus scale (Eisenberg 1982), which was designated to mitigate the effect of outlying values in any one scale produced by the peculiarities of the method .

Methods

Database: A critical problem in classification of protein folding types by neural networks is the limited number of protein examples of known three-dimensional structure. We have partly overcome this problem by using the much larger protein sequence database. Proteins of the same family from different organisms, showing high sequence homology with proteins of known structure were assumed to have the same overall folding pattern and were used to greatly increase the number of input examples for use in network training and testing. Thus, we selected the proteins for which the crystal structure is known to have the fold of interest and then retrieved homologous proteins from the SWISS-PROT sequence database. We chose to begin our studies on four diverse folding patterns for which there exists a relatively large number of known structures and sequence analogs. Other folding motifs can easily be added to the prediction scheme as sufficient examples of known structure are characterized.

Still, in some cases the number of variable parameters did not exceed the number of independent observations which led to a lack of generalization. This problem, in turn, led to our experiments with an alternate input representation specified by fewer parameters.

Protein structure information was from the Brookhaven Protein Data Bank (Bernstein et al. 1977) (PDB) and publications describing proteins of known crystal structure but not yet deposited in the Brookhaven PDB. Sequence information was from the Swiss-Prot database (SP) Release 20 (Bairoch & Boeckmann 1991). Therefore, databases were compiled for each of the following folding patterns 1) 4α-helical bundles (BUNDLE), 2) Eight stranded parallel α/β barrels (BARREL), 3) Nucleotide binding or Rossmann (NBF) fold, 4) Immunoglobulin fold (IGF) and 5) other or unclassified (UNC) folds . A full list of used protein sequences and structures is presented in (Dubchak, Holbrook, & Kim 1993). The number of proteins in each database and average characteristics for each class are given in Table 1. Ribbon drawings of representative proteins from each of the four folding patterns are shown in Figure 1(a-d).

For single domain proteins, only sequences for which the folding pattern of interest comprised more than 50% of the total sequence were used. In the case of multidomain proteins, sequences of a domain were used when the pattern of interest accounted for more than 50% of that domain, even though the overall percentage of the total protein was less than 50%. Only the single domain of. multidomain proteins that contained the folding pattern of interest, was used in our studies.

Neural Networks: The architecture of the networks is shown in Figure 2. The neural networks used in this study were of the feed forward type with either zero (perceptron) or one hidden layer and weights adjusted by conjugate gradient minimization using the computer program package BIOPROP (Muskal & Kim 1992).

The protein sequences were converted to 21 numbers (percent composition of amino acids and number of amino acids in a sequence) in Representation I, or seven (hydrophobic composition) numbers in Representation II. The hydrophobic composition characteristics are: number of amino acid residues in the sequence; overall percent composition of polar residues (Arg, Lys, Glu, Asp, Gln, and Asn); overall percent composition of neutral residues (Gly, Ala, Ser, Thr, Pro, His, and Tyr); overall percent composition of hydrophobic residues (Cys, Val, Leu, Ile, Met, Phe, and Trp); number of sequential transitions from polar to neutral, from polar to hydrophobic, and from neutral to hydrophobic residues. Technically all values were normalized for better convergence during an optimization of neural net weights.

In the networks utilizing percent amino acid composition (as well as number of residues) as input, no single amino acid percentage was seen to be significantly different among the different classes. Likewise, in the networks using hydrophobic composition as input it is apparent that no single characteristic out of 7 for any of 4 folding classes is significantly different.

As illustrated in Figure 2, networks of from one to four outputs were used to classify a protein into one of the four folding patterns or as "other". When one output node is used, the choice is between a particular fold and all others. As additional nodes are added the network makes a decision between the various output patterns, with a low activity to all outputs indicating "other", or that the fold of interest does not correspond to any of the available choices.

