Integration of Multi-Feature Fusion and PLS-DA in Protein Secondary Structure Prediction

Guangzao Huang[1], Meishuang Tang[2], Jinting Guan[1], Sun Zhou[1], Wenbing Zhu[1], and Guoli Ji[1]

[1]*Department of Automation, Xiamen University, Xiamen, Fujian 361005, China*
[2]*Modern Educational Technical and Practical Training Center, Xiamen University, Xiamen 361005, China*

Abstract. Protein structure prediction has become one of the central problems in the field of modern computational biology. Protein secondary structure prediction is the basis of the spatial structure prediction of proteins. This paper presents a novel method for protein secondary structure prediction, which integrates multi-feature fusion and partial least square discriminant analysis (PLS-DA). Multi-feature fusion can make full use of the available information of proteins; however, it also leads to high-dimensional and redundant features. Then PLS-DA is utilized to deal with the fused protein data, which can effectively extract features from the protein data and remove the redundant information. Several benchmark datasets are used to verify the performance of the proposed method. The experiment results show that the proposed method gives satisfying prediction results of protein secondary structure compared with existing methods. Therefore the integration of multi-feature fusion and PLS-DA can fully utilize the available protein information, effectively reduce dimension and achieve robust classification in the multi-category analysis of protein secondary structure.

1 Introduction

Protein is a kind of important biological macromolecules, which plays many vital roles for almost all kinds of biological phenomena. The function of protein is dependent on its spatial structure, which greatly promotes the study of protein structure [1]. The methods for protein structure prediction can be divided into two categories. One is based on actual experiment method, including X-ray diffraction and Nuclear Magnetic Resonance. The other one is based on computing method, including homology modelling [2], folds recognition [3] and ab initio prediction [4]. Though the actual experiment method is more accurate, it takes a long time and is restricted by the technology and equipment. However, the computing method can overcome these limitations and therefore it has great development potential and space. In the field of biological information engineering, protein secondary structure is the basis of the prediction of protein tertiary structure and it is also the first step of many protein researches. Based on the contents of protein secondary structures, the protein structures are classified into four categories, all-α, all-β, α/β and α+β. At present, the application of machine learning in protein secondary structure prediction has become one of the hottest research areas in bioinformatics.

The feature representation of protein has a huge impact on the secondary structure prediction. The usually adopted feature representation of protein includes Amino Acid Composition (AAC) [5], polypeptide composition, functional domain composition [6], physicochemical features, PSI-BLAST profiles [7] and function annotation information [8]. Obviously, making full use of available protein feature information can effectively improve the prediction of protein structures. To this end, this paper focuses on how to integrate multiple protein information including the PSI-BLAST profiles, PROFEAT and Gene Ontology as the feature representation of protein. However, it makes the protein data contain much more features than that of samples; therefore it is difficult to be processed by traditional discriminant analysis. The partial least squares discrimination analysis (PLS–DA) [9] proposed by Errikson et al. is a robust discriminant analysis method. It is particularly suitable for the data of multicollinearity among the explanatory variables, high dimension and small sample, which is the case of the protein data after being integrated with multiple features. In this paper, we study the performance of the proposed method of integration of multi-feature fusion and PLS-DA in the prediction of protein secondary structures.

2 Generating feature vector of protein by multi-feature fusion

In classification, all the information that decides which class a sample falls into is included in the sample feature vector. Therefore, how to create the protein feature vector has a significant influence on the established classification model. In protein secondary structure prediction, in order to make full use of available protein

feature information, in this paper the protein feature vector fuses three kinds of commonly used protein information, which are described in the following.

2.1 Linear predictive coding of the PSSM

The PSSM (Position Specific Scoring Matrix) [10] of a protein sequence represents homolog information affiliated with its aligned sequences, which is an effective feature for the prediction of protein structure and function. We used the PSI-BLAST program to search the NCBI's non-redundant (NR) database under the parameter setting h=0.001 and j=3 to get the PSSM of each protein sequence. The PSSM matrix elements were standardized between [0, 1] by the following standard sigmoid function

$$f(x)=\frac{1}{1+e^{-x}} \tag{1}$$

Then the linear predictive coding (LPC) scheme was used to turn the PSSM into a fixed-length feature vector.

