2007－BIO－8 (2)
2007／3／5

タンパク質フォールド予測におけるアミノ酸残基頻度/比率の比較

田口善弘[1], マイケル グロミハ[2]

[1] 中央大学理工学部物理学科,[2] 産業技術総合研究所生命科学研究センター

タンパク質の3次元構造をアミノ酸残基配列から推定することは計算/ 分子生物学の重要な課題である。タンパク質の3次元構造はフォールドと呼ばれている類似構造ごとに分類できることが知られているが、どのタンパク質がどのフォールドに属するかを判別することができれば、3次元構造推定の精度の向上に大いに有用である。本研究では、アミノ酸残基の「数」を用いた線形判別分析で、30種の異なったフォールドに属する1612種のタンパク質がどのフォールドに属するかを予測する手法を提案する。この方法では、主な30種類のフォールドに属する球状蛋白質のフォールドを敏感度(Sensitivity) 37%で判別が可能であり、この値は従来の方法による値と同等、もしくは、よりよいことが解った。

Comparison of amino acid occurrence and composition for predicting protein folds

Y-h. Taguchi[1], M. Michael Gromiha[2]

[1]Dept. Phys., Chuo. Univ., tag@granular.com
[2]CBRC, AIST, Japan, michael-gromiha@aist.go.jp

Prediction of protein three-dimensional structures from amino acid sequences is a long-standing goal in computational/molecular biology. The successful discrimination of protein folds would help to improve the accuracy of protein 3D structure prediction. In this work, we propose a method based on linear discriminant analysis (LDA) for recognizing proteins belonging to 30 different folds using the occurrence of amino acid residues in a set of 1612 proteins. The present method could discriminate the globular proteins from 30 major folding types with the sensitivity of 37%, which is comparable to or better than other methods in the literature.

Background

Deciphering the native conformation of a protein from its amino acid sequence known as, protein folding problem is a challenging task. The recognition of proteins of similar folds and/or proteins belonging to same structural class is a key intermediate step for protein structure prediction. For the past several decades several methods have been proposed for predicting protein structural classes. These methods include discriminant analysis[1], correlation coefficient[2], hydrophobicity profiles[3], amino acid index [4], Bayes decision rule[5], amino acid distributions[6], functional domain occurrences [7], supervised fuzzy clustering approach[8], amino acid principal component analysis[9] etc. These methods showed that the sensitivity lies in the range of 70-100% for discriminating protein structural classes and the sensitivity mainly depends on the dataset. Wang and Yuan[5] developed a dataset of 674 globular protein domains belonging to different structural classes and reported that methods claiming 100% sensitivity for structural class prediction, predicted only with the sensitivity of 60% with this dataset.

On the other hand, alignment profiles have been widely used for recognizing protein folds[10, 11]. Recently, Cheng and Baldi[12] proposed a machine learning algorithm for fold recognition using secondary structure, solvent accessibility, contact map and β-strand pairing, which showed the pairwise sensitivity of 27%. On the other hand, it has been reported that the amino acid properties are the key determinants of protein folding and are used for discriminating membrane proteins[13], identification of membrane spanning regions[14], prediction of protein structural classes[15], protein folding rates [16], protein stability[17] etc. Towards this direction, Ding and Dubchak[18] proposed

	with re-weighting		without re-weighting	
	over all	fold average	over all	fold average
Occurrence	0.33	0.32	0.37	0.28
Composition	0.26	0.27	0.32	0.24

Table 1: Leave-one-out cross validation results for two types of sensitivities.

a method based on neural networks and support vector machines for fold recognition using amino acid composition and five other properties, and reported a cross-validated sensitivity of 45 %.

In this work, we have used the amino acid occurrence (not composition) of proteins belonging to 30 major folds for recognizing protein folds. We have developed a method based on linear discriminant analysis (LDA), which showed an accuracy of 37% in recognizing 1612 proteins from 30 different folds, which is comparable with other methods in the literature, in spite of the simplicity of our method and the large number of proteins considered.

