Some New Features for Protein Fold Prediction

Nikhil Ranjan Pal and Debrup Chakraborty

Electronics and Communication Sciences Unit
Indian Statistical Institute, Calcutta 700108 , India
{nikhil,debrup_r}@isical.ac.in

Abstract. In this paper we propose several sets of new features for protein fold prediction. The first feature set consisting of 47 features uses only the sequence information. We also define four different sets of features based on hydrophobicity of amino acids. Each such set has 400 features which are motivated by folding energy modeling. To define these features we have considered pair-wise amino acids (AA) interaction potential. The effectiveness of the proposed feature sets is tested using multilayer perceptron and radial basis function networks to solve the 4 class (level 1) and 27 class (level 2) prediction problems as defined in the context of SCOP classification. Our investigation shows that such features have good discriminating powers in predicting protein folds.

1 Introduction

One of the most important and challenging problems of bioinformatics is prediction of protein folds from the amino acid sequences. Researchers have been using machine learning techniques for solving many problems of bioinformatics including the prediction of protein folds[1-4]. Also, there have been several attempts to predict local secondary structures of proteins[2]. Success of any such method depends on the features used to characterize a residue sequence representing a protein. Among other tools, neural networks are the most successful ones for prediction of protein structures. Dubchak et al. [1] point out that when we want a broad structural classification of protein, say into four classes, all alpha, all beta, (alpha+beta) and (alpha/beta) it is easy to get more than 70% prediction accuracy using simpler feature vector for representing a protein sequence. However, the problem becomes more and more difficult as we demand more refined classification into more classes.

So far one of the most successful set of features used for protein folds consists of global description of the chain of amino-acids representing proteins. In this feature set different properties of the amino acids are used as features. For example, they used the relative hydrophobicity of amino acids. They also used information about the predicted secondary structure and predicted solvent accessibility. Here we do not like to use the predicted secondary structure because it has, inherent in it, some incorrect information. In other words, about 30% of the predicted secondary structure is incorrect. Thus it would not be meaningful to use, the predicted secondary structure as a feature. It would be better if we

can use properties of the residue sequence to directly predict the folds. And that is the objective of this paper.

Here we propose five sets of new features and evaluate their performance using neural networks. We compute some features which characterize the spatial distribution of different amino acids on the sequence. For example, for a particular residue, we compute the average separation between two successive occurrences of the same residue. We also compute some entropy based features. In this way, just based on the symbols, not using any of their physico-chemical properties, we computed 47 features and this set alone is found to produce a test prediction accuracy of about 74%. We also computed four sets of features each having 400 members based on hydrophobicity of the amino acids. Here, we have considered amino acid pairs separated by just one position and taken an exponential function of the hydrophobocity of the pair to compute the feature. This is motivated by the fact that in case of structure prediction by ab-initio method hydrophobicity has been successfully used [5]. Moreover, for such approaches researchers have considered pairs, which are in "contact" but separated by at least one residue to avoid complete collapse. Out of these four sets of features two works quite well both for both level 1 and level 2 classification tasks. In fact for classification into 27 folds, both these feature sets outperform the classification accuracy reported in the literature using the popular 125 features based on the various physico-chemical properties and predicted secondary structure of the amino acid sequences.

2 Features for Protein Fold Prediction: Old and New

2.1 Some existing features

One of the most successful set of features for protein fold prediction contains 125 features which are extensively used in the context of SCOP classification [6]. These features are computed using the following characteristics of proteins: composition, predicted secondary structure, hydrophobicity, volume, polarity and polarizability. The amino acids (AAs) are divided into three groups based on hydrophobicity, volume, polarity, polarizability and predicted secondary structure, as shown in Table 1 [1,4]. Now a protein, that is an amino acid sequence, is described using three global descriptors [1,4] : Composition (C), Transition (T) and Distribution (D). These descriptors essentially describe the frequencies with which the properties change along the sequence and their distribution on the chain. Let us illustrate it using hydrophobicity as an example. As stated earlier, based on the hydrophobicity values, AAs are divided into three groups, polar (P), neutral (N) and hydrophobic (H). Then C, the composition descriptor consists of three values giving the percentage of the three types of AAs in the protein. Transition feature is also characterized by three transition probabilities: transition probability of P to N and N to P; transition probability of P to H and H to P and that from N to H and H to N. The computation of the feature values representing distribution is a little complex. Here for each of the three

groups of AAs, five percentages are computed. These five values are: the percentage of the sequence where the first member of that group is located, and the fractions where the first 25%, 50%, 75% and 100% residues of that group are contained. In this way, based on just hydrophobicity we get 3+3+15= 21 feature values. So using five properties of AAs one gets $5 \times 21 = 105$ features and an additional 20 features representing the percentage compositions of amino acids in the protein. Thus in total one gets 125 features. Authors in [4] used various combination of these features. The data sets with these features are available at http://www.nersc.gov/~cding/protein. The 125 features and various subsets of them have been extensively used by researchers for prediction of protein folds.

