PROTEINS: Structure, Function, and Genetics 35:401–407 (1999)

Recognition of a Protein Fold in the Context of the SCOP Classification

Inna Dubchak,[1]* Ilya Muchnik,[3] Christopher Mayor,[1] Igor Dralyuk,[1] and Sung-Hou Kim[1,2]
[1]*E.O. Lawrence Berkeley National Laboratory, University of California, Berkeley, California*
[2]*Department of Chemistry, University of California, Berkeley, California*
[3]*Center for Discrete Mathematics and Theoretical Computer Science (DIMACS), Rutgers University, Piscataway, New Jersey*

ABSTRACT **A computational method has been developed for the assignment of a protein sequence to a folding class in the Structural Classification of Proteins (SCOP). This method uses global descriptors of a primary protein sequence in terms of the physical, chemical, and structural properties of the constituent amino acids. Neural networks are utilized to combine these descriptors in a way to discriminate members of a given fold from members of all other folds. An extensive testing of the method has been performed to evaluate its prediction accuracy. The method is applicable for the fold assignment of any protein sequence with or without significant sequence homology to known proteins. A WWW page for predicting protein folds is available at URL http://cbcg.lbl.gov/. Proteins 1999;35:401–407.** **Published 1999 Wiley-Liss, Inc.**†

Key words: protein fold prediction; computer simulated neural networks

INTRODUCTION

Large-scale sequencing projects produce a massive number of putative protein sequences in contrast to the much slower increase in the number of known three-dimensional (3D) protein structures. This creates both a need and an opportunity for extracting structural information from sequence databases. The direct prediction of a protein's 3D structure from a sequence remains elusive. However, considerable progress has been shown in assigning a sequence to a folding class.[1] There have been two general approaches to this problem. One is to use threading algorithms[2] that solve the inverse folding problem: given a group of structures and a sequence, identify the structure that is most compatible with this sequence. The other is a taxonometric approach which presumes that the number of folds is restricted and thus the focus is on structural predictions in the context of a particular classification of 3D folds. Proteins are defined as having a common fold if they have the same major secondary structures in the same arrangement and with the same topological connections.[5] Recent classifications of known protein structures have identified up to 500 "unique" folds. The classification scheme 3D_ALI, devised by Pascarella and Argos,[3] classifies 254 3D structures and associated sequences from the protein sequence databases into 83 folding classes. A more recent scheme, CATH, is a hierarchical classification of protein domain structures, which clusters proteins at four major levels: class (C), architecture (A), topology (T), and homologous superfamily $(H)_4$. Another scheme, SCOP (Structural Classification of Proteins),[5] provides a detailed and comprehensive description of the structural and evolutionary relationships among all proteins whose structures are known. Proteins possessing the same fold in these classifications sometimes have little similarity at the primary sequence level.

In a broad structural classification (all α, all β, ($\alpha + \beta$), α/β, and irregular),[6] a higher than 70% prediction accuracy can be easily achieved by various methods based on a simple presentation of protein sequences as vectors of a small number of general parameters.[7–9] Broad structural class can also be implied from a predicted secondary structure[14] or a secondary structure content of a protein.[19] It is obvious that the difficulty of prediction grows rapidly with the number of classes because the parameter vectors of the proteins from different classes are located too close to one another in parameter space.[7] The more classes there are in a classification scheme and the more similar they are, the more difficult it is to distinguish between them. The availability of fine-grained classifications of known structures, such as SCOP, has encouraged us to choose a taxonometric approach for the development of the scheme for searching for one or a few protein folds which may be similar to that of a target sequence whose structure is unknown.

We previously developed the protein fold assignment method,[10] which proved to be efficient in distinguishing the members of 83 folds in the 3D_ALI classification.[3] We describe here the results of our current efforts to predict protein folds in the context of the 128 folds of the SCOP classification and present the development, testing, and application of the assignment scheme.

