Proc. Natl. Acad. Sci. USA
Vol. 92, pp. 8700–8704, September 1995
Biophysics

Prediction of protein folding class using global description of amino acid sequence

INNA DUBCHAK*, ILYA MUCHNIK†, STEPHEN R. HOLBROOK‡, AND SUNG-HOU KIM*‡

*Department of Chemistry and ‡Lawrence Berkeley Laboratory, University of California, Berkeley, CA 94720; and †Rutgers Center for Operations Research, Rutgers University, New Brunswick, NJ 08903-5062

Contributed by Sung-Hou Kim, May 31, 1995

ABSTRACT **We present a method for predicting protein folding class based on global protein chain description and a voting process. Selection of the best descriptors was achieved by a computer-simulated neural network trained on a data base consisting of 83 folding classes. Protein-chain descriptors include overall composition, transition, and distribution of amino acid attributes, such as relative hydrophobicity, predicted secondary structure, and predicted solvent exposure. Cross-validation testing was performed on 15 of the largest classes. The test shows that proteins were assigned to the correct class (correct positive prediction) with an average accuracy of 71.7%, whereas the inverse prediction of proteins as not belonging to a particular class (correct negative prediction) was 90–95% accurate. When tested on 254 structures used in this study, the top two predictions contained the correct class in 91% of the cases.**

Examination of three-dimensional (3D) structures of proteins determined by x-ray diffraction and NMR has shown that the variety of folding patterns of proteins is significantly restricted (1, 2). Since protein sequence information grows significantly faster than information on protein 3D structure, the need for predicting the folding pattern of a given protein sequence naturally arises. Since the first relatively full classification of folding patterns of globular proteins (3), researchers have developed various schemes for classification of protein 3D structures (4–6) that are essentially based on the same spatial motifs.

If the prediction is restricted to a small number of structural classes (less than five), a prediction performance >70% can be easily achieved by using various methods based on a simple representation of sequences as vectors of a small number of general parameters. In the simplest classification, proteins are usually described in terms of the following "tertiary super classes:" all α (proteins have only α-helix secondary structure), all β (mainly β-sheet secondary structure), $\alpha+\beta$ (α-helix and β-strand secondary structure segments that do not mix), α/β (mixed or alternating segments of α-helical and β-strand secondary structure), and irregular (7–9). Several statistical methods were developed to predict whether a protein belongs to one of these classes (10–17). In a recent study on predicting protein structural class (all α, all β, or composed of α and β elements) from amino acid composition and hydrophobic pattern frequency information using computer-simulated neural networks (NNs) and statistical clustering, Metfessel *et al.* (18) obtained a prediction accuracy of 80.2%. Consideration of specific features of folding classes in the form of so-called hidden Markov models or probabilistic grammars allows a >2-fold increase in the number of classes of recognition (9). This method accurately predicts 12 classes; however, the study gives test results only for 16 sequences.

It is obvious that difficulty of folding pattern prediction grows rapidly with the number of classes. Even the distinction between $\alpha+\beta$ and α/β classes has serious problems because the parameter vectors of these structure types are located too close in parameter space (11). As the number of classes in the classification system increases, the classes become more similar and it is more difficult to distinguish among them.

Here, we present a prediction method for 83 folding classes that uses a combination of protein sequence descriptors applied to computational NNs. The method was tested by cross-validation.

Data Base and NNs

A variety of folding classifications have appeared in the last years, based not only on the grouping of 3D structures but also on analysis of sequences. Our assignment was based on the classification scheme 3D_ALI of Pascarella and Argos (6) that classifies the majority of known 3D structures (254 proteins and protein domains) into 83 classes, 38 of them having two or more members and the other 45 classes containing only a single protein example. This data base was based on a superposition among protein structures with similar main-chain fold and contains a large number of protein families with low sequence homology. The average sequence identity over all possible aligned Protein Data Bank sequence pairs is 15%.

