Protein folding class predictor for SCOP: approach based on global descriptors

Inna Dubchak[1], Ilya Muchnik[2], Sung-Hou Kim[1]

[1]Lawrence Berkeley National Laboratory, Mailstop: Calvin, Berkeley, CA 94720. E-mail: (ildubchak, shkim)@lbl.gov
[2]RUTCOR - Rutgers University Center for Operations Research
P.O. Box 5062, New Brunswick, NJ, 08903-5062. E-mail: muchnik@rutcor.rutgers.edu

Abstract.
This work demonstrates new techniques developed for the prediction of protein folding class in the context of the most comprehensive Structural Classification of Proteins (SCOP). The prediction method uses global descriptors of a protein in terms of the physical, chemical and structural properties of its constituent amino acids. Neural networks are utilized to combine these descriptors in a specific way to discriminate members of a given folding class from members of all other classes. It is shown that a specific amino acid's properties work completely differently on different folding classes. This creates the possibility of finding an individual set of descriptors that works best on a particular folding class.

Introduction

The direct prediction of a protein three dimensional structure from the sequence alone remains elusive, however considerable progress has been made in assigning a sequence to a folding class. There have been two general approaches to this problem. Threading algorithms attempt to solve the inverse protein folding problem: given a group of structures and a sequence, identify the structure that is most compatible with this sequence.

The second approach has been taxonometric. This approach attempts to bring order to the concept that the number of 3D folds is restricted by providing a set of distinct 3D folds which span all known 3D structures. The most recent classifications are fine-grained, providing ~80 to ~350 folding classes describing 3D protein structure (Pascarella & Argos 1992; Orengo et al. 1993; Murzin et al. 1995). Availability of fine-grained classifications has encouraged us to work on the development of the comprehensive scheme for predicting the protein folding class for a target sequence whose structure is unknown. The information derived from such a class assignment is substantial, guaranteed by the similarity between the 3D structures and the functions of class members. The advantages of this approach are also substantial, since the existence of several representatives within the classes allows one to extract common features of class members which can be used to assert or exclude the membership within the class.

We previously developed the protein folding class prediction method (Dubchak et al. 1995) that used machine learning applied to the intermediate (83 folding classes) 3D_ALI classification scheme (Pascarella & Argos 1992). This method: 1) introduced a global description of a protein sequence in terms of biochemical and structural properties of amino acids; 2) used computer simulated neural networks (NN) to combine these descriptors in specific ways to discriminate members of a given folding class from members of all other classes; 3) used a voting procedure among predictions based on different descriptors to decide on the final assignment.

We concentrate our current efforts on the extensive SCOP classification that provides a good target for the development of prediction algorithms. It is obvious that the complexity of the folding pattern prediction grows rapidly with the number of classes; that is why it was necessary to develop new techniques to complement the existing prediction scheme.

Materials and Methods

General prediction scheme.

All protein sequences in the chosen database are transformed into the inputs for the learning system in two steps:
(a) The sequence of amino acids is transformed into a sequence expressed in terms of a particular local attribute, for example, in terms of hydrophobicity each amino acid is replaced by one of three letters - H (hydrophobic), N (neutral), or P (polar).
(b) The descriptors (C, T, D - see below) are calculated and the vector of the combination of the descriptors is constructed for use as an input to the learning system.

A separate training set is built for each class in the database. Each set consists of two groups of proteins, one contains the proteins from the class (group A), the second, the proteins from all other classes (group B, or 'others'). One NN would allow one to distinguish between proteins of group A and B. After a training series is performed for all classes, NN weights for each class in the database are found. This strategy results in a highly flexible modular

system for recognition where adding more classes to the classification does not require an extensive retraining of the whole system.

Database

The database for protein fold recognition that did not contain highly homologous proteins and adequately represented SCOP classification was created on the basis of 35% cutoff PDB_select set (Hobohm & Sander 1994). These sets are non-redundant lists of PDB sequences each at a different cutoff of pairwise protein similarity. After removing SCOP classes represented by only one protein and the classes of designed polypeptides the database for all our calculations contained 607 proteins from those listed in the 35% PDB_select file. These proteins represented 128 folding classes of SCOP (as of 09/1996 (Murzin et al. 1995)).