2.2 Gene function annotation features

Gene Ontology (GO) is widely used in the field of bioinformatics, which can describe the function of gene and protein. The GO database used in this paper was downloaded from ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/gene_association.goa_uniprot.gz (released on Nov 26, 2014). We searched the GO terms of each protein sequence in the GO database and then all the GO terms related to the entire dataset are used to form a feature set:

$$protein_{GO}=\{g_i\}(i=1,\ 2,\ \cdots,\ \mathrm{n}) \tag{2}$$

$$g_i=\begin{cases}1,\ if\ its\ GO\ terms\ hit\ the\ ith\ term\\ 0,\ if\ not\ hit\end{cases} \tag{3}$$

where g_i is the i-th GO term and n is the number of GO terms of the dataset.

2.3 Extracting structural and physicochemical features from PROFEAT

Sequence structural and physicochemical features have been frequently used in the development of statistical learning models for predicting proteins and peptides of different structural, functional and interaction profiles. PROFEAT is a web server for computing commonly used structural and physicochemical features of proteins and peptides from amino acid sequence [11]. In this paper, we acquired 1497-dimension vector of PROFEAT feature for each protein sequence.

3 Partial least squares discrimination analysis

Partial Least Squares (PLS) algorithm [12] is a widely modeling method in chemometrics and many other fields in recent years [13]. The PLS algorithm can establish the relationship between the two data block, eliminate the redundant information to achieve the goal of dimension reduction. The PLS regression model has many advantages such as simple, robust, less calculation, no need to remove any explanatory variables and easy to interpret. PLS-DA is based on PLS regression algorithm [14]. When dealing with PLS-DA, the class labels of the samples need to be binary encoded. $X_{n\times m}$ is the feature matrix, $Y_{n\times g}$ is the class matrix, n is the number of samples, m is the number of features, g is the number of classes. The element y_{ij} in $Y_{n\times g}$ represents the relationship between the i-th sample x_i in $X_{n\times m}$ and the j-th class , which is expressed as

$$y_{ij}=\begin{cases}1,\ \text{if}\ x_i\in class\ j\\ 0,\ \text{if}\ x_i\notin class\ j\end{cases}\ j=1,\ 2,\ \cdots,\ \mathrm{g} \tag{4}$$

Then PLS regression model is modeled between X and Y in the usual way. The PLS algorithm decomposes the X and Y as follows:

$$X=TP^{'}+E=\sum_{i=1}^{a}t_i p_i^{'}+E \tag{5}$$

where T is the score matrix of X, t_i is the score vector, P is the load matrix, p_i is the load vector, E is the residual matrix and a is the number of latent variables;

$$Y=UQ^{'}+F=\sum_{i=1}^{a}u_i q_i^{'}+F \tag{6}$$

where U is the score matrix of Y, u_i is the score vector, Q is the load matrix, q_i is the load vector and F is the residual matrix. PLS regression algorithm extracts the latent variables from X and Y respectively, which satisfies the following conditions: (i) each group of latent variables extracting maximum variation information from the corresponding matrix; (ii) maximize the covariance between the two groups of latent variables. The general regression equation can be written as

$$Y=XB=XW\left(P^{'}W\right)^{-1}Q^{'} \tag{7}$$

where W is the weight matrix of PLS regression algorithm.

For an unknown samples x_{un} , the corresponding predicted value could be calculated by the following formula

$$y_{un}=x_{un}B=x_{un}W\left(P^{'}W\right)^{-1}Q^{'} \tag{8}$$

The predicted value y_{un} is in a real number and it needs to be translated into class attribute. For instance, if the maximum value of y_{un} is in the j-th column, the sample is predicted as the j-th class.