Results and Discussion

Role of re-weighting for fold recognition

We have computed the occurrence of all the 20 amino acid residues in each protein. The occurrence of 20 types of residues represents the elements of 20 dimensional vectors for each protein. We have applied LDA to these vectors for recognition. Here, we have employed two kinds of LDA, i.e., with and without reweighing. In LDA with re-weighting, each fold equally contributes to the measure of performance irrespective of the number of proteins in each fold; i.e., even if one fold includes hundreds of proteins and another has only few proteins, LDA is optimized to achieve the highest performance equally in each fold. This re-weighting is important especially when the number of proteins included in each fold has large variations.

On the other hand, LDA without re-weighting, tends to achieve the maximum sensitivity for the whole dataset. In this case, folds with less number of proteins have a strong tendency to be ignored. In accordance with this choice, we have defined two kinds of sensitivities, (i) averaged over folding types and (ii) overall. The overall sensitivity is the ratio between the number of correctly predicted proteins (true positives) in each fold and total number of proteins. Folding type sensitivity is computed as the average of sensitivities obtained in each fold.

In Table 1, we presented two types of sensitivities (overall and fold average) with two kinds of LDA (with and without re-weighting). We observed that re-weighting significantly changed the performance. This is due to the divergence in the number of proteins in each fold (min. 25, max. 173, mean 54, see Table 2). Two kinds of sensitivities differ from each other by almost 10%, without re-weighting. We achieved the sensitivity of 37%, which is the best performance to our knowledge, for large number of folds (30) and proteins (1612) considered. Further, the method is extremely simple, which indicates that the physical properties of proteins carry sufficient information instead of sequences.

Prediction of proteins belonging to different folding types

We have examined the ability of the present method for predicting proteins belonging to 30 major folds. In Table 2, we have shown the sensitivity of recognizing 30 different folds. We observed that the sensitivity of folds with fewer proteins has increased after re-weighting. All the folds that have the sensitivity of less than 10 % without re-weighting are the ones with fewer proteins. For example, SAM domain like fold has the sensitivity of 7%, which has only 26 proteins. Similar tendency is also observed for the folds b.2, b.34, c.3, c.47, c.55, d.15 and d.17. On the other hand, many folds

ID	Fold	Fold Description	Number	Sensitivity(%)		
				without re-weighting	with re-weighting	hierarchical
		all-α				
1	a.3	Cytochrome C	25	24	48	48
2	a.4	DNA/RNA binding 3-helical bundle	103	72	49	33
3	a.24	Four helical up and down bundle	26	23	39	34
4	a.39	EF hand-like fold	25	40	44	44
5	a.60	SAMdomain-like	26	7	27	19
6	a.118	α-α superhelix	47	46	45	46
		all-β				
7	b.1	Immunoglobulin-like β-sandwich	173	76	38	13
8	b.2	Common fold of diphtheria toxin/transcription factors/cytochrome f	28	3	29	28
9	b.6	Cupredoxin-like	30	26	37	30
10	b.18	Galactose-binding domain-like	25	20	36	36
11	b.29	Concanavalin A-like lectins/glucanases	26	23	27	26
12	b.34	SH3-like barrel	42	0	28	7
13	b.40	OB-fold	78	21	24	7
14	b.82	Double-stranded a-helix	34	11	18	20
15	b.121	Nucleoplasmin-like	42	52	52	50
		α/β				
16	c.1	TIM barrel	145	44	26	25
17	c.2	NAD(P)-binding Rossmann-fold domains	77	33	31	29
18	c.3	FAD/NAD(P)-binding domain	31	9	16	16
19	c.23	Flavodoxin-like	55	10	5	3
20	c.26	Adenine nucleotide a hydrolase-like	34	11	29	26
21	c.37	P-loop containing nucleoside triphosphate hydrolases	95	43	33	32
22	c.47	Thioredoxin fold	32	9	18	9
23	c.55	Ribonuclease H-like motif	49	4	6	4
24	c.66	S-adenosyl-L-methionine-dependent methyltransferases	34	29	29	29
25	c.69	α/β-Hydrolases	37	35	40	40
		$\alpha+\beta$				
26	d.15	β-Grasp, ubiquitin-like	42	4	21	35
27	d.17	Cystatin-like	25	0	8	20
28	d.58	Ferredoxin-like	118	32	7	16
		small				
29	g.3	Knottins	80	97	88	88
30	g.41	Rubredoxin-like	28	10	71	85

Table 2: Leave-one-out cross validation sensitivity in each fold.