Global sequence descriptors.

Our approach consists in using the association of local and global information about amino acid sequences. We developed the representation for a protein sequence that includes a small number of descriptors based on various physico-chemical and structural properties of amino acids. We used three descriptors, "composition" (C), "transition" (T), and "distribution" (D), to describe the global composition of a given local 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. It was shown (Dubchak et al. 1995) that the introduction of new 'transition' and 'distribution' characteristics significantly enhanced protein folding class prediction.

In this study the vectors of descriptors for all attributes described in the next section contained 21 scalar components. The 20 - dimensional vectors of amino acid composition were also used as descriptors of protein sequences.

Amino Acid Attributes

The 20 amino acids have different physical, chemical and biochemical properties such that the same segment of the protein chain can be described by a variety of property patterns. It is shown (Selbig, Kaden & Koch 1992) that structurally meaningful properties are often not explicit and intermingle with other properties. That is why it is critical to study as many amino acid properties and their mutual combinations as possible. We selected properties from all the main clusters of amino acid indices (Tomii & Kanehisa 1996).

The most accurate-to-date protein secondary structure prediction by Rost and Sander (Rost & Sander 1993) obtained by PHD E-mail server (Rost, 1996) was utilized in our study. This method gives predicted secondary structures (PHD_SS) as three-state models: helix, strand, and coil. Grouping of amino acids based on the other properties was arbitrary. We used the numerical scale of a particular property and separated the 20 amino acids into three groups of approximately equal size according to their numerical values on this scale. The ranges of these numerical values for all selected groups of amino acids taken from the original papers are shown in Table 1.

Neural Networks

Three layer feed forward NN with weights adjusted by the conjugate gradient minimization technique using the BIOPROP software (Muskal & Kim 1992) were used. Ninp was equal to 20 for percent composition of amino acids and 21 for all other attributes, Nhid was equal to 1 and Nout was equal to 2. High activity output to one node indicated assignment to a particular class, and high activity to another node - inclusion to the group of 'others'.

Table 1. Amino acid attributes and the classification of amino acids into three groups according to the attribute.

Property	Group 1	Group 2	Group 3
Hyrophobicity (Chothia & Finkelstein, 1990)	Polar R,K,E,D,Q,N	Neutral G,A,S,T,P,H,Y	Hydrophobic C,V,L,I,M,F,W
Normalized van der Waals volume (Fauchere, 1988)	0 - 2.78 G,A,S,C,T,P,D	2.95 - 4.0 N,V,E,Q,I,L	4.43- 8.08 M,H,K,F,R,Y,W
Polarity (Grantham, 1974)	4.9 - 6.2 L,I,F,W,C,M,V,Y	8.0 - 9.2 P,A,T,G,S	10.4 - 13.0 H,Q,R,K,N,E,D
Polarizability (Charton & Charton,1982)	0 - 0.108 G,A,S,D,T	0.128 - 0.186 C,P,N,V,E,Q,A,L	0.219 - 0.409 K,M,H,F,R.Y,W
Normalized frequency of alpha-helix (Chou & Fasman, 1978)	0.57 - 0.83 G,P,N,Y,C,S,T	0.98 - 1.08 R,H,D,V,W,I	1.11 - 1.51 Q,F,K,L,A,M,E

Results and discussion

Testing new attributes

In order to build testing sets, each of 128 classes was shuffled by a random permutation and divided into two equal parts. One half of its sequences was included in the training set, the other half was involved in testing, and vice versa. Thus, the testing was performed on the proteins which did not participated in training. Two training sets and two corresponding testing sets were assembled for the prediction of each class in terms of every attribute, and accordingly two neural networks were trained. In total 128 * 2 = 256 training-testing sessions were performed to estimate a performance of a particular attribute. Both training and testing set contained N/2 (N-number of proteins in a particular class) proteins of the class and (607-N)/2 proteins in the group 'others'.