4 Materials and Methods

4.1 Datasets

Two groups of datasets were adopted to evaluate the proposed method. One group is the high similarity datasets, including Z277 and Z498. The other group is the low similarity datasets, including 1189 and 25PDB. The homology of proteins in low similarity dataset is lower than 40%. The specific information of these datasets is listed in Table 1.

Table 1. Datasets used to verify the effect of proposed method

Dataset	Protein Number					Feature Number
	All-α	All-β	a/β	α+β	total	
Z277	70	61	81	65	277	3387
Z498	107	126	136	129	498	3407
25PDB	443	443	346	441	1673	6621
1189	223	294	334	241	1092	6137

4.2 Evaluation method

The k-fold cross validation is used to evaluate the effectiveness of proposed method. The accuracy and overall accuracy are used as the evaluation index, which are formulated as follows:

$$accuracy(i) = \frac{TP_i}{m(i)} \tag{9}$$

$$overallaccuracy = \frac{\sum_{i=1}^{M} TP_i}{N} \tag{10}$$

where TP_i denotes the true positives of proteins in class i, $m(i)$ denotes the number of proteins in class i, N denotes the total number of proteins.

5 Experimental results and discussion

After multi-feature fusion, the protein dataset has some common characteristics such as huge data volume, far less number of samples than the feature dimension and mixed with noise interference. Therefore the discriminant analysis method employed needs to filter out noise, extract the features and create a model for protein secondary structure prediction, which is well suited to PLS-DA. One can see from Table 2 to 5 that the predicted results given by the method of integration of multi-feature fusion and PLS-DA are comparable to the existing methods. Besides, another observation is that protein datasets with low homology are harder to be recognized than those of high homology, especially for the recognition of the protein class of $\alpha + \beta$. This is because PLS-DA is a linear discriminant analysis model; therefore it is not fit for nonlinear classification problem.

PLS-DA can also be used as a dimension reduction method. From Figure 1, it can be seen that the extracted features by PLS-DA can significantly increase the similarity between protein samples within same class and the difference between protein samples from different classes. Therefore, the extracted features by PLS-DA can be combined with other nonlinear classifier, such as SVM to improve the recognition of low homology protein datasets. From Figure 2, it can be seen that PLS-DA is also an effective visualizing technique for protein data. Therefore PLS-DA can amalgamate many actions, such as classification, dimension reduction and data visualization in one for protein secondary structure prediction, and integration of multi-feature fusion further strengthens these advantages.

Table 2. Performance comparison of different methods on Z277 dataset.

Method	Prediction accuracy (%)				
	All-α	All-β	a/β	α+β	Overall
Markov-SVM by Qin et al. [15]	90.0	85.2	86.4	81.5	85.9
PLS-DA (10)	85.7	93.4	91.3	76.9	87.0

Note: the number in parentheses represents the optimal number of principal components of PLS-DA.

Table 3. Performance comparison of different methods on Z498 dataset.

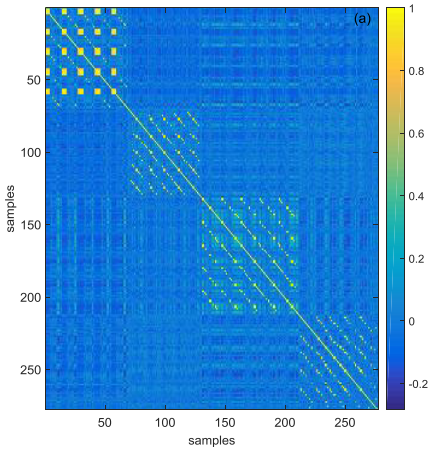
Method	Prediction accuracy (%)				
	All-α	All-β	a/β	α+β	Overall
AAC-PSSM-AC by Liu et al. [16]	94.4	96.8	97.8	93.8	95.8
PLS-DA (11)	93.5	96.0	97.1	88.4	93.8

Table 4. Performance comparison of different methods on 25PDB dataset.