After all amino acid sequences were substituted by property vectors (see below) and an assignment to a definite class was made for each protein in the data base, computer-simulated NNs were applied (for a review of NNs, see ref. 19). This tool has been successful in the prediction of protein structural features (17, 20–23). Three-layer feed forward NNs with weights adjusted by conjugate gradient minimization using the computer program BIOPROP (22) were utilized in our study. Various NN architectures were tested with the number of NN hidden nodes (N_{hid}) varied from 0 to 3, with one or two outputs (N_{out}). The simplest geometry with $N_{\mathrm{hid}} = 1$ and $N_{\mathrm{out}} = 2$ that achieved good performance and had a minimum overall number of nodes [to improve generalization (24)] was chosen for all calculations (Fig. 1). The number of training examples is 254, which is basically enough to avoid overfitting (24) since it is at least 5 times more than a number of adjustable parameters (NN synaptic weights and thresholds) varied from 11 to 46, depending on a number of inputs used.

Physical and Stereochemical Properties of Amino Acids Used

Since the NN and all other methods of pattern recognition require property vectors as input, a sequence of amino acids should be replaced by a sequence of symbols representing local physicochemical properties. The first amino acid property used in this study is relative hydrophobicity of amino acids. The

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: 3D, three-dimensional; NN, neural network; HP, hydrophobicity attribute; SA, solvent accessibility.

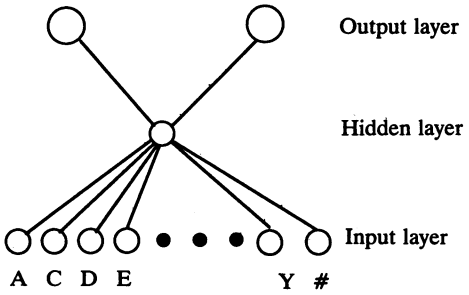

FIG. 1. Schematic diagram of the architecture of the computational NN discussed.

classification of Chothia and Finkelstein (4) was used, where amino acids form three groups: Arg, Lys, Glu, Asp, Gln, and Asn as polar; Gly, Ala, Ser, Thr, Pro, His, and Tyr as neutral; and Cys, Val, Leu, Ile, Met, Phe, and Trp as hydrophobic.

The second amino acid attribute we chose is predicted secondary structure. The methods of secondary structure prediction, including NN predictions, have been significantly improved in recent years and have reached 70% performance (20, 21, 25, 28). Two NN prediction methods (20, 21), one of which (21) was modified by us in the program PROBE (26), have been used in this study. Both of them use a three-state model (helix, strand, and coil). A consensus prediction among the above two methods was introduced as a four-state model—helix, strand, or coil, when both methods give the same assignment for a given residue, and unknown, when the two methods disagree.

Predicted solvent accessibility of amino acids was used as the third amino acid attribute. Trained NN can correctly predict solvent accessibility for 72% of residues in a testing set by using a binary model (buried/exposed) (23). A residue was considered to be exposed if >20% of its surface is exposed.

In summary, amino acids are divided into three groups based on hydrophobicity (hydrophobic, neutral, and polar), three groups based on secondary structure prediction (helix, strand, and coil), four groups based on consensus secondary structure prediction (helix, coil, strand, and unknown), and two groups based on solvent accessibility (buried and exposed).

Global Protein Sequence Descriptors

We use three descriptors, composition (*C*), transition (*T*), and distribution (*D*), to describe the global composition of a given amino acid property in a protein, the frequencies with which the property changes along the entire length of the protein, and the distribution pattern of the property along the sequence, respectively. Some of these were introduced in our earlier work (27). How to construct these descriptors is explained below by using a model amino acid sequence consisting of two kinds of amino acids.

