PROTEINS: Structure, Function, and Genetics 16:79–91 (1993)

Prediction of Protein Folding Class From Amino Acid Composition

Inna Dubchak,[1] Stephen R. Holbrook,[2] and Sung-Hou Kim[1,2]
[1]Department of Chemistry, and [2]Structural Biology Division, Lawrence Berkeley Laboratory, University of California at Berkeley, Berkeley, California 94720

ABSTRACT **An empirical relation between the amino acid composition and three-dimensional folding pattern of several classes of proteins has been determined. Computer simulated neural networks have been used to assign proteins to one of the following classes based on their amino acid composition and size: (1) 4α-helical bundles, (2) parallel $(\alpha/\beta)_8$ barrels, (3) nucleotide binding fold, (4) immunoglobulin fold, or (5) none of these. Networks trained on the known crystal structures as well as sequences of closely related proteins are shown to correctly predict folding classes of proteins not represented in the training set with an average accuracy of 87%. Other folding motifs can easily be added to the prediction scheme once larger databases become available. Analysis of the neural network weights reveals that amino acids favoring prediction of a folding class are usually over represented in that class and amino acids with unfavorable weights are underrepresented in composition. The neural networks utilize combinations of these multiple small variations in amino acid composition in order to make a prediction. The favorably weighted amino acids in a given class also form the most intramolecular interactions with other residues in proteins of that class. A detailed examination of the contacts of these amino acids reveals some general patterns that may help stabilize each folding class.**

Key words: protein structure prediction, neural networks, amino acid composition, protein folding classes, 4α-helical bundles, parallel $(\alpha/\beta)_8$ barrels, nucleotide binding fold, immunoglobulin fold

INTRODUCTION

It is becoming clear that nature has created a limited number of protein folding patterns that serve as the core or scaffold around which variations are added to perform specific protein functions. As new structures are determined by X-ray crystallography or high resolution multidimensional NMR studies it is more and more likely that a protein will belong to one of the previously identified folding patterns. Some of the macro folding patterns are: the globin fold, the immunoglobin 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.

Computational neural networks have recently been applied to a variety of problems in protein structure including secondary structure prediction,[1,2] β-turn classification,[3] prediction of surface accessibility of amino acid residues,[4] and the propensity of cysteine for disulfide bond formation.[5] Attempts at application of neural nets to the prediction of protein tertiary structure, based on the analysis of their sequences[6,7] were successful at prediction of homologous proteins or proteins of similar function, but generally were not able to predict structures unrelated to those on which the network was trained.

Recently, neural networks have been applied to the assignment of specific protein folding patterns based on conserved amino acid sequences which characterize the fold. Hirst and Sternberg have trained a network to distinguish the ATP-binding motif based on a window of 17 residues around the consensus sequence of the phosphate binding motif. This perceptron network correctly classified 78% of the sequences tested, much better than a simple consensus sequence search, but roughly equivalent to the performance of a sophisticated statistical method.[8] Bengio and Pouliot have trained a network containing a hidden layer to identify immunoglobin domains from the amino acid sequences found in the four most conserved β-strands.[9] This network recognized 98% of the immunoglobin test sequences with only 7% false positives.

Several groups[10,11] have shown that not only sequence, but the overall amino acid composition (percent) is relevant to many structural features of pro-

Received August 21, 1992; revision accepted November 17, 1992.
Address reprint requests to Professor Sung-Hou Kim, Melvin Calvin Laboratory, University of California at Berkeley, Berkeley, CA 94720.

teins, including the folding type. Nishikawa and coauthors[12,13] proposed the universal classification of proteins into groups based on their amino acid composition and other characteristics and found a correlation of the amino acid composition of a protein to its structural and biological features.[14] Multiple linear regression was used to obtain a relationship useful for predicting the percentage of secondary structure types (helix, strand, or coil) of a protein from a knowledge of its amino acid composition.[15,16] Recently, Muskal and Kim have shown that neural networks can be used to predict with approximately 95% accuracy the secondary structure composition (percentages) from the amino acid composition and size of the protein as a whole.[17]

This finding led us to question whether information describing the overall three-dimensional fold of a protein may also be hidden in its amino acid composition, i.e., can we predict the core folding motif from the percentage composition of each amino acid and the total number of amino acids? A roadblock to answering this question was the limited number of known three-dimensional structures, as stored in the Brookhaven Protein Data Bank, for any single folding motif. We attempted to overcome this problem by making the assumption that the same protein (as defined by function and overall sequence homology) from different organisms will fold into the same general motif. Thus, we added to the structure database those sequences of the corresponding protein from different organisms as found in the Swiss-Prot sequence database. We then used neural networks trained on this database to predict the folding motif of proteins of known sequence, but unknown tertiary structure.

METHODS

Database

Protein structure information was from the Brookhaven Protein Data Bank (PDB)[18] and from publications describing protein structures not yet deposited in the Brookhaven PDB. Sequence information was from the Swiss-Prot[19] database (SP) Release 20. Since a large number of known examples are necessary in order to train a network to recognize proteins of unknown fold, we chose to begin our studies on four diverse folding patterns for which there exists a relatively large number of known structures and their sequence analogues. 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. It should be noted that these classes are defined only by the organization of secondary structure elements and not by their connectivity. Ribbon drawings of representative proteins from each of the four folding patterns are shown in Figure 1a–d.

For single domain proteins, only sequences for which the folding pattern of interest comprised more than 50% of the total sequence were used. For multidomain proteins, only the domain containing the pattern of interest was considered and used only if the folding pattern accounted for more than 50% of the residues of the domain. Assignment of protein domains was based on annotations in the PDB files and original papers describing the three-dimensional structures. Related proteins from the SP database were assigned the same domain structure. In the case of multidomain proteins of unknown structure, it may be possible to define the domain boundaries by information from biochemistry, genetics or molecular biology, such as partial digestion patterns, intron–exon boundaries, etc. Table I shows the known examples of each folding class selected from the PDB and published structure determinations, together with the number of homologous sequence relatives found in the SP database. It should be noted that there are large differences in the number of proteins in each database of folding motifs. We were especially restricted by the limited size of the set of BUNDLE proteins (84 examples). Besides the four datasets of defined folding patterns, we selected from the PDB and SP databases a set of "unclassified" (UNC) examples—proteins which did not belong to any of the four folding classes which we had chosen. These were then used as "false" or "other" classes in network training and testing.

Neural Networks

The neural networks used in this study were of the feedforward type with either zero (perceptron) or one hidden layers and weights adjusted by conjugate gradient minimization using the computer program package BIOPROP.[17] The architecture of the networks is shown in Figure 2. The protein sequences were converted to amino acid percentages and these together with the normalized total number of amino acids in the protein or domain were used as inputs of the neural network. Each of the four databases was separated into independent training and testing sets, i.e. no protein family from training was included in testing. 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. For optimizing the architecture, networks containing from 0 to 9 hidden nodes were systematically trained and tested numerous times in order to find the deepest minima. The architecture with best performance was chosen for further calculations. However, architectures with more than five hidden nodes contained more weights than examples in database (Table I) and were thus subject to memorization.[17] The networks with the optimal number of hidden nodes were then