Method	Prediction accuracy (%)				
	All-α	All-β	a/β	α+β	Overall
MLR model by Xia et al. [17]	92.6	72.5	71.7	71.0	77.2
PLS-DA (9)	73.8	74.3	82.7	43.1	67.7

Table 5. Performance comparison of different methods on 1189 dataset.

Method	Prediction accuracy (%)				
	All-α	All-β	a/β	α+β	Overall
Markov-SVM by Qin et al. [15]	53.8	79.3	68.3	32.0	60.3
PLS-DA (10)	82.1	75.2	74.3	33.2	63.8



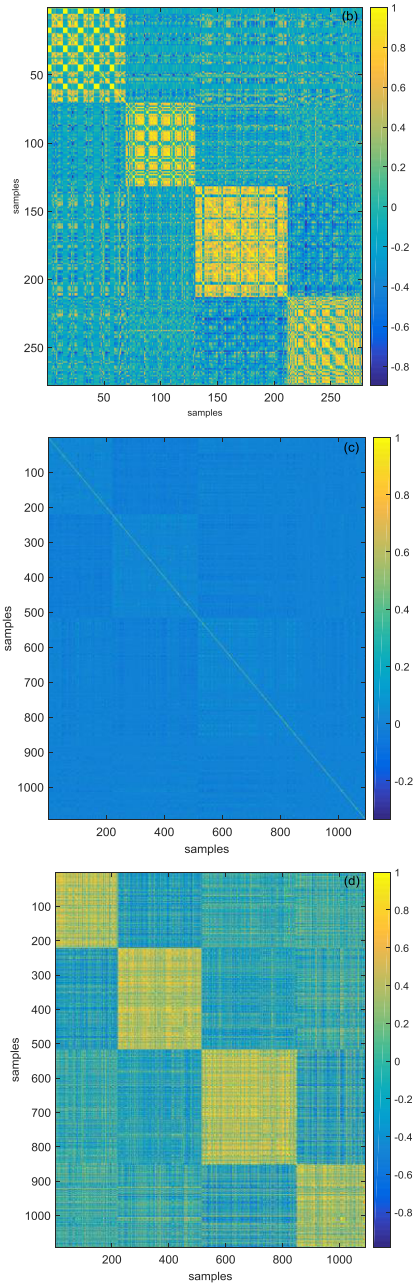
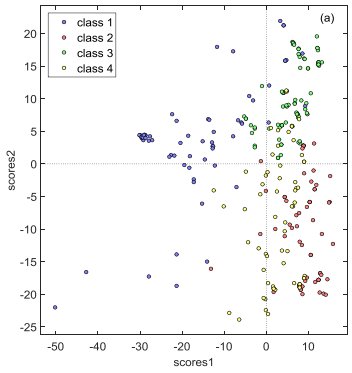


Figure 1. Heat maps of correlation coefficient matrixes of the protein datasets. (a) Z277 with raw features; (b) Z277 with 10 features extracted by PLS-DA; (c) 1189 with raw features;(d) 1189 with 10 features extracted by PLS-DA. For (a) (b), samples of 1-70, 71-131, 132-212, 213-277 and for (c) (d), samples of 1-223, 224-517, 518-851, 852-1092 represent the protein structure of All-α, All-β, a/β, α+β respectively.