The model sequence (Fig. 2) includes 10 type A residues ($n_1 = 10$) and 16 type B residues ($n_2 = 16$). The percent compositions are calculated as follows: $n_1 \times 100.0/(n_1 + n_2) = 38.5\%$ for A and $n_2 \times 100.0/(n_1 + n_2) = 61.5\%$ for B. These two numbers represent the first descriptor, *C*. The second descriptor, *T*, characterizes the percent frequency with which A is followed by B or B is followed by A. In this case, there are 10 transitions of this type, that is (10/25) × 100.0 = 40%. The third descriptor, *D*, deserves a more detailed discussion. For a given property of amino acids, the distribution of the property along the protein chain is described by five chain lengths (in percent), within which the first, 25%, 50%, 75%, and 100% of the amino acids with a certain property are contained. In the example of Fig. 2, the first residue of group A coincides with the beginning of the chain, so the first number of *D* descriptor equals 0.0. Twenty-five percent of all group A residues (rounded to 2 residues) are contained within the first 4 residues of the protein chain, so the second number equals (4/26) × 100.0% = 15.4%. Similarly, 50% of group A residues are within the first 12 residues of the chain; thus, the third number is (12/26) × 100.0% = 46.1%. The fourth and fifth numbers of the distribution descriptor are 73.1% and 100%, respectively. Analogous numbers for group B are 7.5%, 23.1%, 53.8%, 79.9%, and 92.3%, respectively.

In summary, 13 numbers are used to describe the model sequence shown in Fig. 2 with respect to a given property (type A or B): 2 for composition, 1 for transition, and 10 for distribution. Thus, the chain descriptor values to be used as the input for NN are 38.5, 61.5, 40.0, 0.0, 15.4, 46.1, 73.1, 100.0, 7.6, 23.1, 53.8, 76.9, and 92.3.

Combined Protein Sequence Descriptors

For each of the chosen amino acids attributes, all three descriptors (*C, T, D*) were calculated, combined, and used as input parameters for NN training. For the hydrophobicity attribute (HP), the three numbers represent *C*—the percent compositions of polar, neutral, and hydrophobic residues in the protein. The first number of *T* is the percent frequency with which a polar residue is followed by a neutral or a neutral residue by a polar residue. The second number is the frequency of a polar residue followed by a hydrophobic residue or a hydrophobic residue followed by a polar residue. The third number corresponds to the cases where a neutral residue is followed by a hydrophobic residue or a hydrophobic residue is followed by a neutral residue. Distinguishing polar-to-neutral transition from neutral-to-polar transition did not improve prediction results significantly (the same for other types of transitions). The *D* descriptor has five numbers for each of the three residue types (neutral, polar, and hydrophobic). Thus the sequence description of a protein in terms of hydrophobicity consists of 21 numbers.

. Three methods were used for the secondary structure assignment: prediction of Qian and Sejnowski (20) (SS1), modified prediction of Holley and Karplus (21) as modified by Holbrook *et al.* (26) (SS2), and the consensus prediction of the two methods (SS3). For the SS1 and SS2 methods (three-state models), combined descriptors consist of 21 numbers: *C* (3 numbers), *T* (3 numbers), and *D* (15 numbers). For four-state [helix(*h*), strand(*s*), coil(*c*), and unknown(*x*)] consensus prediction (SS3), the sequence descriptor consists of 30 numbers: four numbers for the *C*, six numbers for the *T* (*h* to *s* or *h* to *s*, *h* to *c* or *c* to *h*, *h* to *x* or *x* to *h*, *s* to *c* or *c* to *s*, *s* to *x* or *x* to *s*, and *c* to *x* or *x* to *c*), and 20 numbers for *D*—a distribution of each of four secondary-structure states.

The chain descriptor for solvent accessibility (SA) contains seven numbers: one number for *C*, the percent composition of buried residues; one number for T, the percent frequency with which a buried residue is followed by an exposed residue or an

```
Sequence            A  B  B  A  B  B  B  B  B  A  A  A  A  B  B  B  A  B  A  B  B  B  B  B  A  A
Sequence numbering  1           5              10             15             20             25
Group A numbering   1        2                 3  4  5  6           7     8                 9  10
Group B numbering      1  2     3  4  5  6  7              8  9  10    11    12 13 14 1516
A-B transitions      |        |                          |           |     |
B-A transitions            |                 |                    |     |                 |
```

FIG. 2. Model sequence consisting of two types of residues—A and B (see text for discussion on the sequence descriptors).