Table 1. Grouping of Amino Acids based on attributes (an extended version of Table 1 in [1])

Property	Group 1	Group2	Group 3
Hydrophobicity	Polar	Neutral	Hydrophobic
	R,K,E,D,Q,N	G,A,S,T,P,H,Y	C,V,L,I,M,F,W
Volume	0 - 2.78	2.95 - 4.0	4.43 - 8.08
Polarity	4.9 - 6.2	8.0 - 9.2	10.4 - 13.0
Polarizability	0.00 - 0.108	0.128 - 0.186	0.219 - 0.409
Predicted Secondary Structure	Helix	Strand	Coil

2.2 Some New Features

Since, the native fold of a protein depends only on the residue sequence, we should be able to do a good job using just the sequence information. Keeping this in mind we shall talk about five types of features. The first set of features does not explicitly use any attribute of the AAs but is based on distribution of the residues on the chain. Let us denote the 20 residues by $x_i, i = 1, 2, ..., 20$ and their frequencies by $f_i, i = 1, 2, ..., 20$. Let N be the length of a residue sequence representing a protein $S = s_1\ s_2\ s_3 ... s_N$. Define $P = \{p_i = f_i/N : i = 1, 2, , n\}$, where p_i is the probability of residue x_i on S.

Our first set of 20 features is $F_i = p_i, i = 1, 2, ..., 20$. These 20 features have been used by other researchers also. Next we compute five features that try to characterize the shape of the histogram. These features are summarized in Table 2. The *first order energy* (F_{21}, Table 2) attains the minimum value for a uniform distribution, while the *first order entropy* (F_{22}, Table 2) attains the maximum value for a uniform distribution. The features $F_{23} - F_{25}$ characterize the shape of the histogram. The *second order energy* (F_{26}) and the *second order entropy* (F_{27}) measure the uniformity of distribution of pairs of residues. The remaining 20 features compute the average separation between two successive appearance of the same residue on the AA chain. Note that, the denominator of the *average*

separation of the residues makes it independent of the length of the sequence and the frequency with which the residue type occurs. These 47 features, $F_1 - F_{47}$, constitute our first set of features.

We also compute other sets of features based on the hydrophobicity attribute of the AAs. It is believed that hydrophobicity characteristic of residues is a very important determinant of native structure of a protein. Consequently, hydrophobicity has been used by researchers for threading [5]. Motivated by this fact, we have generated features characterizing interaction between pairs of residues in contact. We consider two residues are in contact if they are separated by exactly one residue on the AA chain. So we compute the interaction potential between two residues (a, b) as

$$P_1(a,b) = \frac{1}{R} \sum_{x_i=a, x_j=b, j=i+2} e^{\frac{(h(x_i)+h(x_j))}{M}}, a, b = 1, 2, ..., 20.$$

Here $h(a)$ is the hydrophobicity of residue a and M is a constant, the role of M is to scale the value of hydrophobicity so that numerical overflow is avoided. We choose M as 24 because the maximum absolute value of hydrophobicity is 12. The same M will be used for scaling the test data, X_{Ts}. R is a normalizing constant defined as :

$$R = max_{S \in X_{Tr}} \left\{ \max_{a,b} \sum_{x_i=a, x_j=b, j=i+2} e^{\frac{(h(x_i)+h(x_j))}{M}} \right\}.$$

Note that, the normalizing constant R is computed taking into account the entire training set. In this way we shall get a feature vector of dimension 400 for each protein sequence. Researchers using energy modeling for threading or ab-initio folding do not consider adjacent residues to avoid complete collapse of the sequence. Keeping this in mind, the feature set P_1 does not consider adjacent residues. However, since we are not using energy modeling, but shall do feature based pattern recognition, we also conducted experiments considering another set of features P_2 which is computed exactly in the same manner as P_1, but taking into account only adjacent pairs of residues.

In P_1 and P_2 we defined the interaction potential for a pair of residues (a, b) in protein S as the sum of interaction potential of every occurrence of the (a, b). However, the potential energy of an ensemble is usually computed as the sum of pair-wise interaction. Keeping this in view, we propose another two sets of features P_3 and P_4 as defined in Table 2. P_3 considers pairs of residues separated by one residue while P_4 considers all adjacent pairs. Note that, P_3 and P_4 are normalized using a different constant Q. We choose

$$Q = max_{S \in X_{Tr}} \left\{ \max_{a,b} \sum_{x_i=a, x_j=b, j=i+1} (h(x_i) + h(x_j)) \right\},$$

X_{Tr} is the training data set. In other words, we find the maximum possible exponent over the entire training data.