MATERIALS AND METHODS

Database

We created a structural database which did not contain highly homologous protein sequences yet adequately repre-

Grant sponsor: Office of Health and Environment, Office of Energy Research, Department of Energy; Grant number: DE-AC03-76SF00098; Grant sponsor: National Science Foundation; Grant number: DBI-9723352.

*Correspondence to: Inna Dubchak, Ph.D., MS 84-171, E.O. Lawrence Berkeley National Laboratory, Berkeley, CA 94720. E-mail: ildubchak@lbl.gov

Received 3 September 1998; Accepted 26 February 1999

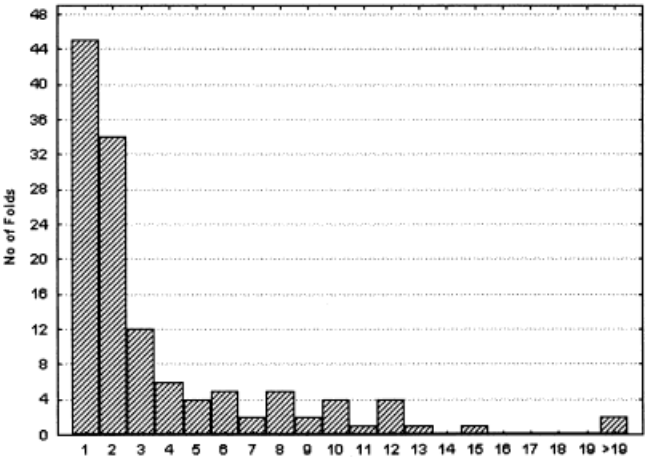

Fig. 1. Distribution of the number of proteins in the 128 folds used in our study.

sented the SCOP classification at its fold level. We used 753 proteins of the PDB_select set[11,12] where two proteins have no more than 35% of the sequence identity for the aligned subsequences longer than 80 residues. The fold according to the SCOP classification was assigned to all members of the set.

Since machine learning requires several examples for a generalization, 147 folds represented by only one protein each were removed from our database. Non-natural proteins, such as designed polypeptides, were removed from the database as well. The final subset of SCOP contained 607 proteins grouped into 128 folds. Figure 1 shows the distribution of the number of proteins in the folds in our database. The α structural class was represented by 19 folds, the β structural class by 26, the $\alpha+\beta$ class by 28, the α/β class by 25, the multiple domain class by seven, the membrane protein class by three, and the small protein class by 19.

Global Descriptors of Amino Acid Sequence

Our approach uses a combination of local and global information about amino acid sequences. A protein sequence is represented by a set of parameter vectors based on various physico-chemical and structural properties of amino acids along the sequence. These parameter vectors were constructed in two steps (for details see Dubchak et al.[10]).

Step 1. The sequence of the amino acids was transformed into sequences of certain physico-chemical or structural properties (attributes) of residues. Twenty amino acids were divided into three groups for each of six different amino acid attributes representing the main clusters of the amino acid indices of Tomii and Kanehisa.[13] Thus, for each attribute, every amino acid was replaced by the index 1, 2, or 3 according to one of the three groups to which it belonged.

The attributes we have used included the predicted secondary structure and the predicted solvent accessibility by Rost and Sander.[14] For the former, the indices 1, 2, and 3 correspond to the helix, strand, and coil, respectively. Similarly, the buried, exposed, and intermediate residues are assigned the indices 1, 2, and 3, respectively. For the other four attributes, those of hydrophobicity,[15] normalized van der Waals volume,[16] polarity,[17] and polarizability,[18] the 20 amino acids were divided into three groups according to the magnitudes of their numerical values. The ranges of these numerical values and the amino acids belonging to each group are shown in Table I.

Step 2. Three descriptors, "composition" (C), "transition" (T), and "distribution" (D), were calculated for a given attribute to describe the global percent composition of each of the three groups in a protein, the percent frequencies with which the attribute changes its index along the entire length of the protein, and the distribution pattern of the attribute along the sequence, respectively.