Table 2. Predictions at a 60% and higher accuracy for different amino acid attributes.

Attribute	Name of the folding class in SCOP	Number of proteins in the class	Correct positive %	Correct negative %
PHD_SS	Alpha; Globin-like	13	84.6	100.0
	Alpha; Long alpha-hairpin	3	66.7	99.7
	Alpha; lambda repressor-like DNA bind.	5	60.0	99.5
	Alpha; Oligomers of long helices	3	66.7	100.0
	Beta; Immunoglobulin-like	30	66.7	85.4
	A/B; beta/alpha (TIM)-barrel	29	69.0	80.2
	A/B; FAD (NAD)-binding motif	11	63.6	85.2
	A+B; Ribonuclease A-like	3	66.7	99.3
	A+B; SH2-like	3	100.0	99.5
	A+B; Histidine-containing	2	100.0	99.5
	Multi; Sugar phosphatases	3	100.0	99.7
	Small; Small inhibitors, toxins, lectins	14	71.4	91.4
	Small; BPTI-like	3	66.7	99.7
	Small; EGF-like module	4	75.0	99.7
Percent Composition of amino acids	Alpha; DNA-binding 3-helical bundle	12	66.7	99.7
	Alpha; lambda repressor-like DNA bind.	5	60.0	99.7
	Alpha; EF-hand	6	100.0	100.0
	Beta; Immunoglobulin-like	30	66.7	88.6
	Beta; Viral coat and capsid proteins	16	75.0	97.6
	A/B; Periplasmic bind. protein-like	11	63.6	92.4
	Small; Metallothionein	3	100.0	100.0
Hydrophobicity	A/B; PLP-dependent transferases	3	66.7	98.7
	A/B; Periplasmic binding protein-like	11	72.7	93.1
	Small; Small inhibitors, toxins, lectins	14	71.4	94.8
Van der Waals volume	Alpha; DNA-binding 3-helical bundle	2	66.7	92.9
	Alpha; Pheromone proteins	3	66.7	97.0
	Alpha; Ferritin like	5	60.0	98.5
Polarizability	Alpha; Pheromone proteins	7	66.7	99.7
	A/B; beta/alpha (TIM)-barrel	29	62.1	82.5
Polarity	A/B; beta/alpha (TIM)-barrel	29	62.1	85.6
	Small; Classic zinc finger	3	66.7	99.7
	Small; Metallothionein	3	100.0	98.3
Alpha-frequency	Alpha; lambda repressor-like DNA bind.	5	60.0	99.7
	Beta; Viral coat and capsid proteins	16	62.5	89.8
	Small; Small inhibitors, toxins, lectins	14	71.4	96.5

After testing two numbers were calculated for combined testing sets - 1) the percentage of correct positive predictions (number of class members correctly assigned to its class) and 2) the percentage of correct negative predictions or rejection accuracy (the number of proteins from the group 'others' correctly not assigned to the class). This procedure was repeated for each class in terms of each attribute.

The number of classes predicted at a 60% and higher accuracy level totaled 25 for all attributes (Table 2). Among them 18 classes were predicted by only one attribute, four classes by two attributes, and three classes by three attributes.

The largest number of classes (14) were predicted by PHD_SS. Among larger classes the best prediction was made for the class of Globins, with a high positive accuracy (84.6 %) and 100 % rejection accuracy . Classes with a small number of proteins (2 - 5) also demonstrate an extremely high level of rejection accuracy (99.3 - 100%). Other bigger classes - Immunoglobulin-like beta-sandwich, (TIM)-barrel, FAD - binding motif, and Small inhibitors - have a much lower rejection accuracy (80,2 - 91.4%), and a positive accuracy in the range of 63.6 - 71.4%.