Each of the four databases for different folding classes was separated into independent training and testing sets, i.e. no protein family from training was included in testing. This was necessary because each database contains a large number of similar (but not identical) examples and therefore random sampling will perform very well simply because of memorization. For each of the four databases, at least two different training-testing set combinations were constructed so as to test the effect of specific proteins on training/testing. The same operation was made for a combined sets of two, three and four folding classes representatives. This procedure is described in detail elsewhere (Dubchak, Holbrook, & Kim 1993).

In order to optimize the number of hidden nodes for each particular training-testing set combination, networks containing from 0 to 9 hidden nodes were systematically trained and tested numerous times to find the architecture which gave the best performance in predicting the testing set. The networks with the optimal number of hidden nodes were then trained repeatedly (10 times) from random initial starting weights to search more completely the parameter space and the weights producing the highest testing scores were chosen as the "optimal network weights". One hundred cycles of conjugate gradient minimization was found optimal to provide maximal generalization (best training without memorization) for most of the sets and was thus used in all training.

Results

The results of network testing for each of the four folding classes with from one to four output nodes are summarized in Table 2. Thus, there are four types of trained networks for each representation, each making a prediction as to whether a protein contains a specific folding pattern. At least two separate training-testing sets were considered for each folding pattern.

As can be seen from Table 2, both approaches give fairly reliable results for all the groups of proteins. In some cases Representation II shows an improvement in prediction compared to Representation I, especially in distinction between folding classes 1/3, 2/3, 1/3/4, and 1/2/3. However, in some cases performance became worse, as in the case of class 1 prediction and distinction between classes 1/2, 2/4 and 1/2/4. Thus, we propose that for maximum confidence of prediction, each given sequence should be tested sequentially on all 30 networks, namely on 15 networks for each of the two Representations - four networks with 1 output node, six 2 output node networks representing all possible pairwise combinations of 4 classes, four 3 output networks allowing assignment of the sequence to one of three folds and 1 net with 4 outputs which makes a decision between all available folding classes. Each network gives one integer number (1-4 for known classes and 5 for unclassified) indicating the most favorable type of folding. Each folding class participates in 8 nets (per Representation), so the maximum number of predictions for each class is 8, if the results of all testings are in accordance. Tests have shown that the fold is strongly predicted if the number of positive predictions for this particular fold is equal to or greater

Table 1. Characteristics of protein database used in this study

Folding class*	Number of crystal structures/ examples Swiss-Prot	Number of amino acids in sequence	Av. % composition of residues in database / stand. deviations			Av. number of sequential transitions** /st. dev.			Average%/St.dev.	
			Polar	Neutral	Hydrophobic				α-helices	β-strands
			P	N	H	P-N	P-H	N-H		
I	12/68	103 - 232	33.6/3.8	36.0/4.9	30.5/3.9	24.6/3.8	21.0/3.6	21.92.9	70.3/2.9	0.0/0.0
II	8/92	194 - 394	30.3/1.5	39.4/2.0	30.2/1.7	22.9/2.3	18.1/2.6	25.1/2.8	45.3/5.7	15.3/5.7
III	10/167	171 - 374	30.4/3.9	37.1/3.5	32.6/2.5	22.0/3.1	20.4/2.4	23.1/4.5	36.3/10.1	16.1/4.5
IV	9/180	105 - 400	24.3/3.3	46.3/2.8	29.4/0.8	20.6/2.7	18.2/1.9	27.1/3.5	7.1/4.5	44.1/4.7

* I - 4α-helical bundle; II -Parallel α/β barrels; III - Nucleotide binding or Rossmann ; IV - Immunoglobulin