Let us consider the hydrophobicity attribute as an example. All amino acids are divided into three groups—polar, neutral, and hydrophobic. The "composition" descriptor C consists of the three numbers—the global percent compositions of polar, neutral, and hydrophobic residues in the protein. The "transition" descriptor T also consists of the three numbers—the percent frequency with which: 1) a polar residue is followed by a neutral residue or a neutral residue by a polar residue; 2) a polar residue is followed by a hydrophobic residue or a hydrophobic residue by a polar residue; and 3) a neutral residue is followed by a hydrophobic residue or a hydrophobic residue by a neutral residue. The "distribution" descriptor D consists of the five numbers for each of the three groups: the fractions of the entire sequence, where the first residue of a given group is located, and where 25%, 50%, 75%, and 100% of those are contained. Thus, the complete parameter vector contains 3 (C) + 3 (T) + 5 $\times$ 3 (D) = 21 scalar components.

Consequently, the six different amino acid attributes produce six parameter vectors each containing 21 scalar components. The seventh parameter vector used was the vector of the percent composition of amino acids.

Neural Networks

Three-layer feed-forward neural networks (NN) were used with the NN weights adjusted by conjugate gradient minimization as implemented in the computer program BIOPROP.[19] Various NN architectures were tested with the number of NN hidden nodes (Nhid) ranging from 0 to 3, with one or two output nodes (Nout). The simplest geometry (Nhid = 1 and Nout = 2) which achieved a good performance and had a minimum overall number of nodes (to improve generalization) was found to be adequate and thus was chosen for all of the calculations. The number of inputs was 20 for the percent composition of amino acids and 21 for each of the other six attributes. High activity output to one node indicated the assignment of the test sequence to a particular fold, and high activity to the other node indicated the assignment to the other folds.

Since the population of these two groups differed significantly in size, in training we used a resampling procedure to change a balance between two types of errors—in predicting the fold and in predicting the group of "others."

TABLE I. Amino Acid Attributes and the Division of the Amino Acids Into Three Groups for Each Attribute

Property	Group 1	Group 2	Group 3
Hyrophobicity[15]	Polar	Neutral	Hydrophobic
	R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, V, L, I, M, F, W
Normalized van der Waals volume[16]	0–2.78	2.95–4.0	4.43–8.08
	G, A, S, C, T, P, D	N, V, E, Q, I, L	M, H, K, F, R, Y, W
Polarity[17]	4.9–6.2	8.0–9.2	10.4–13.0
	L, I, F, W, C, M, V, Y	P, A, T, G, S	H, Q, R, K, N, E, D
Polarizability[18]	0–0.108	0.128–0.186	0.219–0.409
	G, A, S, D, T	C, P, N, V, E, Q, I, L	K, M, H, F, R, Y, W

This procedure, however, can not change a confidence value of predictions. We used a repetition of entities that were important: the inputs of the smaller group were fed into the NN repeatedly to match those of the larger group. In each case, the number of training examples was about 10 times higher than the number of adjustable parameters (NN synaptic weights and thresholds) to avoid "memorization" by overfitting.[20] It is obvious that a fast increase in the number of known protein structures and therefore the proteins in the non-redundant SCOP set will significantly improve learning procedures. Meanwhile we believe that for any sequence in question, the higher reliability level should be assigned to predictions of folds with the larger number of representatives available for training.

Assignment Scheme

For each fold in the database, a separate training set was constructed. The number of these sets equaled the number of folds in the database (let us call this number M). Each set consisted of two groups of proteins: those having the fold in question, and those that did not ("others"). M NNs were trained. Each NN was trained to distinguish between the proteins of a particular fold and the "others." This procedure was performed n times, where n is the number of the parameter vectors (seven at present). The final assignment of a protein sequence in question to a particular fold was based on the voting among n predictions made by the individual parameter vectors.