Table 1. Combined descriptors in general terms of relative hydrophobicity and predicted secondary structure

	Relative hydrophobicity*			Predicted secondary structure†			% of proteins predicted not more than in	No. of NN
Scheme	*C*	*T*	*D*	*C*	*T*	*D*	three classes	inputs
1	X	X					3.5	6
2			X				58.3	15
3	X		X				79.9	18
4		X	X				67.3	18
5	X	X	X				83.8	21
6						X	57.5	15
7				X		X	66.1	18
8					X	X	70.1	18
9				X	X	X	62.2	21
10	X			X	X	X	84.6	36
11	X		X	X	X	X	92.1	39
12	X	X		X	X	X	87.4	39
13	X	X	X	X	X	X	90.2	42

X, particular descriptor is a part of combined descriptor.

*Grouped as neutral, polar, and hydrophobic according to ref. 4.

†Prediction by the method of Qian and Sejnowski (20), a three-state model (helix, strand, coil).

exposed followed by a buried residue; and five numbers for *D*, the distribution of buried residues along the chain. The *C* and *D* descriptors of exposed residues were ignored since in two group property usually data on one group are sufficient.

General Prediction Scheme and Evaluation of Combined Protein Chain Descriptors

For each of the 83 classes in the data base, an independent NN was trained by using the various combination of descriptors as inputs to distinguish this class from all other classes. For an amino acid sequence, each network gives one of two answers: yes, if its 3D structure corresponds to that class, or no, if it belongs to one of the remaining 82. A training set for each class consisted of two groups—the proteins from the class and the group of "others" (i.e., 254 minus the proteins from the class). The population of these two groups differs significantly in size. Since the number of examples in each group should be approximately equal for balanced training (22), we used weighting procedure: the inputs of the smaller group were repeated to match those of the larger group. For output, we used two nodes, where high activity to one node indicated belonging to particular class, and high activity to the other node indicated belonging to the other classes. These series of trainings resulted in 83 sets of NN weights.

After the NN training process, a series of testings was performed on each of the 254 structures. Ideally each protein should be assigned to its own class by that class NN and to others by the other 82 NN. In reality, we usually found not one but several NNs assign a protein to their respective classes. Therefore, we accepted the performance as satisfactory if a given protein sequence is predicted to belong to <3 of 83 classes (≈3%), and one of those is correct. The percent of the proteins that pass this criterion for various combined descriptors is as follows: 83.8% for HP, 62.2% for SS1, 58.6% for SS2 (21 NN input nodes), 50.8% for SS3 (30 NN input modes), and 25.6% for SA (7 NN input nodes).

To select the best combination of descriptors, a series of experiments on training/testing by using various combinations of all HP and SS1 descriptors was conducted. The results of a search for the best chain representation are shown in Table 1. It is obvious that the performance of a NN using various combined descriptors as inputs varies significantly. For HP, a combination of all three descriptors (scheme 5) gives the best performance. One can see that the *D* descriptor is the most important in the scheme, since a combination of the *C* and *T* descriptors by themselves (scheme 1) gives extremely poor performance. However, all the schemes that include *D*s (schemes 2–5) perform fairly well. In the case of predicted secondary structure, the situation is similar, but the best (scheme 8) consists of *C* and *D* descriptors without the *T*s. As to the combined HP and SS1 (schemes 10–13), all give good performance. The best, scheme 11, uses 21 inputs from HP and 18 inputs—the *C* (3 inputs) and *D* (15 inputs)—from SS1. Schemes 5, 9, and 11 were used for further cross-validation.

A summary of NN inputs is shown in Table 2. In the analysis of 13 combined descriptors presented in Table 1, each protein sequence was tested 13 × 83 = 1079 times and, thus, a total number of testings on 254 sequences was 254 × 1079 = 274,066. Additional testing was performed on schemes III, IV, and VI by using SS2, SS3, and SA (Table 2). It required 83 × 3 × 254 = 63,246 testings.

Cross-Validation by Voting

Cross-validation testing was performed for 15 of the largest protein classes with four or more protein members (Table 3). To assemble training and testing sets for a given class, proteins from the other 82 classes were shuffled by random permutation and then divided into 10 subsets. One protein from the class and proteins from 1 subset of other classes were used for testing and all other proteins of the class and 9 remaining subsets were used for training. All possible combinations of proteins from the class and 10 subsets were made, so the overall number of training/testing sessions for each class of proteins was equal to 10*n*, where *n* is a number of proteins in the class. This way each protein in a given class was tested 10 times with different subsets and each protein of the set others was tested *n* times with proteins in the tested class. For example, for folding class 16 containing 32 proteins, it means 32 × 10 = 320 training/testing sessions. After all training and testing sets were assembled, each protein sequence was transformed to a string of descriptors used as inputs for NN training and testing according to schemes I–VI from Table 2.