with less than 30 proteins have the sensitivity of more than 20% even without re-weighting (e.g., a.3, a.24, a.39 etc.). As there are 30 folds, the expected sensitivity is only 3.3 % if classification is supposed to be random. Hence the sensitivity of 20% obtained for several folds is significantly higher than that of random for fold recognition. Interestingly most of the folds, which have more than 20% sensitivity, in spite of less number of proteins, belong to either all-α or all-β. This might be due to the fact that the proteins belonging to all-α and all-β classes have different secondary structural patterns and hence they are easy to discriminate them. In addition, folds in these classes are near-by each other in amino acid occurrence vector space, which caused high sensitivity. The comparison between experimental vs predicted folds is shown in Fig. 1. In this figure, dark block indicates the presence of relatively higher number of proteins and the data are normalized so that the total percentage of true fold is 100 %. We noticed that before re-weighting (Fig. 1(a)), the folds, to which many proteins are misclassified, are the ones with more number of proteins (e.g., a.4, b.1, c.1 and d.58). On the other hand, after re-weighting, the trend has been changed: the misclassified proteins mainly accommodates within the same structural class. Especially, in $\alpha+\beta$, the block diagonal region is filled almost uniformly, which is partially caused by re-weighting. Since each fold is equally weighted, $\alpha+\beta$ class is less weighted than other classes. This causes inter-class misclassification between $\alpha+\beta$ and other classes, because $\alpha+\beta$ class includes only three folds. This problem can be clearly seen in Table 3(a) where we have shown true vs predicted classes with

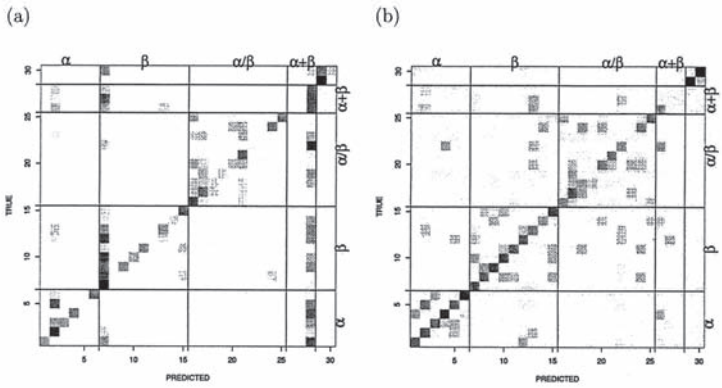


Figure 1: Comparison between predicted and experimental folds in 1612 proteins. The diagonal elements show the correctly predicted proteins. Dark block indicates the presence of more number of proteins and solid line indicates boundary between five classes as shown in Table 2, i.e., all-α, all-β, α/β, and $\alpha+\beta$ and small proteins. (a)without reweighing. (b) with reweighing.

re-weighting. Here, the classes are not evenly sampled and α/β class keeps almost three times as large as $\alpha+\beta$. Further, neither all-α nor all-β are mainly misclassified into $\alpha+\beta$.

Hierarchical re-weighting

In order to resolve this problem, we proposed the scheme of hierarchical re-weighting. In this method, weight is equally distributed to 5 classes (all-α, all-β, α/β, $\alpha+\beta$, and small) then it is re-distributed to each fold. For example, fold c.37 gets 0.02 weight since it gets one tenth of weight 0.2, which is delivered to α/β. The results obtained with hierarchical re-weighting is also included in Table 2. The comparison of results obtained with and without re-weighting showed that the folds with less number of proteins increase the sensitivity after re-weighting and vice-versa. However, the trend is different between simple and hierarchical re-weighting. For example, although all folds in all-α class have the same weight with hierarchical re-weighting only three folds (a.3, a.39, and a.118) have similar or better sensitivity compared with simple re-weighting. On the other hand, sensitivity of fold a.4 drastically decreased from 49% to 33%. This might be due to the fact that several proteins belonging to all-α, all-β and α/β proteins are misclassified into $\alpha+\beta$. Further, the data presented in Table 3(b) showed that folds belonging to both all-α and all-β classes are misclassified into $\alpha+\beta$ class.