RESULTS AND DISCUSSION

Testing of Individual Amino Acid Attributes

In order to build the training and testing sets, the members of the 128 folding classes were shuffled by random permutation and divided into two equal parts. One-half of them were used in the training, and the other half in the testing; then the sets were switched. Thus, the testing was performed on the proteins, which were not included in the training. For each fold, two training sets and two testing sets were assembled, and, thus, two NNs were trained for every attribute. In total, $128 \times 2 = 256$ training-testing sessions were performed to estimate the performance of a particular attribute. Both the training and the testing sets contained N/2 proteins of the fold and $(607 - N)/2$ proteins of the group "others" (where N is the number of proteins in a particular fold). After training, the trained NN was tested on the test set, and three numbers were calculated: 1) the percentage of the correct positive predictions or the sensitivity (the percentage of the members of the tested fold correctly assigned to its fold); 2) the percentage of the correct negative predictions or the rejection accuracy (the percentage of the proteins from the group "others" correctly not assigned to the fold); and 3) the percentage of true positives among true and false positives, or the selectivity. This procedure was repeated for each fold and for each attribute. The accuracy of a random correct prediction to a particular fold equaled N/607 and varied from $2/607 = 0.003$ (0.3%) to $30/607 = 0.05$ (5%) for the least and the most populated folds in the database.

The number of folds predicted at 60% and higher accuracy levels (that we considered satisfactory) totaled 24 for all attributes (Table II). Among them, 17 folds were predicted by only one attribute, six folds by two attributes, and one fold by three attributes. Table II shows that each attribute worked completely differently on the different folds. Assignment to thirteen folds (10% of all folds) was possible by the predicted secondary structure attribute alone. These 13 folds represent all four structural classes. Among the folds with large number of members, the best prediction was made for the globin fold, with a high positive accuracy of 84.6% and 100% rejection accuracy. Some folds with a small number of members (2–5) also demonstrated very high levels of rejection accuracies (99.3%–100%). Other folds with a large number of members, such as the immunoglobulin-like β-sandwich fold, the TIM-barrel fold, the FAD-binding motif fold, and the small inhibitor fold, had much lower rejection accuracies of 80.2%–91.4% (and accordingly low selectivity of predictions) and positive accuracies in the range of 63.6%–71.4%.

The percent composition of amino acids has a significant correlation to the broad structural class of proteins.[19,21] Our earlier work[22] showed that the percent composition of amino acids possesses certain predictive power for much more detailed classification. As seen in Table II, the percent composition of amino acids performed well on seven folds (5% of all folds), six of which have five or more members. Two folds in this group, the Lambda repressor-like DNA-binding domain fold and the immunoglobulin fold, were predicted by percent composition alone, with the same or better accuracy than predicted by the attribute of the predicted secondary structure. The first of these folds demonstrates the same positive and rejection accuracies, while the other shows an even higher rejection accuracy

TABLE II. Predictions at Higher Than 60% Positive Accuracy for the Different Amino Acid Attributes