** Normalized by division on number of amino acids in a protein

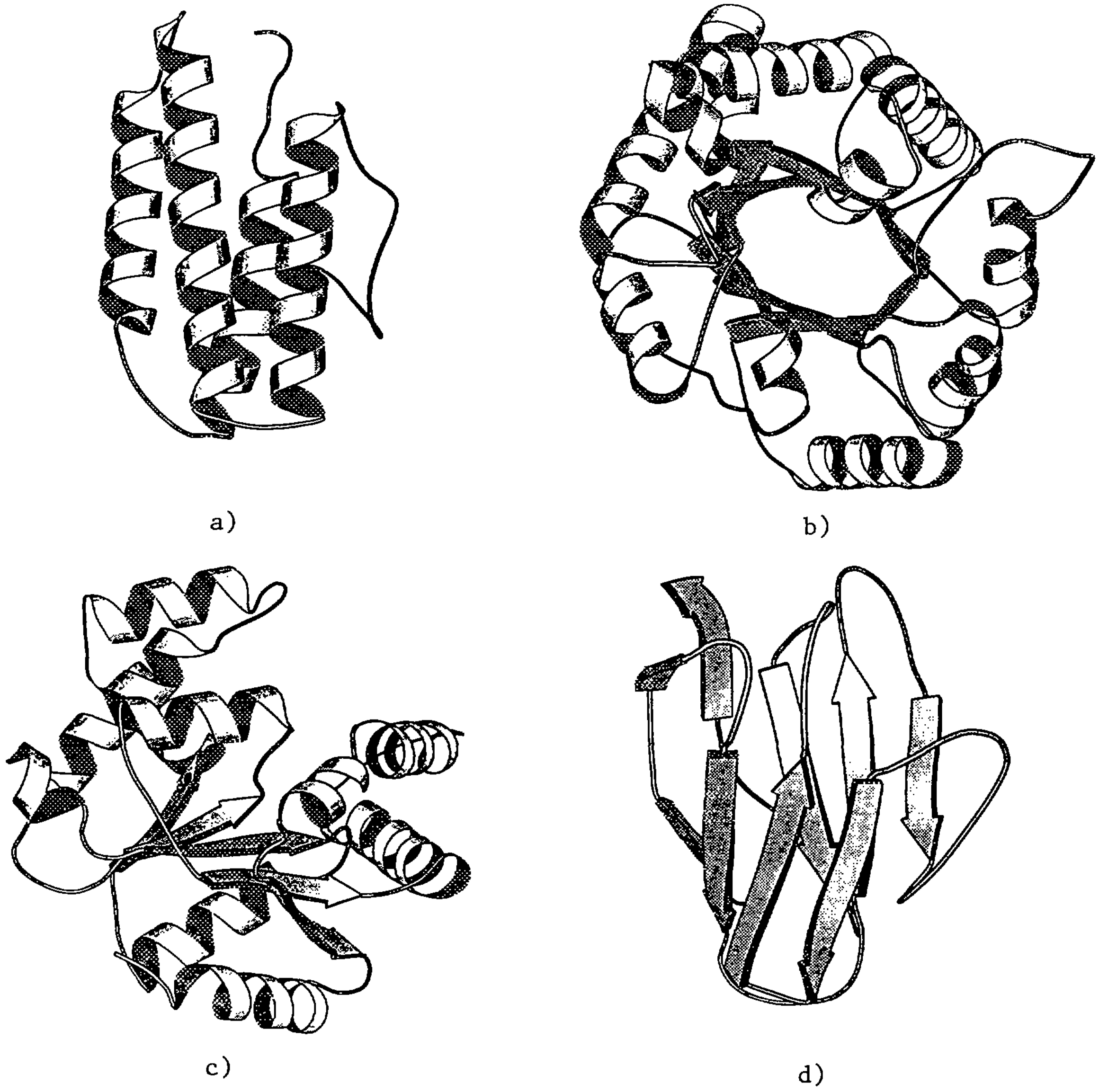

Figure 1. Schematic drawings of members of the four folding patterns predicted. a) Four α-helical bundle fold (BUNDLE): The specific example shown is Myohemerythrin (PDB code: 2MHR, 118 amino acids); b) Eight stranded parallel (α/β) barrel (BARREL): The specific example shown is: Triosephosphate isomerase (PDB code: 1TIM, 248 amino acids); c) Nucleotide binding or Rossmann fold (NBF): The specific example shown is: Adenylate kinase (PDB code: 3ADK, 194 amino acids); d) Immunoglobin fold (IGF): The specific example shown is the variable portion of the Bence-Jones immunoglobin (PDB code: 1REI, 107 amino acids). Arrows represent β-strands and coils represent helical regions.