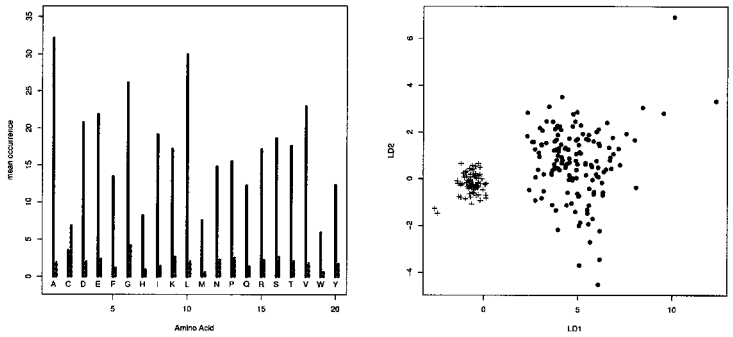
(a) simple reweighting

	predicted					
True	all-α	all-β	α/β	$\alpha+\beta$	small	total
all-α	178	32	13	20	9	252
all-β	41	320	56	42	19	478
α/β	61	89	413	25	1	589
$\alpha+\beta$	54	34	34	44	19	185
small	3	3	0	1	101	108

(b) hierarchical reweighting

	predicted					
True	all-α	all-β	α/β	$\alpha+\beta$	small	total
all-α	144	4	3	84	17	252
all-β	20	180	27	203	48	478
α/β	62	61	372	90	4	589
$\alpha+\beta$	30	8	12	109	26	185
small	0	1	0	2	105	108

Table 3: Leave-One-out results of true vs predicted structural classes (a) with simple and (b) hierarchical re-weighting.



(a) (b)

Figure 2: (a)Comparison between mean amino acid occurrences of the most distant pair of folds, TIM barrel (black) and knottins (red). (b) Distribution of these two folds over the first two discriminant functions with re-weighting.

Comparison among different re-weighting procedures

The results presented in Tables 2 and 3 showed that the sensitivity of recognizing protein folds differs significantly between different prediction methods (without, simple and hierarchical re-weighting). Hence, it would be difficult to choose the best method for fold recognition. However, it may be selected based on the interest of the users, whether the prediction can be done for proteins that are within a specific structural class or whole dataset and/or obtaining the accuracy of each fold or overall.

Usually, training and test sets of data are obtained from sequence and structure databases and are culled with sequence identity. However, these datasets do not always reflect proper representatives of all proteins in different folds, e.g., protein population in each fold. Further, the proteins available in databases such as, PDB are biased with the proteins that can be solved experimentally, which may be different from the proportion of real proteins. Hence, considering these aspects would help to develop "good" methods for protein fold recognition in future.

In essence, based on the methods and datasets used in the present work, we suggest that the performance with simple re-weighting is better than that without and hierarchical re-weighting.

Influence of amino acid occurrence in recognizing protein folds

The importance of amino acid occurrence is illustrated with Figure 2(a).

In this figure we show the occurrence of the 20 types of amino acid residues in TIM barrel fold and knottins. We noticed that TIM barrel fold has eight alpha helices and eight beta strands and hence the occurrence of all the residues except Cys is higher than that of knottins. Knottin is a small protein and hence it has lower occurrence of all the residues and due to the importance of Cys it has more number of Cys residues than TIM barrel fold. In Figure 2(b), we have shown the distribution of residues in "amino acid occurrence" space. It is clearly seen that the two folds are separated well in this space. We observed similar results about the variation of amino acid occurrences among different folds in our data set.

In addition, we have tested the performance of the method using amino acid composition (i.e., amino acid occurrence/total number of residues) in each protein. We noticed that the overall sensitivity without re-weighting decreased to 32% indicating the importance of amino acid occur-

Fold Description	Number	Sensitivity (%)	
		without re-weighting	with re-weighting
Cytochrome C	16	56	94
DNA/RNA binding 3-helical bundle	32	75	56
Four helical up and down bundle	15	33	33
EF hand-like fold	15	53	53
Immunoglobulin-like β-sandwich	74	66	31
Cupredoxin-like	21	29	38
Concanavalin A-like lectins/glucanases	13	38	38
SH3-like barrel	16	0	50
OB-fold	32	16	28
TIM barrel	77	40	25
FAD/NAD: (P)-binding domain	23	22	30
Flavodoxin-like	24	8	13
NAD: (P)-binding Rossmann-fold domains	40	40	35
P-loop containing nucleoside triphosphate hydrolases	22	23	18
Thioredoxin fold	17	18	35
Ribonuclease H-like motif	22	5	18
α/β-Hydrolases	18	33	39
β-Grasp, ubiquitin-like	15	0	33
Ferredoxin-like	40	23	3
over all	532	36	32
fold average		30	35