Attribute	Name of the fold in SCOP	Number of proteins in the subset of SCOP	Sensitivity of predictions %	Rejection accuracy %	Selectivity of predictions %
Predicted secondary structure	Alpha; Globin-like	13	84.6	100.0	100.0
	Alpha; Long alpha-hairpin	3	66.7	99.7	50.0
	Alpha; Lambda repressor-like DNA binding	5	60.0	99.5	50.0
	Alpha; Oligomers of long helices	3	66.7	100.0	100.0
	Beta; Immunoglobulin-like	30	66.7	85.4	19.2
	α/β; (TIM)-barrel	29	69.0	80.2	15.3
	α/β; FAD (NAD)-binding motif	11	63.6	85.2	7.3
	α + β; Ribonuclease A-like	3	66.7	99.3	42.8
	α + β; SH2-like	3	100.0	99.5	40.0
	α + β; Histidine-containing	2	100.0	99.5	40.0
	Multi; Sugar phosphatases	3	100.0	99.7	60.0
	Small; Small inhibitors, toxins, lectins	14	71.4	91.4	16.4
	Small; BPTI-like	3	66.7	99.7	50.0
	Small; EGF-like module	4	75.0	99.7	60.0
Percent composition of amino acids	Alpha; DNA-binding 3-helical bundle	12	66.7	99.7	66.6
	Alpha; Lambda repressor-like DNA binding	5	60.0	99.7	66.6
	Alpha; EF-hand	6	100.0	100.0	100.0
	Beta; Immunoglobulin-like	30	66.7	88.6	23.5
	Beta; Viral coat and capsid proteins	16	75.0	97.6	46.1
	α/β; Periplasmic binding, protein-like	11	63.6	92.4	25.0
Hydrophobicity	α/β; PLP-dependent transferases	3	66.7	98.7	20.0
	α/β; Periplasmic binding protein-like	11	72.7	93.1	14.2
	Small; Small inhibitors, toxins, lectins	14	71.4	94.8	25.0
Van der Waals volume	Alpha; DNA-binding 3-helical bundle	12	66.7	92.9	16.0
	Alpha; Pheromone proteins	3	66.7	97.0	10.0
	Alpha; Ferritin like	5	60.0	98.5	25.0
Polarizability	Alpha; Pheromone proteins	3	66.7	99.7	50.0
	α/β; (TIM)-barrel	29	62.1	82.5	15.1
Polarity	α/β; (TIM)-barrel	29	62.1	85.6	17.8
	Small; Classic zinc finger	3	66.7	99.7	50.0
	Small; Metallothionein	3	100.0	98.3	23.1

(88.6% compared to 85.4%) when predicted by the percent composition. Four more folds containing 6–30 protein members were predicted with positive accuracies of 63.6%–75% and rejection accuracies of 88.6%–100%, which is higher than for the same size folds predicted by the secondary structure attribute. Among the folds with a large number of members, the percent composition of amino acids worked the best for the EF-hand fold, where both positive and negative accuracies equal 100%.

The hydrophobicity attribute worked satisfactorily on three folds (2% of all folds). Two of them were α/β structural not predicted by any other attribute classes (the PLP-dependent transferase fold and the periplasmic binding protein-like fold). Three other attributes together (the normalized van der Waals volume, the polarity, and the polarizability of amino acids) worked satisfactorily on eight folds, where the pheromone protein fold, the classic zinc finger fold, and the ferritin fold were also uniquely predicted. The predicted solvent accessibility of amino acids did not demonstrate prediction performance above 60% for any fold.

Each parameter vector works differently on different folds. It appears to be important to discover an individual set of descriptors that works best for a particular fold of interest. Our future study will include a number of new physical, chemical, and structural properties, including various different hydrophobicity scales, as well as different descriptors to increase the accuracy and the specificity of the assignment.

Cross-Validation Test

As was mentioned, the reliability of assignment to a particular fold is directly related to the number of the fold representatives for training. That is why we selected the 27 most populated folds (seven or more proteins) from the non-redundant SCOP subset to use in a cross-validation test. Each protein of the fold and of the group of "others" was consequently separated from the complete set of 607 proteins. The training for recognition of a particular fold was performed on 606 proteins and the separated protein was tested to determine whether it belongs to this fold or to the group of "others." For every protein, training and the

TABLE III. Cross-Validation of the Database Folds Containing Seven or More Proteins and Testing of the Proteins Having the Same Folds From the PDB40D SCOP Set[†]