For each set, six parallel trainings based on six different combined chain descriptors were performed and testing was performed on corresponding testing sets. Two numbers rep-

Table 2. Main combined descriptors used in the cross-validation

Scheme			Total no. of NN	Components of combined descriptor, no. of corresponding NN inputs		
No.	Name	Amino acid attribute	inputs	*C*	*T*	*D*
I	HP	Relative hydrophobicity	21	3	3	15
II	SS1	Predicted secondary structure	21	3	3	15
III	SS2	Predicted secondary structure	21	3	3	15
IV	SS3	Constituent predicted secondary structure	30	4	6	20
V	HP+SS1	Scheme 11 from Table 1	39	3+3	3(HP)	15+15
VI	SA	Predicted solvent accessibility	7	1	1	5

Table 3. Summary on 3D_ALI classes (6), used in cross-validation testing

Class no.	3D_ALI code	Protein Data Bank codes	n_k
2	AC_PROT	1CMS(1-175), 1CMS(176-323), 4APE(2-174), 4APE(175-326), 2APP(1-174), 2APP(175-323), 2APR(1-178), 2APR(179-325), 4PEP(1-174), 4PEP(175-326)	10
4	BINDING	1ABP, 2LIV, 1LBP, 2GBP	4
6	CA_BIND	3CLN, 5CPV, 3ICB, 4TNC, 5TNC	5
7	GCR	1GCR(1-39), 1GCR(40-87), 1GCR(88-128), 1GCR(129-174), 2GCR(1-39), 2GCR(40-87), 2GCR(88-128), 2GCR(129-174)	8
8	CYTC	451C, 1CCR, 1CYC, 5CYT, 3C2C, 155C	6
14	FERREDOX	1FDX(1-54), 1FDX(27-54), 4FD1(1-106), 4FD1(31-57), 1FXB	5
15	GLOBIN	4HHB(α), 4HHB(β), 2MHB(α), 2MHB(β), 1FDH(γ), 1MBD, 1MBS, 2LHB, 1PMB, 1MBA, 1ECA, 1LH1	12
16	IGB	2FB4(L1-109), 2FB4(L110-214), 2FB4(H1-118), 2FB4(H119-221), 2FBJ(L1-106), 2FBJ(L107-213), 2FBJ(H1-118), 2FBJ(J119-218), 1FC2(238-339), 1FC2(340-443), 1MCP(L1-113), 1MCP(H1-122), 1PFC, 3FAB(L1-109), 3FAB(L104-214), 3FAB(H1-117), 3FAB(H118-220), 2HFL(H1-116), 2FHL(H117-213), 2HFL(L1-105), 1F19(L1-108), 1F19(L109-215), 1F19(H1-123), 1F19(H124-220), 1CD4, 1REI, 3HLA(A183-270), 3HLA(B1-99), 4FAB(L1-112), 3HFM(L1-108), 1MCW, 2RHE	32
18	INHIBIT	1TGS, 3SGB, 2OVO, 1OVO	4
20	NBD	4MDH, 2LDB, 1LDM, 5LDH, 2LDX, 1LLC, 8ADH, 3GPD, 1GPD, 1GD1, 1FX1, 4FXN, 2SBT, 3ADK, 8ATC	15
25	RDX	3RXN, 4RXN, 1RDG, 6RXN	4
28	SBT	CSE, 1SBT, 1TEC, 1PRK	4
29	S_PROT	1TON, 2PKA, 2PTN, 2TRM, 4CHA, 3EST, 1HNE, 2RP2, 1SGT, 2SGA, 3SGB, 2ALP	12
31	VIRUS	4RHV(VP1), 4RHV(VP2), 4RHV(VP3), 4SBV, 2MEV(VP1), 2MEV(VP2), 2MEV(VP3), 2TBV, 2STV, 2PLV(VP1), 2PLV(VP2), 2PLV(VP3), 1R1A(VP1), 1R1A(VP2), 1R1A(VP3)	15
32	WGA	7WGA(1-43), 7WGA(44-86), 7WGA(87-129), 7WGA(130-171), 9WGA(1-43), 9WGA(44-86), 9WGA(87-129), 9WGA(130-171)	8