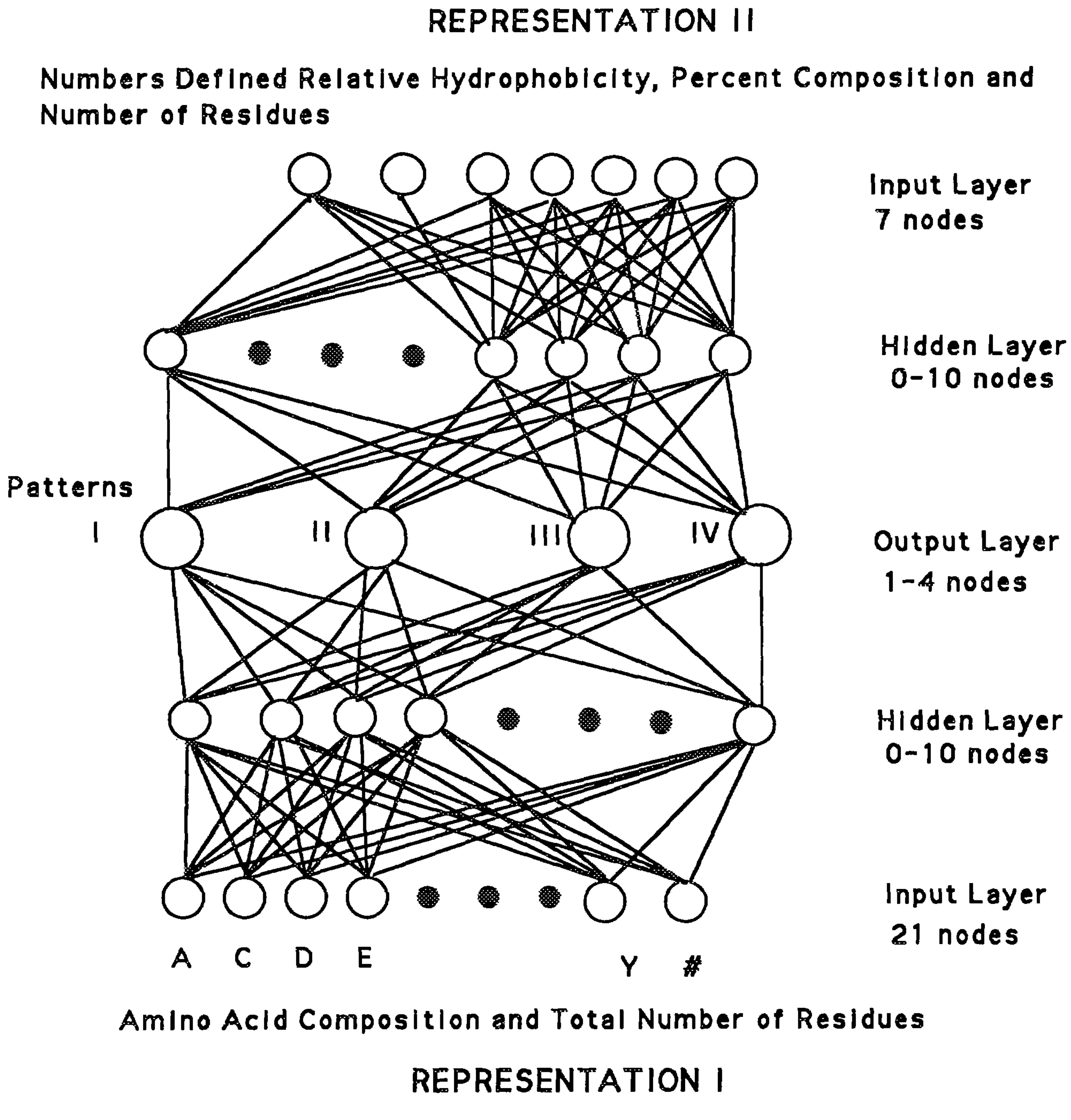

Figure 2. Schematic diagram of the architecture of the computational neural networks discussed in the text. The circles represent computational nodes and the lines correspond to weighted links between the nodes.