	Cross-validation				Testing proteins from the PDB40D set		
Fold	Number of proteins in the database	Correctly assigned to the fold by 4 or more votes	Percentage of correct assignments to the fold	Percentage of correct assignments to "others"	Number of proteins in the PDB40D set	Correctly assigned to the fold	Perentage of correct assignments to the fold
Alpha; Globin-like	13	9	69.2	100.0	6	6	100.0
Alpha; Cytochrome C	7	3	42.8	96.2	9	3	33.3
Alpha; DNA-binding 3-helical bundle	12	6	50.0	98.7	20	5	25.0
Alpha; 4-helical up-and-down bundle	7	2	28.5	100.0	8	2	25.0
Alpha; 4-helical cytokines	9	4	44.4	99.3	9	8	88.9
Alpha; EF-hand	7	2	28.5	98.5	9	3	33.3
Beta; Immunoglobulin-like	30	20	66.6	88.1	45	15	33.3
Beta; Cupredoxins	9	1	11.1	100.0	12	3	25.0
Beta; Viral coat and capsid proteins	16	12	75.0	100.0	13	7	53.8
Beta; ConA-like lectins/glucanases	7	2	28.5	100.0	6	3	50.0
Beta; SH3-like barrel	8	1	12.5	100.0	8	4	50.0
Beta; OB-fold	13	5	38.4	87.6	19	7	36.8
Beta; beta-Trefoil	8	1	12.5	100.0	4	2	50.0
Beta; Trypsin-like serine proteases	9	2	22.2	100.0	4	1	25.0
Beta; Lipocalins	9	3	33.3	100.0	7	2	28.5
α/β; (TIM)-barrel	29	24	82.7	100.0	48	35	72.9
α/β; FAD (also NAD)-binding motif	11	4	36.3	100.0	12	5	38.4
α/β; Flavodoxin-like	11	1	9.0	100.0	13	3	23.0
α/β; NAD(P)-binding Rossmann-fold	13	7	53.8	100.0	27	14	51.8
α/β; P-loop	10	6	60.0	100.0	12	3	25.0
α/β; Thioredoxin-like	9	4	44.4	98.5	8	3	37.5
α/β; Ribonuclease H-like motif	10	3	30.0	98.8	14	4	28.5
α/β; Hydrolases	11	6	54.5	100.0	7	4	57.1
α/β; Periplasmic binding protein-like	11	8	72.7	100.0	4	1	25.0
α + β; beta-Grasp	7	2	28.5	100.0	8	2	25.0
α + β; Ferredoxin-like	13	4	30.7	100.0	27	5	18.5
Small; Small inhibitors, toxins, lectins	14	11	78.5	100.0	27	9	33.3

[†]None of the tested proteins were included in the training set.

corresponding fold assignment was performed seven times using the seven parameter vectors derived from the six amino acid attributes and the amino acid composition. Thus, the training-testing procedure was repeated 607 × 7 times for each particular fold.

The protein sequence was predicted as having the fold if four (more than one-half) of the NNs positively predicted the fold and the same majority rule was used to assign proteins to the group of "others." The percentage of accurately predicted fold representatives and "others" was calculated for each of 27 folds. The results of this testing are presented in the left half of Table III. We estimated that the number of proteins correctly predicted by 4, 5, 6, and 7 votes were approximately the same. This means that all predictions made by a majority of votes should be taken into consideration. As seen from Table III, prediction performance varies widely for different folds–from 9% for the Flavodoxin-like fold to 82.7% for the (TIM)-barrel. The four largest folds (Immunoglobulin-like, Viral coat and capsid proteins, (TIM)-barrel and Small inhibitors) demonstrated the highest accuracy in the test (66.6–82.7%). This fact again shows the critical importance of the number of fold representatives in accurate fold assignment. Rejection accuracy was extremely high, and reached 100% for 19 folds.

Performance of the Method on Testing

To test our prediction scheme on the independent data, we used the PDB40D set recently developed by the authors of the SCOP database.[5] This set contains the SCOP sequences having less than 40% identity with each other.