Table 2. The new features

Histogram	$F_i = p_i, i = 1, 2, ..., 20$
First order energy	$F_{21} = \sum_{i=1}^{20} p_i^2$
First order entropy	$F_{22} = \sum_{i=1}^{20} -p_i log\ p_i$
Histogram difference	$F_{23} = \sum_{i=1}^{19} \lvert p_i - p_{i+1} \rvert$
Weighted histogram difference-1	$F_{24} = \sum_{i=1}^{19} \lvert p_i - p_{i+1} \rvert p_{i+1}$
Weighted histogram difference-2	$F_{25} = \sum_{i=1}^{19} \lvert p_i - p_{i+1} \rvert p_i$
Second order energy	$F_{26} = \sum_{i=1}^{20} \sum_{j=1}^{20} p_{ij}^2$
Second order entropy	$F_{27} = -\sum_{i=1}^{20} \sum_{j=1}^{20} p_{ij} log\ p_{ij}$
Average separation of residues	$F_{i+27} = \frac{1}{N f_i} \sum_{s_j = x_i, s_k = s_i, s_l \neq x_i, \forall j<l<k} (j-k), \quad if \quad f_i \neq 0$ $= 0, \quad otherwise, \quad i = 1, ...20$
Pairwise interaction Potential -type1	$P_1(a,b) = \frac{1}{R} \sum_{x_i = a, x_j = b, j = i+2} e^{\frac{(h(x_i)+h(x_j))}{M}}, \ a, b = 1, ..., 20$
Pairwise interaction Potential -type2	$P_2(a,b) = \frac{1}{R} \sum_{x_i = a, x_j = b, j = i+1} e^{\frac{(h(x_i)+h(x_j))}{M}}, \ a, b = 1, ..., 20$
Pairwise interaction Potential -type3	$P_3(a,b) = e^{\frac{1}{Q} \sum_{x_i = a, x_j = b, j = i+2} (h(x_i)+h(x_j))}, \ a, b = 1, ..., 20$
Pairwise interaction Potential -type4	$P_4(a,b) = e^{\frac{1}{Q} \sum_{x_i = a, x_j = b, j = i+1} (h(x_i)+h(x_j))}, \ a, b = 1, ..., 20$

3 Results

In our experiments we use a set of 698 proteins divided into 313 training and 385 test instances as used by Dubchak et al. [4]. The training data set do not have proteins with more than 35% of sequence identity for aligned subsequences [4]. Similarly, the test data contains SCOP sequences having less than 40% identity with each other. Also it does not contain any sequence with more than 35% identity with the training data. There are two levels of classifications of this data set. First, a coarse classification into four levels as shown in Table 3 are available. Each of these four classes are then further classified. All-alpha is classified into 6 folds, all-beta is classified to 9 classes, alpha/beta and alpha+beta are respectively grouped to 9 and 3 classes resulting in total 27 folds. In this investigation we shall consider classification at both levels using neural networks as the machine learning tools. We shall use both the multilayer perceptron and radial basis function network as the classifiers.

Table 3. Number of patterns in training and test sets

Fold Types	Number of Training Pattern (X_{Tr})	Number of Test Pattern (X_{Ts})
All Alpha	55	61
All Beta	109	117
Alpha/Beta	115	145
Alpha+Beta	34	62
Total	313	385

Table 4 presents the performance of the proposed features on the training and test data sets when MLP is used as the classifier. In order to compare the performance of the proposed features we also trained networks with the 125 features of Dubchak et al. [1,4]. We call this feature set D_{125}. D_{125} uses 21 features based on predicted secondary structures. We excluded these 21 features and generated a feature set named D_{104}. As we mentioned earlier that use of features which are guaranteed to have about 30% incorrect information is not desirable. This is the reason for considering D_{104}. We also experimented with this feature set. Table 4 shows that the feature set $F_1 - F_{47}$ have considerable discriminatory power, giving a classification accuracy of nearly 74%. Of these five new feature sets, the feature set P_3 seems to be the best.

The prediction accuracy obtained by the RBF networks is depicted in Table 5. We see that using RBF also P_3 performs the best among all the features we calculated. But, still while classifying into 4 folds D_{125} seems to be the best.