Number of classes is in 3D_ALI classification (6).

resent a final result of each testing on one combined descriptor: (*i*) the NN output of one or zero indicates whether the protein sequence tested was correctly assigned to the class being tested or not and (*ii*) how many proteins from the complementary subset were correctly assigned to other classes. After obtaining all six predictions for each testing set, it was possible to make a final conclusion based on voting. Two systems of voting were used. In system A (Fig. 3), a protein is assigned to the class when not less than three out of six schemes (I–VI) simultaneously predict a protein to be in the class; in system B, a protein is assigned when two out of three basic schemes (I, II, and VI) predict it.

The overall number of training/testing sessions for each class of proteins equals $T_k = p10n_k$, where k is the number of the class, n_k is a number of proteins in the class k, p equals 3 or 6 depending on the scheme of voting, and 10 is a number of subsets (see above). We derive two numbers representing the results of testing for a particular class: M_k, percent of proteins from class k that are correctly predicted to be in k, and L_k, percent of proteins from any of the other 82 classes that are correctly predicted to be in one of those other classes. J_k, which is analogous to L_k, but only for the 15 largest classes. The equations used and the results of cross-validation are summarized in Table 4.

Results and Discussion

Assigning a protein sequence to 1 of 83 folding classes is very difficult because of the similarity among the classes. In the present study, we used several different protein chain repre-

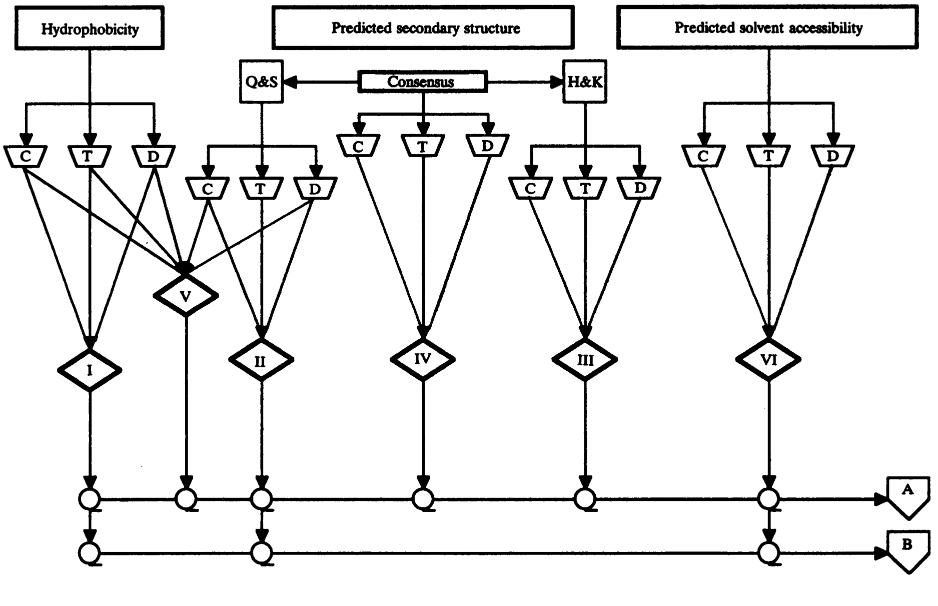
Hydrophobicity
Predicted secondary structure
Predicted solvent accessibility
Q&S
Consensus
H&K
C
T
D
V
I
II
IV
III
VI
A
B

FIG. 3. General scheme of protein folding class prediction. Rectangles, amino acid attributes; trapezoids, protein sequence descriptors; C, composition; T, transition; D, distribution; rhombuses, prediction scheme (combined descriptor); pentagons, voting system; A, consensus prediction based on prediction schemes I–VI; B, based on prediction schemes I–III.