Table 4: Predictive ability of our method to the independent dataset of proteins used in Ding and Dubchak[18].

rence (un-normalized composition) in each fold (Table 1). Similar tendency is also observed for discriminating β-barrel membrane proteins[16]. Hence, we suggest to use un-normalized composition for better prediction results. In fact, the normalization of amino acid composition produced the problem of co-linearity, i.e., diversity of vectors is not sufficient compared with the number of proteins.

Comparison with other methods

We have compared the performance of our method with other related works in the literature. Ding and Dubchak[18] introduced a combined method for predicting the folding type of a protein. They have used six parameters, amino acid composition, secondary structure, hydrophobicity, van der Waals volume, polarity and polarizability as attributes, and neural networks and support vector machines for recognition. The features have been combined with the number of votes in each method. They reported the sensitivity of 56% in a test set of 384 proteins and 10-fold cross validation sensitivity of 45% in a training set of 311 proteins from 27 folding types. We have used the same dataset of 311 proteins and assessed the performance of our method. We observed that our method could predict with the leave-one-out cross validation accuracy of 44%, which is similar to that (45%) reported in Ding and Dubchak[18].

In addition, we have selected the proteins from the folds that are common in both the studies and tested the performance of our method (trained with our dataset of 1612 proteins) in predicting the folding types of the proteins used in Ding and Dubchak[18]. The results are presented in Table 4. Interestingly, our method could predict the proteins belonging to cytochrome C fold to the sensitivity of 94 %. Further, our method with re-weighting could correctly identify the folding types with the sensitivity of more than 30 % in 13 among the 19 considered folds. The average sensitivity is similar to the one that we reported with the dataset of 1612 proteins. Although our method is optimized with different dataset it has the power to predict the folding type of independent dataset of proteins with similar sensitivity.

Further, there are several advantages in our method: (i) only one feature, amino acid occurrence

is sufficient for prediction rather than six features. The comparison of results obtained with only one feature showed that the performance of our method (45%) is significantly better than that of Ding and Dubchak[18] reported with amino acid composition (20-49%), (ii) voting procedure is not necessary and our method can be directly used for multi-fold classifications, (iii) our method uses LDA, which requires significantly less computational power compared with SVM. In SVM one has to diagonalize the matrix with the size of (protein number) $\times$ (protein number); on the other hand, LDA requires only diagonalization of 20 (the number of kinds of amino acid residues) $\times$ 20 matrix independent of number of proteins and (iv) although they have reported the dependency of fold specific sensitivities upon number of proteins in each fold, it is difficult to compensate this effect without modifying the complicated voting systems; our method has freedom to compensate it as shown in the previous sections.

Recently, Shen and Chou[19] reported better sensitivity for the same data set of Ding and Dubchak[18]. However, the results are biased with training set of data. We have evaluated the sensitivity of identifying proteins belonging to the folds, four helical up and down bundle (a.24) and EF hand-like (a.39) and we observed that the sensitivity is 30.5 % and 24 %, respectively. Our predicted accuracies (39 % and 44 %) are better than that of Shen and Chou[19].

Fold recognition on the web

We have developed a web server for recognizing protein folds from amino acid sequence. It takes the amino acid sequence as input and displays the folding type in the output. Further, the server has the feasibility of selecting the method, with, without and hierarchical re-weighting. It is freely available at http://granular.com/PROLDA/ [20].

Conclusions

In this paper, we have proposed a simple method for protein fold prediction, where both the number of folds and the number of proteins are extensive. Interestingly, the simplest method is the best method for the truly complicated problems. Although complicated methods have several possibilities for tuning they generate over fitting to the data set. Further, the simple method proposed in this work is better than or comparable to other complicated methods, such as, amino acid principal principal component analysis, neural networks and support vector machines proposed in the literature for fold recognition. In addition, our method has several advantages including the less computational time and classifying the folds at a single run rather than pairwise comparisons. We have developed a web server[20], which takes the amino acid sequence as the input and displays the folding type in the output.