Next we consider the problem of detailed classification into 27 folds. For this we use here only RBF networks. Table 6 depicts effectiveness of different feature sets in conjunction with RBF networks (as implemented in MATLAB neural network toolbox). Table 7 reports the results using D_{125} by MLP, General

Table 4. Performance of MLP on different feature sets for first level classification

Feature sets	Network Size	Training error	Test error
$F_1 - F_{47}$ (47 features)	47:50:20:10:4	3.83%	26.75%
P_1 (400 features)	400:80:40:10:4	8.62%	31.60%
P_2 (400 features)	400:80:40:10:4	12.46%	31.90%
P_3 (400 features)	400:80:40:10:4	1.91%	26.49%
P_4 (400 features)	400:80:40:10:4	0.00%	29.35%
D_{125} (125 features)	125:80:40:10:4	0.00%	19.48%
D_{104} (104 features)	125:80:40:10:4	3.83%	22.80%

Regression Neural Network (GRNN) and support vector machines (SVM). Table 7 have been adapted from [7]. Comparing Tables 6 and 7 it is clear that in the 2nd level of classification, i.e., classification into 27 folds, P_3 outperforms D_{125} for all four classifiers tried. The performance of P_4 is marginally less than that obtained by using SVM on D_{125} but is better than D_{125} for all other classifiers tried. The performance of P_1 and P_2 is also quite good .

Table 5. Performance of RBF on different feature sets for first level classification

Feature sets	Network Size	Training error	Test error
$F_1 - F_{47}$ (47 features)	40	19.80%	29.35%
P_1 (400 features)	150	3.19%	23.89%
P_2 (400 features)	150	2.55%	25.71%
P_3 (400 features)	150	1.91%	22.33%
P_4 (400 features)	150	2.23%	25.71%
D_{125} (125 features)	100	1.27%	18.18%
D_{104} (104 features)	80	8.31%	27.01%

4 Conclusions

We have proposed five sets of features that are defined only using characteristics of the residues and the distribution of residues on the AA sequence representing a protein. As the learning machine, we used multilayer perceptron network and radial basis function network. We used the SCOP database and considered classification to four folds and twenty seven folds. Our experimental results revealed that the proposed features have reasonably good discrimination power. Of the five sets of features P3 and P4 are more effective than the other three types of features. Our investigation also revealed that while computing interaction potential type features use of adjacent pairs and pairs separated by exactly one residue produces more or less the same performance. Our next step would be to

Table 6. Performance of RBF on different feature sets for classification into 27 folds

Feature sets	Network Size	Training error	Test error
$F_1 - F_{47}$ (47 features)	40	39.3%	57.50%
P_1 (400 features)	150	4.79%	50.64%
P_2 (400 features)	150	6.38%	51.42%
P_3 (400 features)	150	6.70%	45.97%
P_4 (400 features)	150	5.43%	48.83%
D_{125} (125 features)	100	10.86%	49.87%
D_{104} (104 features)	80	14.37%	51.68%

Table 7. The Protein-fold prediction error using D_{125} for level 2 classification[7]

Method	MLP	GRNN	RBFN	SVM
Classification accuracy	51.2%	55.8%	50.6%	48.6%

combine these feature sets and use some connectionist online feature selection scheme to select the best set of features[8, 9].

References

1. I. Dubchak, I. Muchnik, C. Mayor, I. Dralyuk and S-H Kim, "Recognition of a Protein Fold in the context of the SCOP Classification. PROTEINS: Structure, Function and Genetics, vol. 35, pp. 401-407, 1999.
2. P. Baldi and S. Brunak, Bioinformatics: the Machine Learning Approach, MIT Press, 1998.
3. I. Dubchak, I. Muchnik, S. R. Holbrook, and S. H. Kim, "Prediction of protein folding class using global description of amino acid sequence", Proc. Natl. Acad. Sci., USA, Vol. 92, pp. 8700-8704, 1995.
4. I. Dubchak and C. H. Q. Ding, "Multi-class protein fold recognition using support vector machines and neural networks," Bioinformatics, Vol. 17, No. 4, pp. 349-358, 2001.
5. Antônio F. Pereira de Araújo, "Folding protein models with simple hydrophobic energy function: the fundamenta importanve of monomer inside/outside segregation", Proc. Natl. Acad. Sci., USA, vol 96, no 22, pp. 12482-12487.
6. A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia, "SCOP: A structural classification of proteins database for the investigation of sequence and structures. Journal of Molecular Biology, vol. 247, pp. 536-540, 1995.
7. I-Fang Chung, Chuen-Der Huang, Ya-Hsin Shen and Chin-Teng Lin, "Recognition of Structure Classification of Protein Folding by NN and SVM Hierarchical Learning Architecture, Proceedings of ICONIP 2003.
8. N.R. Pal and K.K. Chintalapudi, "A connectionist system for feature selection", Neural, Parallel & Scientific Computations, vol 5. No. 3, 359-381, 1997.
9. D. Chakraborty and Nikhil R. Pal, "Integrated feature analysis and fuzzy rule-based system identification in a neuro-fuzzy paradigm", IEEE Trans. on Systems Man Cybernetics B, vol 31, no 3, pp. 391-400, 2001.