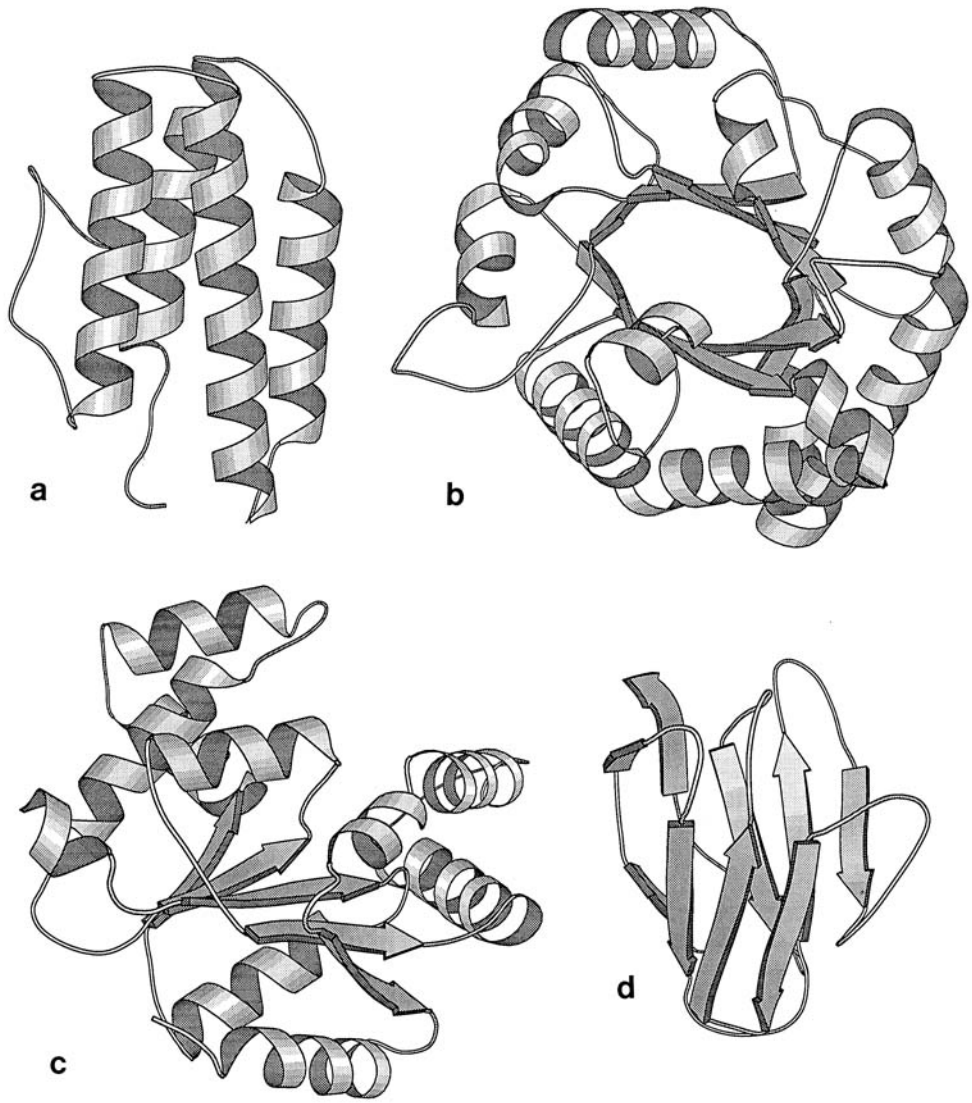

Fig. 1. Schematic (MOLSCRIPT)[39] drawings of members of the four folding patterns predicted. (**a**) Four α-helical bundle fold (BUNDLE): the specific example shown is Myohemerythrin (PDB code: 2MHR, 118 amino acids). (**b**) Eight-stranded parallel (α/β) barrel (BARREL): the specific example shown 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.

trained repeatedly (10 times) from random initial starting weights to search more completely the parameter space and the weights producing the highest testing scores chosen as the "optimal network weights." One hundred cycles of conjugate gradient minimization was found to provide the maximal generalization (best training without memorization) as judged by performance on independent testing sets, i.e., for most testing sets chosen, prediction accuracy was seen to degrade after this point. This number of iterations was therefore used for all network training.

Networks of from one to four outputs were used to classify a protein into one of the four folding patterns or as unclassified (UNC). 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 "UNC" (the fold of interest does not correspond to any of the available choices).

RESULTS

The PDB and SP databases and the four databases corresponding to 4α-helical bundles, eight stranded parallel α/β barrels, nucleotide binding fold, and immunoglobulin fold were analyzed in terms of their average amino acid composition. The average per-

TABLE I. Protein Database Used in This Study

Proteins	Number of residues	Number of examples from Swiss-Prot Database	Code in PDB or reference
With bundle folding			
Cytochrome b_{562}	103	1	256B
Cytochrome c'	128	12	2CCY
Myohemerythrin	118	1	2MHR
Hemerythrin	113	2	1HMQ
	113		1HRB
Somatotropin (growth hormone)		31	28
Ferritin (light chain)		5	29
Protein disk of tobacco mosaic virus	158	10	30
ColE1 ROP	63		31
Apolipoprotein E precursor (NH_2-terminal domain)	191	6	32
Tar(asp) receptor	232		33
Macrophage colony-stimulating factor		2	34
Granulocyte-macrophage colony-stimulating factor		3	35
Granulocyte colony-stimulating factor		2	
With barrel folding			
Tryptophane synthase (α-subunit)	270	13	1WSY
Aldolase	363	15	0ALD
D-Xylose isomerase	394	9	1XIA
Taka-amylase, main domain	247		2TAA
Pyruvate kinase, domain A	220	11	1PYK
Ribulose bisphosphate carboxylase (active site)	251	24	2RUB
2-Hydroxy-acid-oxidase	359		1GOX
Bifunctional enzyme *N*-(5′-phosphoribosyl)antranilate isomerase-indole-3-glycerol-phosphate synthase	~450	15	36
(FT-domain)	~270		
Triosephosphate isomerase	248	12	1TIM
Mandelate racemase (central domain)	182		37
Muconate cycloisomerase (central domain)	194	3	1MLE
With nucleotide binding fold domains			
NAD-binding proteins			
Lactate dehydrogenase	331	24	2LDX
	297		1LDB
	320		1LLC
	329		5LDH
Alcohol dehydrogenase	374	52	5ADH
Glyceraldehyde-3-phosphate dehydrogenase	334	40	1GD1
	333		1GPD
Adenylate kinase	194	11	3ADK
GTP superfamily			
Products of *ras* oncogene	171	25	2P21
			38
GTPases used in ribosomal synthesis	190	13	1EFM
			38
With immunoglobulin fold domains			
IG γ chains	330	16	1FC1
IG heavy chains	126	25	2FB4
	119		2FBJ
	222		1MCP
IG λ chains	105	25	2MCG
	117		3FAB
IG κ chains	108	25	1REI
Myelin-associated glycoprotein		4	
Histocompatibility antigen			
Class 1	365	30	1HLA
Class 2		30	
T-cell surface glycoprotein	458	30	1CD4

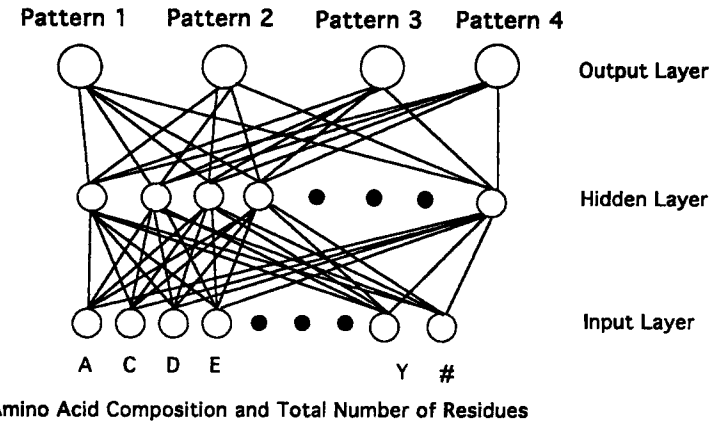

Fig. 2. Schematic diagram of the architecture of the computational neural networks discussed in the text. The circles represent computational nodes and the lines correspond to weighted links between the nodes. The input level nodes supply the percent composition of each amino acid to the network. The hidden level nodes sum the inputs from the input layer times the weights and send a signal based on a sigmoidal response to this summation. These signals multiplied by the weights linking the hidden and output nodes are received by the output layer, summed, and the most active node is identified.

cent composition and standard deviation of each of the 20 amino acids are shown in Table II. It is apparent that no single amino acid percentage for any of the four motif classes is significantly different than that found in the overall databases.