Dataset

We have used a dataset of 1612 globular proteins belonging to 30 major folding types obtained from SCOP database[21] for recognizing protein folds. This dataset has been constructed with the following criteria: (i) there should be at least 25 proteins in each fold and (ii) the sequence identity between any two proteins is not more than 25%. The amino acid sequences of all the proteins are available at [20].

References

[1] Klein P: **Prediction of protein structural class by discriminant analysis.** *Biochim. Biophys. Acta.* 1986, **874**:205–215.

[2] Chou KC, Zhang CT: **Diagrammatization of codon usage in 339 human immunodeficiency virus proteins and its biological implication.** *AIDS Res. Hum. Retroviruses* 1992, **8**:1967–1976.

[3] Gromiha MM, Ponnuswamy PK: **Prediction of protein secondary structures from their hydrophobic characteristics**. *Int. J. Pept. Protein Res.* 1995, **45**:225–240.

[4] Bu WS, Feng ZP, Zhang Z, Zhang CT: **Prediction of protein (domain) structural classes based on amino-acid index,**. *Eur. J. Biochem.* 1999, **266**:1043–1049.

[5] Wang ZZ, Yuan Z: **How good is prediction of protein structural class by the component-coupled method?** *Proteins* 2000, **38**:165 – 175.

[6] S KT, M GM, N PM: **Structural class prediction: an application of residue distribution along the sequence**. *Biophys Chem.* 2000, **88**:81–101.

[7] Cai YD, Chou KC: **Predicting subcellular localization of proteins in a hybridization space**. *Bioinformatics* 2004, **20**:1151–1156.

[8] Shen HB, Yang J, Liu XJ, Chou KC: **Using supervised fuzzy clustering to predict protein structural classes**. *Biochem. Biophys. Res. Commun.* 2005, **334**:577–581.

[9] Du QS, Jiang ZQ, He WZ, Li DP, Chou KC: **Amino Acid Principal Component Analysis (AAPCA) and its application in protein structural class prediction**. *J. Bio. Str. Dyn.* 2006, **23**:635–640.

[10] Shi J, Blundell TL, Mizuguchi K: **FUGUE: sequence-structure homology recognition using environment-specific substitution tables and structure-dependent gap penalties,**. *J Mol. Biol.* 2001, **310**:243–257.

[11] Zhou H, Y Z: **Fold recognition by combining sequence profiles derived from evolution and from depth-dependent structural alignment of fragments**. *Proteins* 2005, **58**:321–328.

[12] Cheng J, Baldi P: **A machine learning informationretrieval approach to protein fold recognition**. *Bioinformatics* 2006, **22**:1456–63.

[13] Gromiha MM, Suwa M: **A Simple statistical method for discriminating outer membrane proteins with better accuracy**. *Bioinformatics* 2005, **21**:961–968.

[14] Hirokawa T, Boon-Chieng S, Mitaku S: **SOSUI: classification and secondary structure prediction system for membrane proteins**. *Bioinformatics* 1998, **14**:378–379.

[15] Chou KC: **Prediction of protein structural classes and subcellular locations**. *Curr. Protein Pept. Sci.* 2000, **1**:171–208.

[16] Gromiha MM, Selvaraj S, Thangakani AM: **A Statistical method for predicting protein unfolding rates from amino acid sequence**. *J. Chem. Inf. Model* 2006, **46**:1503–1508.

[17] Gromiha MM, Oobatake M, Kono H, Uedaira H, Sarai A: **Relationship between amino acid properties and protein stability: Buried Mutations**. *J. Protein Chem.* 1999, **18**:565–578.

[18] Ding HQD, Dubchak I: **Multi-class protein fold recognition using support vector machines and neural networks,**. *Bioinformatics* 2001, **17**:349–358.

[19] Shen HB, Chou KC: **Ensemble classifier for protein fold pattern recognition**. *Bioinformatics* 2006, **22**:1717–1722.

[20] **PROLDA** [[http://granular.com/PROLDA/]].

[21] Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures**. *J. Mol. Biol.* 1995, **247**:536–540.

[22] R Development Core Team: *R: A language and environment for statistical computing*. http://www.R-project.org 2005.