The percent composition of amino acids provides a significant correlation with the broad structural class of proteins (Chou 1989). Our earlier work (Mayoraz, Dubchak & Muchnik 1995) showed that percent composition of amino acids possesses certain predictive power for much more detailed classification . As seen from the Table 2, the percent composition of amino acids performed well on 7 classes, 6 of them having 5 or more proteins.

The hydrophobicity attribute worked satisfactorily on three classes. Two of them were alpha/beta classes (PLP-dependent transferases and Periplasmic binding protein-like) not predicted by any other attribute. Four other attributes (the normalized van der Waals volume, polarity, polarizability of amino acid, and alpha frequency) worked satisfactorily on eleven classes altogether, among them three - Pheromone proteins, Classic zinc finger, and Ferritin were also uniquely predicted.

This work demonstrates that a specific attribute works differently on different classes. It is necessary to emphasize the importance of finding an individual set of descriptors which works best among all others on a particular folding class in the comprehensive classification. The study shows a possibility of such a solution. A number of new physical, chemical, and structural properties including various hydrophobicity scales, as well as different types of protein sequence descriptors should be studied in order to increase the number of classes for the recognition. Cross-validation and blind testing will be necessary to develop a final prediction scheme.

Acknowledgments. The authors thank Nikolai N. Alexandrov for providing them with the protein database, Stanley Goldman for helpful discussions and Chris Mayor for his assistance in the preparation of the manuscript.

References

Charton, M.; and Charton, B. I. 1982. The structural dependence of amino acid hydrophobicity parameters. *J. Theor. Biol.* 99: 629-644.

Chothia, C.; and Finkelstein, A. V. 1990. The classification and origins of protein folding patterns. *Annu. Rev. Biochem.* 59: 1007-1039.

Chou, P. Y. 1989. Prediction of Protein Structure and Principles of Protein Conformation. New York, Plenum Press. 549-586.

Chou, P. Y.; and Fasman, G. D. 1978. Prediction of the secondary structure of proteins from their amino acid sequence. *Adv. Enzymol.* 47: 45-148.

Dubchak, I.; Muchnik, I.; Holbrook, S. R.; and Kim, S.-H. 1995. Prediction of protein folding class using global description of amino acid sequence. *Proc. Natl. Acad. Sci. USA* 92: 8700-8704.

Fauchere, J. L.; Charton, M.; Kier, L. B.; Verloop, A.; and Pliska, V. 1988. Amino acid side chain parameters for correlation studies in biology and pharmacology. *Int. J. Peptide Protein Res.* 32: 269-278.

Grantham, R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185: 862-864.

Hobohm, U.; and Sander, C. 1994. Enlarged representative set of proteins. *Protein Science* 3: 522-524.

Mayoraz, E.; Dubchak, I.; and Muchnik, I. 1995. Relation between protein structure, sequence homology and composition of amino acids. Third International Conference on Intelligent Systems for Molecular Biology., 240-248, Cambridge, UK: AAAI/MIT Press.

Murzin, A. G.; Brenner, S. E.; Hubbard, T.; and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Molec. Biol.* 247: 536-540.

Muskal, S. M.; and Kim, S.-H. 1992. Predicting protein secondary structure content: a tandem neural network approach. *J. Mol. Biol.* 225: 713-727.

Orengo, C. A.; Flores, T. P.; Taylor, W. R.; and Thornton, J. M. 1993. Identification and classification of protein fold families. *Prot. Engng.* 6: 485-500.

Rost, B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. *Methods Enzymol.* 266: 545-549

Pascarella, S.; and Argos, P. 1992. A data bank merging related protein structures and sequences. *Prot. Engng.* 5: 121-137.

Rost, B.; and Sander, C. 1993. Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* 232: 584-599.

Selbig, J.; Kaden, F.; and Koch, I. 1992. Applying machine learning methods for finding significant amino acid properties in proteins. *FEBS* 3: 241-246.

Tomii, K.; and Kanehisa, M. 1996. Analysis of amino acid indices and mutation matrices for sequence comparison and structure prediction of proteins. *Prot. Eng.* 9: 27-36.