The results of network training for each of the four motif classes with from one to four output nodes are summarized in Tables III–VI. Table III gives the performance for prediction of structural class using a single output node. Thus, there are four trained networks, each making a prediction as to whether an unknown protein contains a specific folding pattern. Three separate training/testing sets were considered for BUNDLE proteins. Two of these training/testing sets were similar in prediction, but the third, in which the family of cytochrome c' proteins was separated for testing, strongly overpredicted these proteins as "UNC," i.e., as not containing a BUNDLE motif. For this particular case a 22nd input node—heme presence/absence turned out to be very important. This input was also found to be important in the prediction of secondary structure composition by Muskal and Kim.[17] After the addition of this characteristic to inputs the performance for cytochrome c' family predictions increased to 80% for "true" examples. At the same time the results of prediction of other training/testing sets didn't become worse. Performance of single output node networks for other folding patterns was consistent, regardless of how the training/testing sets were constructed. The correlation coefficients listed in this and other tables are a better measure than percent correct prediction, since they take into account the bias of over- or underprediction.

Since it has been demonstrated[17] 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. We tested this by constructing a test set of proteins of high α-helix (> 50% helix) and low β-strand composition and using the network trained on 4α-helical bundles to predict whether these proteins belong to that folding pattern or not. Of 19 proteins which did not contain the BUNDLE fold, 14 (74%) were correctly predicted to not belong to this class, a clear indication that the networks are recognizing information specific to the folding patterns and not the secondary structure composition.

Table IV lists the results of prediction of folding pattern with 2 output nodes. There are six networks representing all the pairwise combinations of folding choices. Two separate training/testing sets were constructed for each network. One of these presents mixtures of approximately equal amounts of the two database representatives and UNC examples, while the other contains only mixtures of "true" examples and does not contain UNC examples; that is, the first type of network distinguishes one pattern from another and also from other proteins, while the second network distinguishes only between the two patterns of interest. From the data of Table III it is evident that the performance of the networks is dependent on the presence of UNC examples in the testing/training set. The correlation coefficient and performance are systematically higher for sets without UNC examples. One additional difference between the two types of sets is the number of hidden nodes in the trained network. To get high performance for sets with UNC examples we needed 5 hidden nodes except for the cases of pairs of classes which differ significantly in structure (BUNDLE and NBF, BARREL and IGF). In these cases and for sets without UNC examples, 3 hidden nodes turned out to be sufficient. Additional hidden nodes did not improve the correlation coefficients and the perfor-

TABLE II. Average Amino Acid and Secondary Structure Compositions and Standard Deviations for Different Protein Classes (%)

	PDB*	SP seq. Database†	Bundle‡	Barrel‡	NBF‡	IGF‡
Number of examples	259	1900	84	126	178	260
Residue						
Ala	8.3/2.7	8.2/3.0	9.6/5.9	10.6/2.0	8.9/2.9	6.4/2.3
Arg	3.6/1.9	4.5/1.9	5.3/2.6	5.3/1.4	4.0/1.6	4.8/2.0
Asn	4.8/1.7	3.7/1.6	4.2/2.3	3.3/1.1	4.3/1.6	3.7/1.6
Asp	5.7/2.0	4.7/1.7	5.7/2.1	5.6/1.4	6.1/1.5	4.3/1.6
Cys	2.9/2.6	1.6/1.2	1.7/1.0	1.3/0.7	1.7/1.2	2.5/1.4
Gln	3.7/1.5	3.6/1.6	5.7/2.2	3.4/1.2	3.1/1.8	4.7/1.9
Glu	5.3/2.3	5.6/2.0	6.8/2.8	7.0/1.2	5.5/1.9	5.5/2.1
Gly	8.6/2.4	7.0/2.4	5.3/2.2	8.9/1.6	9.1/2.8	7.0/2.6
His	2.1/1.2	1.9/1.0	1.9/1.3	2.4/0.9	1.8/1.1	2.0/1.2
Ile	5.0/1.9	5.3/2.0	3.8/1.7	5.5/1.1	6.8/1.7	4.3/1.7
Leu	7.7/2.3	8.8/2.4	11.5/4.0	8.8/1.4	7.4/2.5	10.1/2.9
Lys	6.2/2.5	5.3/2.2	5.9/3.3	5.6/1.7	7.4/1.8	5.3/2.4
Met	1.8/1.0	2.4/1.2	2.3/1.1	2.0/0.9	2.4/1.1	2.0/1.0
Phe	3.5/1.4	3.7/1.5	4.7/1.7	3.7/1.1	3.9/1.3	3.8/1.7
Pro	4.7/1.6	4.9/2.1	3.7/1.6	4.5/1.0	3.4/1.4	5.3/2.0
Ser	7.3/2.8	6.2/2.0	6.9/2.9	4.9/1.4	5.7/1.8	8.8/3.4
Thr	6.3/2.0	5.3/1.8	5.4/2.5	5.5/1.2	5.9/1.7	7.1/2.0
Trp	1.5/0.9	1.3/0.9	1.2/0.9	1.1/0.8	0.8/0.5	1.9/0.9
Tyr	3.8/1.7	2.9/1.3	3.0/1.2	3.1/1.1	2.7/1.4	3.6/1.7
Val	7.1/2.0	6.3/1.9	5.5/2.2	7.5/1.8	9.8/2.7	6.7/2.0
α-Helices§			70.3/2.9	45.3/5.7	36.3/10.1	7.1/5.1
β-Strands§			28.7/2.9	15.3/5.7	16.1/4.5	44.1/4.7

*All nonidentical proteins.
†Randomly selected.
‡Set of proteins used in this study.
§Only for PDB examples of each class.

mance. The percent of correct predictions for UNC examples is generally lower than for other members of testing set. Overprediction of UNC is apparent in case 3-1. This overprediction is accompanied by significant underprediction of main class representatives. The opposite situation is observed in case 1-3, where unclassified cases are underpredicted and proteins from classes B (fold 2) and A (fold 2), respectively, are overpredicted.

Results of the networks with 3 output nodes are shown in Table V. Four network types were trained, one for each combination of protein classes. The optimal number of hidden nodes was five in each case. Two separate training/testing sets were explored for each type of network, one containing UNC proteins and one including only examples belonging to the folding patterns to be distinguished. The results are similar to those obtained with networks having two output nodes. Again, we observe a more unreliable prediction when UNC examples, i.e., those not belonging to either class, are present (see cases 3-1 and 4-1).

Finally, we list the results of a network with four outputs, one for each protein class, in Table VI. Two different training/testing sets are shown, one including UNC examples. The problem of over/under prediction between classes is apparent for the testing set which includes these proteins.