Table 4. Results of cross-validation for 15 classes having four or more proteins

k	n_k	p	T_k	S_k	M_k	L_k	J_k
2	10	6	600	14,640	75.0	79.1	90.7
		3			70.0	85.9	89.1
4	4	6	240	6,000	75.0	88.1	98.1
		3			75.0	92.3	99.8
6	5	6	300	2,470	44.0	95.9	98.7
		3			50.0	96.9	99.1
7	8	6	480	11,856	77.5	86.1	99.2
		3			60.0	92.3	98.3
8	6	6	360	8,964	60.0	95.9	96.9
		3			61.7	95.7	96.4
14	5	6	300	7,470	58.0	96.1	99.6
		3			46.0	94.4	98.2
15	12	6	720	17,568	87.5	95.8	94.9
		3			68.3	94.9	93.6
16	32	6	1920	42,624	86.3	92.9	91.7
		3			75.9	94.8	83.2
18	4	6	240	6,000	75.0	88.6	100
		3			40.0	85.0	100
20	15	6	900	21,510	76.7	94.6	81.9
		3			70.0	91.8	79.5
25	4	6	240	6,000	65.0	96.5	99.8
		3			72.5	94.3	99.7
28	4	6	240	6,000	50.0	82.7	98.7
		3			47.5	77.2	98.8
29	12	6	720	17,568	70.8	85.9	80.1
		3			73.7	85.7	81.7
31	15	6	900	21,510	75.3	91.0	89.7
		3			61.3	88.6	89.5
32	8	6	480	11,856	100	97.8	99.8
		3			100	92.5	99.8

k, Number of class; n_k, number of proteins in class k; p, number of chain descriptions participated in voting; $T_k = p10n_k$, the overall number of training/testing sessions for class k proteins; $S_k = p(254 - n_k)n_k$, total number of testings of proteins not belonging to k; M_k, percent of the proteins from class k correctly predicted to be in k; L_k, percent of the proteins from other classes correctly assigned to one of those other classes; J_k, percent of class k proteins correctly assigned to the group others being in complement subsets to other 14 classes. The overall number of testings for 15 classes equals $\Sigma\, T_k + \Sigma\, S_k = 221{,}676$.

sentations to overcome this difficulty. We show that increasing the number of independent chain representations leads to an improved protein class prediction without greatly increasing the number of parameters for chain description.

In our work, we chose to do validation tests in a large number of ways for the 15 classes with four or more members. As evident from Table 4, prediction performance M_k for different classes varies significantly. This means that certain types of folds are best recognized by specific parameter types. Scheme A with p equal to 6 performs better than scheme B with p equal to 3 for 9 classes out of 15, in 4 cases it works worse, and in 2 classes it gives the same result. The average numbers for M_k for A and B are 71.7 and 64.8%, respectively. What this means is that as more chain descriptions or groups of parameters are used in the voting scheme, better class prediction performance is achieved. As to the recognition performance L_k and J_k, both voting schemes give very close results: L_k = 91.2% for A and 90.8% for B; J_k = 94.6% for A and 93.7% for B. L_k, as well as J_k, characterizes the ability of our prediction scheme not to recognize a protein as belonging to an inappropriate class or, in other words, not to assign a protein to a particular class incorrectly ("false positive"). The main difference between L_k and J_k is in the number of proteins for which these characteristics were calculated. L_k identifies a performance for all 254 proteins, while J_k deals only with proteins from the classes having four or more proteins. A similarity of these two numbers for both voting schemes proves that the performance of this approach in distinguishing between a particular class and others is extremely high and depends neither on a number of proteins in this class nor on their sequence similarity.

When tested on 254 structures used as the data base in our study, the top predicted class was correct in 59% of the cases, the top two predictions contained the correct folding class in 91% of the cases, and the top three predictions included the correct class in 97% of the cases.

For a larger set of folding classes, additional chain descriptors and combined descriptors are needed. We assume that further addition of other parameter sets can help to increase the performance and to add more classes for recognition. Additional amino acid attributes, such as charge, side-chain bulk, backbone flexibility, hydrophobic moment, and various types of descriptors, could be used.