This non-redundant set was produced by performing an all-against-all comparison and discarding any files that exceeded the relevant threshold identity.[5] From this set we selected proteins of the 27 folds which were represented by larger training sets in our database (the same folds were used for the cross-validation test). All PDB40D proteins that had higher than 35% identity with the proteins of the training set were excluded from the testing set. For all tested proteins, assignments to all 128 folds were made by seven parameter sets followed by voting; thus the sequence in question received 128 individual assignments. Ideally, the sequence should obtain a positive prediction for its correct fold and negative predictions for all other folds. Yet in practice, the prediction of several folds as well as the prediction of none can take place. The former happens when several predicted folds are similar.

We tested a total of 386 proteins (right half of Table III). We classify a protein as correctly predicted if it was assigned to a single correct fold or to several folds with a correct fold predicted by a higher number of votes. There were a total of 161 (41.7%) correctly assigned proteins, and the accuracy of prediction varied in the range 18.5–100% for different folds. These numbers can be considered high taking into consideration the low probability of a correct random assignment (1–5%, see above). It should be noted that although this prediction performance is lower than that of earlier work,[7–9] where predictions were made in the context of four to five structural classes, the number of folds to which a new protein could be assigned was an order of magnitude higher.

For 32 proteins (7.5%) an equal number of votes predicted two or three folds, and among them one fold was correct. These predictions of more than one fold also appeared to be very important. In spite of the fact that they do not show a unique fold, they can significantly restrict the list of possible functions for the predicted sequence.

Differentiation Between Two Similar Folds

We tested a modified version of our expert system that allows one to make a correct selection between two similar folds, i.e. we approached a reduced fold recognition problem, where the choice is limited to two folds. It is relatively easy to distinguish between the all-alpha and all-beta classes of proteins, but it is much harder to pick a correct fold within the same class.[21] Theoretically, the most general fold recognition problem can be solved by a series of binary decisions, where at each step one needs to relate an unknown protein to one of two broad classes. We were interested in the later steps of this chain of binary decisions, i.e., in the differentiation between two close protein folds.

We have chosen three pairs of folds from different structural classes. They were the Globin-like fold and 4-helical cytokines from the α-class, the Immunoglobulin-like and OB-fold from the β-class, the (TIM)-barrel and NAD(P)-binding Rossman folds from the α/β class. Proteins of each pair have a similar secondary structure content and are often predicted together in cases when the sequence in question is assigned to more that one fold. We used representatives of these folds from our SCOP subset for training and the PDB40D proteins of the same folds for testing. In order to distinguish between two particular folds, seven NNs based on seven sets of parameters were trained accordingly. Each NN permitted assignment of any protein sequence to one of the two folds. A majority rule was used in voting, i.e., the protein was assigned to a particular fold if more than half of the parameter sets predicted it to be in the fold. Results of this testing are presented in Table IV. As evident from Table IV "the fine tuning" of the general prediction scheme in order to distinguish between two similar folds demonstrated high prediction accuracy for all three pairs of folds. Such predictions in the context of two similar folds can be used as the second step in our fold prediction method. When all pairs of folds not distinguishable by the general prediction scheme (the assignment of the sequence in question to one fold in the context of 100 or more folds) are discovered, appropriate NNs can be trained to assign a sequence to one of the two folds of each particular pair.

TABLE IV. Results of Differentiation Between Two Similar Folds

Pair of folds	Fold	Number of proteins in		Correctly predicted proteins	
		Training set	Testing set	Number	%
I	Globin-like	13	6	6	100
	4-helical cytokines	9	9	9	100
II	Immunoglobulin-like	30	45	43	95.5
	OB-fold	8	19	16	84.2
III	(TIM)-barrel	29	48	45	93.7
	NAD(P)-binding Rossman	13	27	20	74.0

CONCLUSION

Our method performed well in the protein fold prediction category of the 1996 Critical Assessment of Techniques for Protein Structure Prediction (CASP2) experiment 23. One prediction was recognized as one of the best for a particular target,[24] and one more target was predicted in the correct fold. For another target, the correct fold was one of two folds predicted with equal probability.