Analysis of Network Weights

The weights associated with the links between nodes in a neural network are the variable parameters incorporating the information, extracted from the training examples, which the network uses to make a decision. In an effort to understand the basis of neural network decisions as to protein folding pattern, we trained perceptron (no hidden layer) networks for each of the four classes of fold with one output node (Table III shows these nets *with* a hidden layer). All members of the folding class were included in training, i.e. no testing class was separated out. Figure 3a–d shows the magnitude of these weights for each amino acid connection to the output node. Comparison of these weights with the statistical distribution of amino acids shown in Table II indicates that the larger weights are generally associated with amino acids which are overabundant in that class (albeit not in a statistically significant way) and that the most negative weights correspond to residues that are underrepresented. A combination of these relatively small differences in amino acid composition appears, therefore, to provide a

TABLE III. Performance of Independent Testing Set for 1 Output Node (Input Nodes = 21, Hidden Nodes = 3, Threshold = 0.5)

Folding class* A	Number of examples in training/ testing set	Correlation coefficient†		Testing examples/ predicted examples/ predicted correct‡		Percent of correct predictions§		Testing set**
		Training set	Testing set	A	UNC	A	UNC	
1††	122/14	.96	.99	7/7/7	7/7/7	100	100	1-1
	124/16	.88	.78	8/8/7	8/8/7	87.5	87.5	1-2
	114/26	.91	.50	13/10/8	13/16/11	80.0	68.8	1-3
2‡‡	124/32	.83	.80	16/13/13	16/19/16	100	81.2	2-1
	126/14	.83	.94	7/6/6	7/8/7	100	87.5	2-2
3	317/30	.96	.68	15/13/13	15/17/15	100	88.2	3-1
	320/55	.87	.97	27/25/25	28/30/28	100	93.3	3-2
4	174/76	.95	.50	38/31/25	38/45/32	80.6	71.1	4-1
	174/76	.93	.60	38/36/29	38/40/31	80.6	77.5	4-2
	196/11	.88	.68	57/61/49	57/53/45	80.3	84.9	4-3

*Classes of proteins: 1, BUNDLE; 2, BARREL; 3, NBF; 4, IGF.

†Correlation coefficient defined as $r = (N\Sigma x_i y_i - \Sigma x_i \Sigma y_i)/(\sqrt{N\Sigma x_i^2 - (\Sigma x_i)^2} * \sqrt{N\Sigma y_i^2 - (\Sigma y_i)^2})$ where x_i is the observed value of the output node in the training or testing set (0.0 for UNC, 1.0 for a member of A) and y_i is the activity at the output node calculated by the trained network.

‡Half the testing examples were members of the folding class (A) and half were other unclassified folds (UNC).

§Percent of the *predicted examples* which are correct for classified (A) and unclassified (UNC) folds.

**Group of proteins chosen for testing set, all sets include UNC folds as false examples: 1-1, M-CSF, GM-CSF, G-CSF; 1-2, Tar(Asp) receptor, cytochrome b_{562}, hemerythrin; 1-3, cytochrome c'; 2-1, aldolase; 2-2, indole-3-glycerol phosphate synthase; 3-1, ras-related proteins RAB; 3-2, lactate dehydrogenase; 4-1, T-cell surface glycoprotein; 4-2, histocompatibility antigen; 4-3, IG chains.

††22 input nodes.

‡‡Hidden nodes = 5.

strong predictor of folding pattern. Sorting the weights by known properties of the amino acids themselves gives further insight into the composition requirements of each protein fold (Fig. 3). For example, the networks weights for BUNDLE proteins sorted by helix propensity of the associated amino acids show a weak general trend of increasing with increased helical propensity; suggesting that helix bias is a contributing, but not the primary factor in formation of a BUNDLE fold. The BARREL folding pattern is generally favored (high weights) by hydrophobic residues and disfavored by polar residues. In contrast, IGF folding is generally compatible with hydrophilic residues and incompatible with hydrophobic residues. For the NBF fold a fairly strong correlation (0.3) is observed between network weights and a sum of helix and strand propensity.

Clearly, networks with hidden nodes are able to improve on this by forming combinations of amino acid compositions for each protein and then using combinations of these combinations to make the prediction. We have examined the weights from single output networks (Table III) including a hidden layer of three nodes. Generally, the same tendencies are observed as found for the perceptron networks, i.e., amino acid types which have positive weights in the perceptrons also make positive contributions to the hidden layer networks and vice versa for amino acid types making negative contributions. However, some exceptions are observed. For instance, in the perceptron networks trained on the IGF fold, histidine is strongly favorable (a high positive weight) and valine unfavorable (a negative weight). In the hidden node network trained on the IGF fold, histidine makes a negative contribution and valine a positive contribution.

Since the magnitude of the weights and the relative over- or underrepresentation of an amino acid are correlated for the folding classes, the implication is that these amino acids are performing specific structural roles which may stabilize the various folding patterns. Although the preference of particular folds for certain types of amino acids, such as hydrophobic ones or those with high helical propensity, is clearly understandable, one can also ask the question: Are intramolecular contacts between specific amino acids preferentially observed for each folding type and if so, do they play a unique role in structural stabilization? To answer this question we calculated the non-sequential (contacting residues must be separated by at least 3 other residues) intramolecular contacts for members of each folding type and classified them by amino acid type. Amino acid residues were considered to be in contact if the distance between any two atoms of the residues was less than 4.0 Å. No correction was made for residue size. Polar residues are less likely to appear in such a list since they are at the surface contacting water or possibly adjacent residues.

For the BARRELS, we computed intramolecular

TABLE IV. Performance of Independent Testing Set for 2 Output Nodes (Input nodes = 21)

Folding classes*		Examples in training		Correlation coefficient training set		Correlation coefficient testing set			Taken examples/ predicted examples/ predicted correct§			Percent of correct predictions			Testing
A	B	testing set	H†	A	B	A	B	T‡	A	B	UNC	A	B	UNC	set**
1	2	184/24	5	.87	.89	.97	.68	.3	8/8/8	8/12/8	8/4/4	100	66.7	100	1-3
		120/16	3	.97	.97	.88	.88	.5	8/7/7	8/9/8	—	100	88.9	—	1-4
1	3	200/28	3	.88	.94	.59	.58	.4	8/7/5	10/7/6	10/14/10	71.4	85.7	71.4	2-3
		134/18	3	.99	.99	.55	.56	.5	8/10/6	10/6/6	—	60.0	100	—	2-4
1	4	170/30	5	.95	.98	.63	.56	.5	8/5/5	12/6/5	10/19/8	100	83.3	42.1	3-1
		151/20	3	.99	.99	.69	.69	.5	8/4/4	12/16/12	—	100	75.0	—	3-2
2	3	219/44	3	.82	.88	.79	.84	.4	16/21/15	12/9/9	16/17/10	71.4	100	58.8	4-1
		149/28	3	.96	.96	.92	.92	.4	16/13/13	12/11/11	—	100	100	—	4-2
2	4	210/50	3	.77	.88	.87	.96	.5	16/16/14	18/18/18	16/16/14	87.5	100	87.5	5-3
		149/34	3	.99	.99	.96	.96	.5	16/17/16	18/18/18	—	94.1	100	—	5-4
3	4	266/40	5	.86	.90	.95	.79	.4	15/12/10	13/12/10	12/16/8	83.3	83.3	50.0	6-3
		178/28	3	.98	.98	.97	.96	.4	15/15/15	13/12/12	—	100	100	—	6-4