A computer program for the prediction of protein folding class described here is available on request from the authors.

We thank Dr. Steven M. Muskal for providing the BIOPROP NN software. This work was supported by the U.S. Department of Energy and, in part, by Grant NSF-CCR-91-1999 from National Science Foundation and Grant F49620-93-1-0041 from Air Force Office of Sponsored Research, Department of Defense.

1. Chothia, C. (1992) *Nature (London)* **357,** 543–544.
2. Finkelstein, A. V. & Ptitsyn, O. B. (1987) *Prog. Biophys. Mol. Biol.* **50,** 171–190.
3. Richardson, J. (1981) *Adv. Protein Chem.* **34,** 167–339.
4. Chothia, C. & Finkelstein, A. V. (1990) *Annu. Rev. Biochem.* **59,** 1007–1039.
5. Orengo, C. A., Flores, T. P., Taylor, W. R. & Thornton, J. M. (1993) *Protein Eng.* **6,** 485–500.
6. Pascarella, S. & Argos, P. (1992) *Protein Eng.* **5,** 121–137.
7. Levitt, M. & Chothia, C. (1976) *Nature (London)* **261,** 552–558.
8. Richardson, J. S. & Richardson, D. C. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. Fasman, G. D. (Plenum, New York), pp. 1–98.
9. White, J. V., Stultz, C. M. & Smith, T. F. (1994) *Math. Biosci.* **119,** 35–75.
10. Chou, P. Y. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation*, ed. Fasman, G. D. (Plenum, New York), pp. 549–586.
11. Nakashima, H., Nishikawa, K. & Ooi, T. (1986) *J. Biochem. (Tokyo)* **99,** 152–162.
12. Klein, P. & Delisi, C. (1986) *Biopolymers* **25,** 1569–1672.
13. Klein, P. (1986) *Biochim. Biophys. Acta* **874,** 205–215.
14. Chou, K. C. & Zhang, C. T. (1992) *Eur. J. Biochem.* **207,** 429–433.
15. Chou, K. C. & Zhang, C. T. (1993) *J. Protein Chem.* **12,** 169–178.
16. Zhang, C. T. & Chou, K. C. (1992) *Protein Sci.* **1,** 401–408.
17. Dubchak, I., Holbrook, S. R. & Kim, S.-H. (1993) *Proteins* **16,** 79–91.
18. Metfessel, B. A., Saurugger, P. N., Connelly, D. P. & Rich, S. S. (1993) *Protein Sci.* **2,** 1171–1182.
19. Holbrook, S. R., Muskal, S. M. & Kim, S.-H. (1993) in *Artificial Intelligence and Molecular Biology*, ed. Hunter, L. (AAAI/MIT Press, Menlo Park, CA), pp. 161–194.
20. Qian, N. & Sejnowski, T. J. (1988) *J. Mol. Biol.* **202,** 865–884.
21. Holley, L. H. & Karplus, M. (1989) *Proc. Natl. Acad. Sci. USA* **86,** 152–156.
22. Muskal, S. M. & Kim, S.-H. (1992) *J. Mol. Biol.* **225,** 713–727.
23. Holbrook, S. R., Muskal, S. M. & Kim, S.-H. (1990) *Protein Eng.* **3,** 659–665.
24. Hertz, J., Krogh, A. & Palmer, R. G. (1992) *Introduction to the Theory of Neural Computation* (Addison–Wesley, Redwood City, CA), pp. 147–157.
25. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
26. Holbrook, S. R., Dubchak, I. & Kim, S.-H. (1993) *Biotechniques* **14,** 984–989.
27. Dubchak, I., Holbrook, S. R. & Kim, S.-H. (1993) in *Proceedings, First International Conference on Intelligent Systems in Molecular Biology*, eds. Hunter, L., Searls, D. & Shavlik, J. (AAAI/MIT Press, Menlo Park, CA), pp. 118–126.
28. Kneller, D. G., Cohen, F. E. & Langridge, R. (1990) *J. Mol. Biol.* **214,** 171–182.