Table 2. Examples of performance of independent testing sets for Representations I & II.

Fold	Number of outputs	Groups of proteins in testing set*	Number of examples in training/ testing sets	Percent of known proteins in testing set predicted correct	
				Prediction based on percent composition Representation I	Prediction based on relative hydrophobicity Representation II
1	1	A	122/14	100	71.4
2	1	B	124/32	81.2	100
3	1	C	320/55	86.6	89.6
4	1	D	174/76	65.7	77.5
1/2	2	E,F	120/16	100/88.9	88.9/61.5
1/3	2	E,C	134/18	87.5/60.0	100/87.5
1/4	2	E,D	151/20	50.0/100	70.2/93.0
2/3	2	B,H	149/28	81.2/91.6	100/100
2/4	2	B,D	178/28	100/100	92.8/85.7
3/4	2	C,I	178/28	100/89.2	100/92.3
1/2/3	3	G,B,K	179/36	61.5/81.8/83.3	69.2/90.9/83.3
1/3/4	3	E,K,D	205/33	72.7/85.7/100	85.7/85.7/100
2/3/4	3	B,K,D	209/36	54.5/100/76.9	72.7/100/72.7
1/2/4	3	E,B,D	205/30	88.8/90.9/63.6	77.7/90.9/72.7
1/2/3/4	4	E,F,G,I	290/41	62.5/100/75.0/69.2	75.0/87.5/75.0/76.9

* Independent groups of proteins were separated into testing sets: A. M-CSF,GM-CSF,G-CSF; B. Aldolase; C. Lactate dehydrogenase; D. T-cell surface glycoprotein; E. Tar(asp) receptor, cytochrome B-562, hemerythrin; F. D-xylose isomerase; G. Lactate dehydrogenase; H. RAB proteins; I. Histocompatibility antigen, class 1; G. Cytochrome C'; K. Glyceraldehyde-3-phosphate dehydrogenase;

than 6 out of 8 and no more than 4 networks predict another fold; a weak prediction is defined as 5 out of 8 positive predictions per Representation.

After sequential testing on 30 nets we compare predictions by the two methods and judge them reliable when they are in accordance, even when one of them is weak. However, if a prediction of protein fold by at least one method is strong we can use this as a first approximation and take into consideration other chemical and biochemical information. For example, the testing of five interleukin 2 cytokines (Bairoch & Boeckmann 1991) from different organisms gives a strong indication of BUNDLE fold by Representation I, namely 3 nets with 2 outputs and 3 nets with 3 outputs assign these molecules to this fold but 1 output nets recognize them as Rossmann fold and 4 output net as UNKNOWN. Representation II gives only a weak indication of the same fold (1 by the single output nets, two by 2 output nets and two by 3 output nets). We believe that a final prediction of this fold as BUNDLE is reliable based on a hybrid strategy.

In order to learn strengths and weaknesses of both Representations, we undertook testing of 250 proteins from the Brookhaven Protein Databank (Bernstein et al. 1977). The advantage of these testing samples was known tertiary structure which allowed us to evaluate where either one or both methods don't work.