*Classes of proteins: 1, BUNDLE; 2, BARREL; 3, NBF; 4, IGF.
†H, number of hidden nodes.
‡T, threshold value.
§One-third of the testing examples were members of the folding class A, one-third B, and one-third were other unclassified folds (UNC).
**Group of proteins chosen for testing set: 1-3, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (D-xylose isomerase) + unclassified; 1-4, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (D-xylose isomerase); 2-3, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (lactate dehydrogenase) + unclassified; 2-4, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (lactate dehydrogenase); 3-1, (Tar(Asp) receptor, M-CSF, GM-CSF, G-CSF) + (T-cell surface glycoprotein) + unclassified; 3-2, (Tar(Asp) receptor, M-CSF, GM-CSF, G-CSF) + (T-cell surface glycoprotein); 4-1, (aldolase) + (RAB protein) + unclassified; 4-2, (aldolase) + (RAB protein); 5-3, (aldolase) + (T-cell surface glycoprotein) + unclassified; 5-4, (aldolase) + (T-cell surface glycoprotein); 6-3, (lactate dehydrogenase) + (histocompatibility antigen, class I) + unclassified; 6-4, (lactate dehydrogenase) + (histocompatibility antigen, class I).

contacts of triose phosphate isomerase (1TIM) and a bifunctional enzyme 2-hydroxy-acid-oxidase (1GOX). The pattern of these contacts was very similar between these two proteins. Alanine, isoleucine, leucine, and valine make the most contacts with other amino acids, with most contacts of the following types A-I, A-L, A-V, V-I, V-L. The number of such contacts is much greater in BARREL proteins than in proteins of different fold, but similar size and composition of alanine, isoleucine, valine, and leucine. These types of contacts form part of the interface surface between the β-strands and α-helices. Further examination of these contact sites reveals a pattern in which alanine or valine extending out from the strands of the beta barrel are interacting with leucines or isoleucines pointing in from the helices. Each of these amino acids are assigned a positive (favorable) weight in the prediction of the BARREL folding class in both perception (Fig. 3) and hidden node networks.

As examples of proteins containing the NBF fold we examined the intramolecular contacts of adenylate kinase (1ADK) and glyceraldehyde-3-phosphate dehydrogenase (1GD1). In the NBF class, as in the related BARREL class, Ile and Val residues are assigned a high positive weight in the neural networks. These amino acids also are involved in the most intramolecular contacts. As in the BARREL class, these amino acids are primarily at the interface between the helices and strands with most valines pointing out from the strands and isoleucines pointing in from the helices.

In IGF proteins, analysis of 1REI, 1MCP (heavy chain), and 1MCP (light chain) (Table I) contacts show that serine and threonine are involved in numerous intramolecular contacts, including a significant number of contacts between themselves. Serine and threonine are also assigned highly favorable weights both by perceptron and hidden layer networks. Examination of these residues by computer graphics reveals that serine and threonine are often located across from each other in the beta strands or at the end of turns and make interresidue hydrogen bonds either between the sidechain hydroxyls located on the same face of the beta sheet or one hydroxyl and the carbonyl oxygen of the opposite residue. In some cases a simple rotation about the Cα–Cβ bond would serve to orient the side chains in hydrogen bonding position.

The BUNDLE proteins do not present a clear pattern of intramolecular contacts for the amino acids with high weights. This may be due to the structural diversity of the proteins having this fold, i.e., the proteins containing heme cofactors and those with

TABLE V. Performance of Independent Testing Set for 3 Output Nodes (Input Nodes = 21, Hidden Nodes = 5, Threshold = 0.5)

Folding classes*			Number of examples in training/ testing set	Correlation coefficient						Taken examples/ predicted examples/ predicted correct†				Percent of correct predictions			Testing set‡
				Training set			Testing set										
A	B	C		A	B	C	A	B	C	A	B	C	UNC	A	B	C	
1	2	3	239/51	.75	.73	.67	.88	.45	.29	13/7/6	11/21/9	12/5/3	15/18/12	85.7	42.9	60.0	1-1
			179/36	.96	.83	.81	.82	.61	.87	13/8/8	11/16/10	12/12/11	—	100	62.5	91.7	1-2
1	3	4	265/43	.87	.71	.80	.80	.84	.54	8/7/6	12/14/12	13/4/3	10/18/7	85.7	85.7	75.0	2-1
			205/33	.95	.91	.94	.89	.89	.80	8/9/8	12/14/12	13/8/8	—	72.7	85.7	100	2-2
2	3	4	274/49	.63	.48	.90	.47	.35	.77	11/25/11	12/7/2	13/14/12	13/3/3	44.0	28.6	85.7§	3-1
			209/36	.78	.73	.92	.77	.72	.94	11/6/6	12/17/12	13/10/10	—	100	70.6	100	3-2
1	2	4	270/40	.78	.76	.86	.51	.62	.63	8/12/8	11/13/10	11/8/7	10/7/3	66.7	76.9	87.5**	4-1
			205/30	.92	.97	.94	.58	.96	.70	8/9/8	11/10/10	11/7/7	—	88.9	100	100**	4-2

*Classes of proteins: 1, BUNDLE; 2, BARREL; 3, NBF; 4, IGF.
†Approximately one-fourth of the testing examples were members of the folding class A, one-fourth B, one-fourth C, and one-fourth were other unclassified folds (UNC).
‡Proteins chosen for testing set: 1-1, (cytochrome *c'*) + (aldolase) + (glyceraldehyde-3-phosphate dehydrogenase) + unclassified; 1-2, (cytochrome *c'*) + (aldolase) + (glyceraldehyde-3-phosphate dehydrogenase); 2-1, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (glyceraldehyde-3-phosphate dehydrogenase) + (T-cell surface glycoprotein) + unclassified; 2-2, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (glyceraldehyde-3-phosphate dehydrogenase) + (T-cell surface glycoprotein); 3-1, (aldolase) + (glyceraldehyde-3-phosphate dehydrogenase) + (T-cell surface glycoprotein) + unclassified; 3-2, (aldolase) + (glyceraldehyde-3-phosphate dehydrogenase) + (T-cell surface glycoprotein); 4-1, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (aldolase) + (T-cell surface glycoprotein) + unclassified; 4-2, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (aldolase) + (T-cell surface glycoprotein).
§Threshold = 0.3.
**Threshold = 0.4.

TABLE VI. Performance of Independent Testing Set for 4 Output Nodes (Hidden Nodes = 5, Threshold = 0.4)*

Number of examples in training/ testing set	Correlation coefficient								Taken examples/ predicted examples/ predicted correct†					Percent of correct predictions				Testing set‡
	Training set				Testing set													
	1	2	3	4	1	2	3	4	1	2	3	4	UNC	1	2	3	4	
353/53	.87	.88	.91	.92	.70	.65	.65	.66	8/6/5	8/20/8	12/15/9	13/12/10	12/2/0	83.3	40.0	60.0	83.3	1-1
290/41	.89	.79	.87	.86	.56	.57	.60	.65	8/5/5	8/14/8	12/11/9	13/9/9	—	100	57.1	81.8	100	1-2

*Classes of proteins: 1, BUNDLE; 2, BARREL; 3, NBF; 4, IGF.
†Approximately one-fifth of the testing examples were members of the folding class A, one-fifth B, one-fifth C, one-fifth D, and one-fifth were other unclassified folds (UNC).
‡Proteins chosen for testing set: 1-1, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (D-xylose isomerase) + (lactate dehydrogenase) + (histocompatibility antigen, class I) + unclassified; 1-2, (Tar(Asp) receptor, cytochrome b_{562}, hemerythrin) + (D-xylose isomerase) + (lactate dehydrogenase) + (histocompatibility antigen, class I).

nonheme irons utilize histidine residues for cofactor binding which probably is responsible for the high weight of this amino acid in the neural networks. In some members of this class the highly favorable amino acid phenylalanine forms stacks within the hydrophobic core in the center of the bundle. The large negative weight for glycine likely arises from its low helical propensity. Interestingly, proline is not unfavorable for BUNDLE formation.