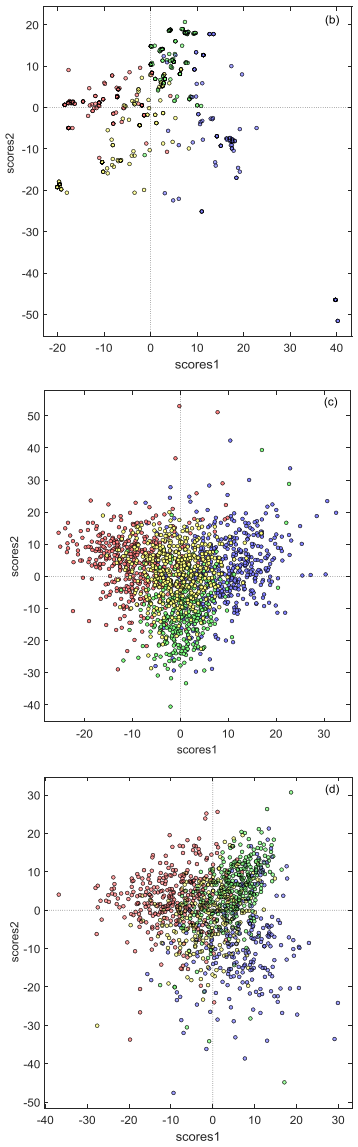


Figure 2. The two scores extracted by PLS-DA. (a) Z277; (b) Z498; (c) 25PDB; (d) 1189.

6 Conclusion

Multi-feature fusion can achieve the purpose of making full use of available protein information. PLS-DA can effectively extract features and remove redundant information. Integration of multi-feature fusion and PLS-DA can effectively deal with the problem of protein secondary structure prediction. Besides, PLS-DA is a linear discriminant analysis model; therefore it is not fit for nonlinear classification problem, such as recognizing the protein secondary structure of $\alpha + \beta$ in low similarity dataset. However, PLS-DA can still be used as a dimension reduction method; when it is combined with a regular nonlinear classifier, a more powerful nonlinear classifier can be constructed.

Acknowledgment

This work was supported by Natural Sciences and Engineering Research Council of Canada, National

Natural Science Foundation of China (No. 61573296, 61304141, 61375077), Fujian Provincial Industry-University-Research Cooperation Major Projects in China (2014H6025), Fujian Province Natural Science Foundation in China (No. 2014J01252), the Specialized Research Fund for the Doctoral Program of Higher Education of China (No. 20130121130004), the Fundamental Research Funds for the Central Universities in China (Xiamen University: No. 2015Y1115, 201412G009, 2014X0217).

References

1. DE Duve, C., Nature **333**, (1988)
2. Schwede, T., Kopp, J., Guex, N., Peitsch, M.C., Nucleic Acids Res. **31**, 13 (2003)
3. Micsonai, A. et al., Proc. Natl. Acad. Sci. **112**, 24 (2015)
4. Islam, M., Chetty, M., Evol. Comput. IEEE Trans. On **17**, 4 (2013)
5. Chou, K., Proteins Struct. Funct. Bioinforma. **43**, 3 (2001)
6. Rao, H., Zhu, F., Yang, G., Li, Z., Chen, Y., Nucleic Acids Res. **39**, suppl 2 (2011)
7. Song, J., Burrage, K., Yuan, Z., Huber, T., BMC Bioinformatics **7**, 1 (2006)
8. Marchler-Bauer, A. et al., Nucleic Acids Res. **39**, suppl 1 (2011)
9. Eriksson, L., Johansson, E., Kettaneh-Wold, N., Wold, S., Umetrics (2001)
10. Ahmad, S., Sarai, A., BMC Bioinformatics **6**, 1 (2005)
11. Li, Z.R. et al., Nucleic Acids Res. **34**, Web Server (2006)
12. DE Jong, S., Chemom. Intell. Lab. Syst. **18**, 3 (1993)
13. Wold, S., Ruhe, A., Wold, H., Dunn, I., WJ, SIAM J. Sci. Stat. Comput. **5**, 3 (1984)
14. Barker, M., Rayens, W., J. Chemom. **17**, 3 (2003)
15. Qin, Y.-F. et al., Protein Pept. Lett. **19**, 4 (2012)
16. Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., Amino Acids **42**, 6 (2012)
17. Xia, X.-Y., Ge, M., Wang, Z.-X., Pan, X.-M., PloS One **7**, 6 (2012)