Since it has been demonstrated (Muskal & Kim 1992) that neural networks can accurately predict secondary structure composition from amino acid percentage, we felt it important to verify that the folding pattern predictions were not just reflecting this ability. In the group of 19 proteins of high alpha-helix (>50% alpha-helix) which we tested, only 3 of them were incorrectly predicted by two methods as belonging to BUNDLE fold. They are: DNA binding regulatory protein (1WRP code in Protein Databank) with a fold consisting of 6 non-parallel helices, cytochrome C2 (code 2C2C, 5 helices in structure) and calcium-binding parvalbumin (4CPV, 6 helices). No protein which was not high in helical content was predicted as a 4α-helical bundle. In the case of myohemerythrin the fold is correctly predicted as BUNDLE only by Representation II. Based on these examples, the presence of cofactors (three of the four cases) in the structure may disturbs a prediction, probably because of the special amino acids necessary for cofactor ligation.

The importance of a combined approach is obvious in the prediction of the α-chain of globins (1FDH, 1HBS, 1HDS, 1PMB, 1ECD in the Protein Data Bank). After correction of the neural network weights by addition of an extra node for the presence or absence of the heme cofactor, still three out of five proteins were incarcerate predicted as BUNDLE by Representation 1, but all five of them assigned to UNKNOWN fold by Representation II. The group of globins might be the next candidate for addition to our scheme of prediction because of significant number of known structures for training.

As for proteins with very low or zero helical content, they are sometimes incorrectly predicted as belonging to the immunoglobulin fold. Examples are: hydrolase (1RNT) , containing 16% helix and 56% strand, alpha-bungarotoxin (2ABX) - 0 and 89% and actinoxanthin (1ACX) - 0 and 56%. This reflects a prejudice in the networks to classify low helical proteins as immunoglobin fold, likely due to a lack of proteins of other folding patterns with low helix content which could be used as false examples.

We can classify all failures into two groups - those which assign a protein to a definite folding class to which it does not belong and those which lack the ability to recognize folding pattern and therefore classify a protein as belonging to unknown fold. While we have observed incorrect assignment to each of the 4 folds we have studied, the $(\alpha\beta)_8$ barrel folding motif is the one which most often gives ambiguous results.

The prediction of the BARREL fold presents a situation which needs more detailed analysis. Structures of aspartate aminotransferase (2AAT), leucine-binding protein (2LBP), dihydrofolate reductase (4DFR) and phosphofructokinase (3PFK) represent combinations of helices and strands in approximately the same quantitative relation as in $(\alpha\beta)_8$ barrels, but with different relative position of helices and strands. Still, all of them were predicted as BARREL by both Representations. The common features of all of them are percent composition of helices and strands is close to those of barrel structures and number of helical and strand fragment in structure is not less than 8 for both, sometimes 10 and more. Another source of confusion arises from the prediction of multidomain proteins. So, the two-domain structures of L-arabinose-binding protein (1ABP) and rhodanese (1RHD) were classified as $(\alpha\beta)$ barrels by both methods. However, when predictions were made on the individual domains of these proteins no one network gives such a prediction and the conclusion can be made that they belong to unclassified fold.

In the case of NBF and BARREL prediction, incorrect results with both Representations have been obtained on some mixed α/β structures. For example, lysozyme (2LZM) was incorrectly, but strongly, predicted as a nucleotide binding fold by Representation I and correctly as unknown by Representation II. Other members of the lysozyme family (1LYM, 1LZT) were predicted as unknown by Representation I, but incorrectly as BARREL by Representation II. Another member of the family (1LZ2) was correctly predicted as UNKNOWN by both Representations. This emphasizes the importance of examining several members of a family when possible.

Conclusions

We have designed and implemented two types of neural networks for the prediction of protein folding pattern from sequence. The input representations in these two network types utilize a different number of parameters and contain a different level of information. Representation I, which requires about three times the number of variable parameters may be able to pick up specific features characteristic of certain folding types. On the other hand Representation II should be superior in prediction by generalization or extrapolation. For the most reliable prediction each of these methods should consistently predict the same fold, whether using one, two, three or four outputs and the two representations should agree with each other.