DISCUSSION

We have shown by extensive testing that computational neural networks trained on the amino acid composition and size of proteins known to belong to certain folding classes are able to predict whether an unknown protein belongs to one of these classes. These predictions are restricted to classes for which enough members are known to have that structure to efficiently train the networks. Also, the folding pattern must account for at least 50% of the residues in the protein or domain. The size of the folding motifs we have considered thus restricts the size of the proteins which are candidates for a structural prediction to a minimum of about 100 amino acids and to less than 400 amino acids unless information concerning the domain structure is available. Even in proteins of this size range knowledge of the domain structure is critical to application of this method.

The finding that the three-dimensional folding type of a protein can be accurately predicted from the amino acid composition of the protein implies not only that the percentages of amino acids deter-

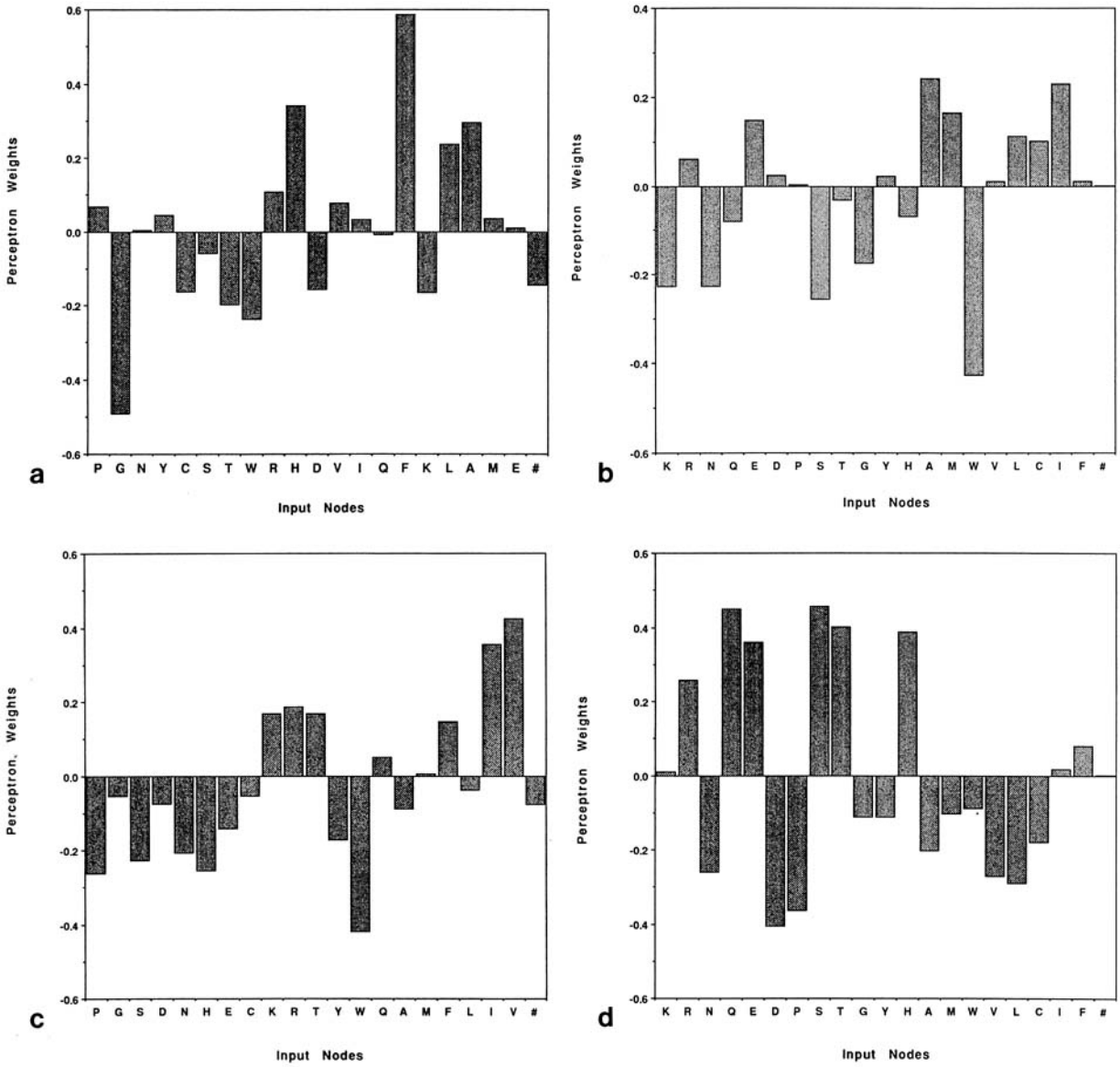

Fig. 3. Perceptron weights of single output networks trained for each of the four folding patterns. (**a**) Four α-helical bundles, weights for each amino acid type are sorted by helix propensity; (**b**) eight-stranded parallel (α/β) barrels, weights for each amino acid type are sorted by hydrophobicity; (**c**) nucleotide binding fold, weights for each amino acid type are sorted by a combination of their helix and strand propensity; (**d**) immunoglobin fold, weights for each amino acid type are sorted by hydrophobicity. The one letter code for the 20 amino acids whose compositions are supplied as input and the number of amino acids in the protein/domain (#) are listed along the horizontal axis. The weight determined for each node is shown on the vertical axis. These weights are multiplied by the input values and summed to give the output activity. A higher weight indicates a stronger contribution to prediction of that fold and a more negative weight indicates opposition to prediction of the fold.

mine the secondary structure content,[17] but that tertiary interactions requiring specific intramolecular contacts also are favored by specific patterns of amino acid content. For example, BARRELS generally have a higher proportion of hydrophobic amino acids (Table II) since they must form a stable hydrophobic interface between the helices and β-barrel. The specific amino acids which receive high weights in the perceptron networks and networks with a hidden layer are those which also show the greatest number of intramolecular contacts.

Analysis of the amino acid content of the complete sequence databases and that of the protein folding classes shows that differences in the compositions of *individual* amino acids are not statistically significant (Table II). However, the *overall* distribution of amino acids clearly is dependent on folding class. The most abundant amino acids[10] for α-helical proteins—alanine, lysine, and leucine (11.6, 10.1, and 8.9%), are also abundant in BUNDLE proteins (Ala = 10.5%, Lys = 6.6%, and Leu = 12.6%), however, leucine is now the most frequent and lysine is much

TABLE VII. Network Prediction of Proteins From Four Folding Classes*

	Bundle	Barrel	NBF	IGF	UNC
Single output	I5†	A5, H	I5†, H	T	
Two outputs					
Bundle–barrel	I5	A5, T, H			
Bundle–NBF	I5		A3, H		A2, T
Bundle–IGF	I5			A1	A4, T, H
Barrel–NBF		A5, T, H	I5		
Barrel–IGF		A5, H		I5	T
NBF–IGF			H	I5, T	A5
Three outputs					
Bundle–barrel–NBF	I5	A5, H	—		
Bundle–NBF–IGF	I5		A5, H	T	
Barrel–NBF–IGF		A5, H	I5	T	
Bundle–barrel–IGF	I5	A5, H		T	
Four outputs					
Bundle–barrel–NBF–IGF	—	A5, H	—	T	