Further improvement of the prediction performance greatly depends on the growth in the number of known protein structures. Extension of the voting scheme by including more predictions based on different sets of parameters representing different groups of physicochemical and structural properties, can also increase an accuracy of presented method.

ACKNOWLEDGMENTS

We thank Dr. Nikolai N. Alexandrov for providing us with the protein database. This work has been supported by the Office of Health and Environment, Office of Energy Research of Department of Energy (DE-AC03-76SF00098

to S.-H.K.) and by the grant from the National Science Foundation (DBI-9723352 to S.-H.K.).

REFERENCES

1. Marchler-Bauer A, Bryant, SH. A measure of success in fold recognition. Trends Biochem Sci 1997;22:236–240.
2. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–169.
3. Pascarella S, Argos P. A data bank merging related protein structures and sequences. Protein Eng. 1992;5:121–137.
4. Orengo CA, Flores TP, Taylor WR, Thornton JM. Identification and classification of protein fold families. Protein Eng 1993;6:485–500.
5. Hubbard TJP, Bart A, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 1999;27:254–256.
6. Levitt M, Chothia C. Structural patterns in globular proteins. Nature 1976;261:552–558.
7. Chou PY. In: Fasman GD, editor. Prediction of protein structure and principles of protein conformation. New York: Plenum Press; 1989. p 549–586.
8. Nakashima H, Nishikawa K, Ooi T. The folding type of a protein is relevant to the amino acid composition. J Biochem 1986;99:152–162.
9. Dubchak I, Holbrook SR, Kim S-H. Prediction of protein folding class from amino acid composition. Proteins 1993;16:79–91.
10. Dubchak I, Muchnik I, Holbrook SR, Kim, S-H. Prediction of protein folding class using global description of amino acid sequence. Proc Natl Acad Sci USA 1995;92:8700–8704.
11. Hobohm U, Scharf M, Schneider R, Sander C. Selection of a representative set of structures from the Brookhaven Protein Data Bank. Protein Sci 1992;1:409–417.
12. Hobohm U, Sander C. Enlarged representative set of proteins. Protein Sci 1994;3:522–524.
13. Tomii K, Kanehisa M. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. Protein Eng. 1996;9:27–36.
14. Rost B, Sander C. Combining evolutionary information and neural networks to predict protein secondary structure. Proteins 1994;20:216–226.
15. Chothia C, Finkelstein AV. The classification and origins of protein folding patterns. Annu Rev Biochem 1990;59:1007–1039.
16. Fauchere JL, Charton M, Kier LB, Verloop A, Pliska V. Amino acid side chain parameters for correlation studies in biology and pharmacology. Int J Pept Protein Res 1988;3:269–278.
17. Grantham R. Amino acid difference formula to help explain protein evolution. Science 1974;185:862–864.
18. Charton M, Charton BI. The structural dependence of amino acid hydrophobicity parameters. J Theor Biol 1982;99:629–644.
19. Muskal SM, Kim S-H. Predicting protein secondary structure content: a tandem neural network approach. J Mol Biol 1992;225: 713–727.
20. Hertz J, Krogh A. Introduction to the theory of neural computations. Redwood City, CA: Addison-Wesley; 1992. 326 p.
21. Chou K-C, Zhang C-T. Prediction of protein structural classes. Crit Rev Biochem Mol Biol 1995;30:275–349.
22. Mayoraz E, Dubchak I, Muchnik I. Relation between protein structure, sequence homology and composition of amino acids. In: Raweing C, Clark D, Altman R, Hunter L, Lengauer T, Wodak S, editors. Third international conference on intelligent systems for molecular biology. Cambridge, UK: AAAI/MIT Press; 1995.
23. Moult J. The current state of the art in protein structure prediction. Curr Opin Biotechnol 1996;7:422–27.
24. Levitt M. Competitive assessment of protein fold recognition and alignment accuracy. Proteins Suppl. 1997;1:92–104.