Several points should be kept in mind when predicting protein fold using the networks described above: 1) it is very important to use only the sequence of a single domain in making the prediction (because we are only predicting a domain structure), 2) the method is applicable to proteins of size range 100-400 amino acids, since this is the range of the proteins used in training and 3) testing of several members of a protein family can greatly increase reliability of results and should be attempted whenever possible. We are in the process of automating protein family prediction and expect the prediction accuracy to be greatly increased.

Finally, another approach to checking the predicted model is to look for consistency with other predictions of protein structure. With this in mind we have developed an integrated neural network package for protein structure prediction. This program package, which we call PROBE (PROtein prediction at BErkeley) (Holbrook, Dubchak, & Kim 1993), not only makes predictions of the folding pattern by the two methods discussed above, but also compares them with the secondary structure predictions by the method (Qian & Sejnowski 1988) and a modified version of the networks of Holley and Karplus (Holley & Karplus 1989), as well as the overall secondary structure composition (Muskal & Kim 1992). Agreement of these various networks leads to a high confidence in protein fold prediction.

References.

Bairoch, A; and Boeckmann, B. 1991. The SWISS-PROT sequence data bank. *Nucl. Acids. Res.* 19 Suppl.: 2247-49.

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

Bohr, H.; Bohr, J.; Brunak, S.; Cotterill, R.M.J.; Fredholm, H.; Lautrup, B.; and Petersen, S.B. 1990. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural network. *FEBS letters.* 261(1): 43-46.

Chothia, C.; and Finkelstein, A.V. 1990 .The classification and origins of protein folding patterns. *Annual Review of Biochemistry* 59: 1007-39.

Dubchak, I.; Holbrook, S. R.; and Kim, S.-H. 1993. Prediction of Protein Folding Class from Amino Acid Composition. *Proteins: Structure, Function, and Genetics.* Forthcoming

Eisenberg, D. 1984. Three-Dimensional Structure of Membrane and Surface Proteins. *Annu. Rev. Biochem.* 53: 595-623.

Eisenberg, D.; Weiss, R.M.; and Terwilliger, T.C. 1982. Hydrophobic Moment and Protein Structure. *Faraday Symp. Chem. Soc.* 17: 109-120.

Finkelstein, A.V.; and Ptitsyn, O.B. 1987. Why Do Globular Proteins Fit the Limited Set of Folding Patterns? *Prog. Biophys. Molec. Biol.* 50: 171-190.

Holbrook, S.R.; Dubchak, I.; and Kim, S.-H. 1993. PROBE: A Computer Program Employing an Integrated Neural Network Approach to Protein Structure Prediction. *BioTechniques..* Forthcoming.

Holbrook, S.R.; Muskal, S.M.; and Kim, S.-H. 1990. Predicting surface exposure of amino acids from protein sequence. *Prot. Eng.* 3(8): 659-665.

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

Kneller, D.G.; Cohen, F.E.; and Langridge, R. 1990. Improvement in Protein secondary structure prediction by an enhanced neural network. *J. Mol. Biol.* 214(1): 171-182.

Levitt, M.; and Chotia, C. 1976. Structural Patterns in Globular Proteins. *Nature* 261: 552-557.

Muskal, S.M; and Kim, S.-H. Predicting protein secondary structure content: a tandem neural network approach. *J. Mol. Biol.* 225: 713-727.

Ponnuswamy, P.K. 1993. Hydrophobic Characteristics of Folded Proteins. *Prog. Biophys. Molec. Biol.* 59(1): 57-103

Qian, N.; and Sejnowski, T.J. 1988. Predicting the Secondary Structure of Globular Proteins Using Neural Network Models. *J.Mol.Biol.* 202: 865-884.

Richardson, J.S. 1977. β-Sheet Topology and the Relatedness of Proteins. *Nature* 268: 495-500, 1977

White, S.H.; and Jacobs, R.E. 1990. Statistical Distribution of Hydrophobic Residues Along the Length of Protein Chains. Implication for Protein Folding and Evolution. *Biophys. J.* 54(4): 911-921.