*A, Prediction of the folding pattern of 5 members of the aldose reductase family, an example of the BARREL fold. H, Prediction of the folding pattern of the heat shock 70-related protein 1 (HSP70) from mitochondria, an example of the NBF fold. I, Prediction of the folding pattern of 5 members of the interleukin 2 family, an example of the BUNDLE fold. T, Prediction of the folding pattern of the T-cell receptor β-chain precursor V region, an example of the IGF fold. The number following the one letter code indicates how many members of the family were predicted for that category (T only contains one member). All sequences are from Swiss-Prot Release 21 A threshold 0.4 in output activity was used († indicates that the proteins tested above threshold in two classes).

less common. The most common amino acids for β-proteins,[10] Gly (9.9%) and Ser (9.5%), are not so abundant in the immunoglobulin family where leucine is the most common amino acid (Table II). Our results indicate that these small differences, although each statistically insignificant, can serve to distinguish specific folding types from both members of the same class (i.e., BUNDLES from all α proteins) and from other folding types generally. Neural networks can "learn" *combinations* of small systematic deviations in each class of examples and use that knowledge to predict the folding motif of an unknown protein.

How can we best use the trained neural networks described above to make predictions of protein tertiary folding motif? As a test, we have chosen four proteins, belonging to the various folding classes from the SWISS-PROT sequence database and predicted their fold using each of the types of neural networks we have trained. For instance, the enzyme aldose reductase for which the three-dimensional structure was recently determined by X-ray crystallography[20] is an example of BARREL fold. Predictions were made for each of these proteins using networks with from 1 to 4 output nodes. The results are summarized in Table VII. In general, each protein is predicted as belonging to its correct folding class. It is clear from the table that in order to be confident about a prediction, there should be a consensus among the networks. For example, the interleukin 2 cytokine[21] is predicted as both BUNDLE and NBF by the single output networks, but is clearly BUNDLE when all networks are considered. Also, when BUNDLE is not an option such as in the two output networks, BARREL-IGF and NBF-IGF, interleukin 2 is not predicted as unclassified (UNC), but rather as IGF. The bovine heat shock cognate protein (HSC70) ATPase fragment[22] is almost equally predicted as BARREL and NBF. These two classes often give ambiguous predictions, likely due to the similarities in their folding. In a case such as this the prediction would be that it belongs to one of these two classes. Once a prediction has been made, it may be tested by other empirical methods which utilize the sequence itself such as secondary structure prediction[1,23,24] and profile analysis.[25]

As a test of our procedure we predicted the folding class of a variety of proteins of unknown structure. A few of these gave strong indications of belonging to one of the four classes for which we have implemented networks. A specific example is the tumor suppressor p53 protein.[26] The human p53 protein is composed of 393 residues which may be subdivided into three putative domains: an N-terminal, acidic and highly charged α-helical domain of about 70 residues, a C-terminal basic domain, containing DNA binding motifs of about 75 residues, and a central domain. We have excluded residues in both the acidic (N-terminal) and basic (C-terminal) domains in calculation of the amino acid composition of the central domain. All networks, when tested on the amino acid composition of the central domain of p53 predict it to belong to the immunoglobin folding class (IGF) as shown in Table VIII. These predictions for p53 are consistent with secondary structure pre-

TABLE VIII. Prediction of the Folding Pattern of the p53 (Central Domain) Tumor Suppressor Protein

	Bundle	Barrel	NBF	IGF	UNC
Single output				p53	
Two outputs					
Bundle–barrel	—	—			p53
Bundle–NBF	—		—		p53
Bundle–IGF	—			p53	—
Barrel–NBF		—	—		p53
Barrel–IGF		—		p53	
NBF–IGF			—	p53	
Three outputs					
Bundle–barrel–NBF	—	p53	—		
Bundle–NBF–IGF	—		—	p53	
Barrel–NBF–IGF		—	—	p53	
Bundle–barrel–IGF	—			p53	
Four outputs					
Bundle–barrel–NBF–IGF	—	—	—	p53	

diction[1] which indicates 8–9 β-strands in this domain.

There are several areas which may immediately benefit from the ability to predict the folding motif of a protein from its amino acid composition. First, an improvement in secondary structure prediction can be expected if the protein fold is known. This is because the number, length, and organization of structural elements within the various folding classes is more or less known. Although there can be variation in the lengths of the helices in BUNDLE proteins within certain limits; the length of the β-strands in IGF proteins and the lengths of the helices and strands in NBF motifs and BARRELS are fairly constant. Langridge and co-workers[27] have shown that knowledge of the tertiary structural class (all α, all β, α/β) of a protein enables a significant increase in accuracy of secondary structure prediction. An accurate secondary structure prediction coupled with the three-dimensional topology specified by the predicted folding pattern may allow construction of actual three-dimensional models of proteins from their amino acid sequences. Once the secondary structure elements have been fit into the predicted folding motif, insertions or deletions in variable loop regions connecting the secondary structural elements may be constructed by standard molecular modeling techniques and the entire structure refined by molecular dynamics and energy minimization.

Another area in which knowledge of folding pattern could be of great value is in the determination of new protein structures by X-ray crystallography using the method of molecular replacement. In this method a search or probe molecule which is expected to have similar structure to the unknown protein is rotated and translated throughout the crystallographic asymmetric unit until the proper position is located, phases are then calculated from the properly positioned probe for the diffracted intensities and used to make an electron density map of the new protein. A model of the new protein is then built to this map. Typically, probe molecules are determined by sequence homology to the protein being studied crystallographically. In the absence of such sequence homology, this method has not been practical. Prediction of the folding pattern from the amino acid composition would enable the crystallographer to use members of this class as search probes in molecular replacement.

Finally, a predicted folding motif may provide one with insight into the biological functions of an uncharacterized protein sequence. For example, proteins containing the Rossmann fold may be expected to bind nucleotides, a property which could easily be checked and specifically defined biochemically. At least it will be possible to rule out certain functions which are associated only with specific folds, for example, the DNA binding proteins containing helix–loop–helix elements do not belong to any of the classes which are discussed in this manuscript.

Even though the four folding classes which we have investigated represent only an incomplete set of all possible protein folds, it should be clear that as more examples of other folding patterns are accumulated by X-ray crystallographic or NMR methods, it will be straightforward to include them in our predictions.

ACKNOWLEDGMENTS

The authors are grateful to Dr. Steven M. Muskal for providing the BIOPROP neural network software and to Dr. P.J. Kraulis for providing the MOLSCRIPT graphics software. We acknowledge the support of the Health Effects Research Division, Health and Environmental Research, Office of Energy Research of the U.S. Department of Energy.

NOTE ADDED IN PROOF

A computer program, incorporating the neural networks discussed in this manuscript to make predictions of protein folding pattern from sequence, is now available. This program, PROBE, also uses neural networks to predict other features of protein structure. For further information contact the authors by mail or electronically at either ILDUBCHAK@LBL.BITNET or SRHOLBROOK@LBL.BITNET.

REFERENCES

1. Quan, N., Sejnowski, T.J. Predicting the secondary structure of globular proteins using neural network models. J.Mol.Biol. 202:865–884, 1988.
2. Holley, L.H., Karplus, M. Protein secondary structure prediction with a neural network. Proc. Natl. Acad.Sci. U.S.A. 86:152–156, 1989.
3. McGregor, M.J., Flores, T.P., Sternberg, M.J. Prediction of beta-turns in proteins using neural networks. Protein Eng. 2(7):521–526, 1989.
4. Holbrook, S.R., Muskal, S.M., Kim, S.-H. Predicting surface exposure of amino acids from protein sequence. Prot. Eng. 3(8):659–665, 1990.
5. Muskal, S.M., Holbrook, S.R., Kim, S.-H. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng. 3(8):667–672, 1990.
6. Bohr, H., Bohr, J., Brunak, S., Cotterill, R.M.J., Fredholm, H., Lautrup, B., Petersen, S.B. A novel approach to prediction of the 3-dimensional structures of protein backbones by neural network. FEBS Lett. 261(1):43–46, 1990.
7. Friedrichs, M.S., Goldstein, R.A., Wolynes, P.G. Generalized protein tertiary structure recognition using associative memory hamiltonians. J. Mol. Biol. 222:1013–1034, 1991.
8. Hirst, J.D., Sternberg, M.J.E. Prediction of ATP-binding motifs: A comparison of a perceptron-type neural network and a consensus sequence method. Prot. Eng. 4:615–623, 1991.
9. Bengio, Y., Pouliot, Y. Efficient recognition of immunoglobulin domains from amino acid sequences using a neural network. CABIOS 6(4):319–324, 1990.
10. Nakashima, H., Nishikawa, K., Ooi, T. The folding type of a protein is relevant to the amino acid composition. J. Biochem. 99:153–162, 1986.
11. Chou, P.Y. In "Prediction of Protein Structure and the Principles of Protein Conformation." Fasman, G.D. (ed.). New York: Plenum Press, 1989:549–586.
12. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. I. Angular distribution. J. Biochem. 94: 981–985, 1983.
13. Nishikawa, K., Kubota, Y., Ooi, T. Classification of proteins into groups based on amino acid composition and other characters. J.Biochem. 94:997–1007, 1983.
14. Nishikawa, K., Ooi, K. Correlation of the amino acid composition of a protein to its structural and biological character. J. Biochem. 91:1821–1824, 1982.
15. Davies, D. A correlation between amino acid composition and protein structure. J.Mol.Biol. 9:605–609, 1964.
16. Krigbaum, W.R., Knutton, P. Prediction of the amount of secondary structure in a globular proteins from its amino acid composition. Proc. Natl. Acad. Sci. U.S.A. 70(10): 2809–2813, 1973.
17. Muskal, S.M., Kim, S.-H. Predicting protein secondary structure content: a tandem neural network approach. J. Mol. Biol. 225:713–727, 1992.
18. Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F.J., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. The Protein Data Bank: A computer-base archival file for macromolecular structures. J. Mol. Biol. 112:535–542, 1977.
19. Bairoch, A., Boeckmann, B. The SWISS-PROT sequence data bank. Nucl. Acids. Res. 19: Suppl.:2247–2249, 1991.
20. Wilson, D.K., Bohren, K.M., Gabbay, K.H., Quiocho, F.A. An unlikely sugar substrate site in the 1.65A structure of the human aldose reductase holoenzyme implicated in diabetic complications. Science 257:81–84, 1992.
21. Brandhuber, B.J., Boone, T., Kenney, W.C., McKay, D.B. Three-dimensional structure of interleukin-2. Science 238: 1707–1709, 1987.
22. Flaherty, K.M., DeLuca-Flaherty, C., McKay, D.B. Three-dimensional structure of the ATPase fragment of a 70K heat-shock cognate protein. Nature (London) 346:623–628, 1990.
23. Chou, P.Y., Fasman, G.D. Prediction of protein conformation. Biochemistry 13:222–245, 1974.
24. Sternberg, M.J. Prediction of protein structure from amino acid sequence. Anticancer Drug Design 1:169–178, 1986.
25. Luthy, R., McLachlan, A.D., M., Eisenberg, D. Secondary structure-based profiles: Use of structure-conserving scoring tables in searching protein sequence databases for structural similarities. Proteins 10:229–239, 1991.
26. Levine, A.J., Momand, J. Tumor suppressor genes: the p53 and retinoblastoma sensitivity genes and gene products. Biochim. Biophys. Acta 1032:119–136, 1990.
27. Kneller, D.G., Cohen, F.E., Langridge, R. Improvement in protein secondary structure prediction by an enhanced neural network. J. Mol. Biol. 214(1):171–182, 1990.
28. Abdel-Meguid, S.S., Shien, H.-S., Smith, W.W., Dayringer, H.E., Violand, B.N., Bentle, L.A. Three-dimensional structure of a genetically engineered variant of porcine growth hormone. Proc. Natl. Acad. Sci. U.S.A. 84:6434–6437, 1987.
29. Clegg, G.A., Stansfield, R.F.D., Bourne, P.E., Harrison, P.M. Helix packing and subunit conformation in horse spleen apoferritin. Nature (London) 288:298–300, 1980.
30. Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R., Klug, A. Protein disk of tobacco mosaic virus at 2.8 A resolution showing the interactions within and between subunits. Nature (London) 276:361–368, 1978.
31. Banner, D.W., Kokkinidis, M., Tsernoglou, D. Structure of the Col E1 protein at 1.7 resolution. J. Mol. Biol. 196:657–675, 1987.
32. Wilson, C., Wardell, M.R., Weisgraber, K.H., Mahley, R.W., Agard, D.A. Three-dimensional structure of the LDL receptor-binding domain of human apolipoprotein E. Science 252:1817–1822, 1991.
33. Milburn, M.V., Prive, G.G., Milligan, D.L., Scott, W.G., Yeh, J., Jancarik, J., Koshland, D.E.J., Kim, S.-H. Three-dimensional structure of the ligand-binding domain of the bacterial aspartate receptor with and without a ligand. Science 254(5036):1342–1347, 1991.
34. Pandit, J., Bohm, A., Jancarik, J., Halenbeck, R., Koths, K., Kim, S.-H. Three-dimensional structure of dimeric human recombinant macrophage-colony stimulating factor. Science 258:1358–1362, 1992.
35. Diederichs, K., Boone, T., Karplus, A. Novel fold and putative receptor binding site of granulocyte-macrophage colony-stimulating factor. Science 254:1779–1882, 1991.
36. Priestle, J.P., Grutter, M.G., White, J.L., Vincent, M.G., Kania, M., Wilson, E., Jardetsky, T.S., Kirschner, K., Jansonius, J.N. Three-dimensional structure of the bifunctional enzyme N-S'-phosphoribosyl)antranilate isomerase-indole-3-glycerol-phosphate synthase from Escherichia coli. Proc.Natl.Acad.Sci.U.S.A. 84:5690–5694, 1987.
37. Neidhart, D.J., Howell, P.L., Petsko, G.A., Powers, V.M., Li, R.S., Kenyon, G.L., Gerlt, J.A. Mechanism of the reaction catalyzed by mandelate racemase. 2. Crystal structure of mandelate racemase at 2.5 resolution: Identification if the active site and possible catalytic residues. Biochemistry 30(38):9264–9273, 1991.
38. Valencia, A., Kjeldgaard, M., Pai, E.F., Sander, C. GTPase domains of ras p21 oncogene protein and elongation factor Tu: Analysis of three-dimentional structures, sequence families, and functional sites. Proc. Natl. Acad. Sci. U.S.A. 88:5443–5447, 1991.
39. Kraulis, P.J. MOLSCRIPT, a program to produce both detailed and schematic plots of protein structures. J. Appl. Crystallogr. 24:946–950, 1991.