Scholars Academic Journal of Biosciences (SAJB)
Sch. Acad. J. Biosci., 2017; 5(3):249-256
©Scholars Academic and Scientific Publisher
(An International Publisher for Academic and Scientific Resources)
www.saspublishers.com

ISSN 2321-6883 (Online)
ISSN 2347-9515 (Print)

DOI : 10.36347/sajb.2017.v05i03.021

Original Research Article

A Protein Structural Classes Prediction Method based on Various Information Fusion

Lifeng Lou, Baoguang Tian
[1]School of Mathematics and Physics, Qingdao University of Science and Technology, Qingdao 266061, China

***Corresponding author**
Lifeng Lou
Email: 1156586464@qq.com

Abstract: Protein structural class's knowledge plays an important role in understanding the folding mode of protein. The prediction of protein structural classes as a transitional stage of the secondary structure of the protein to the tertiary structure is considered to be an important and challenging task. In this paper, PSI-BLAST profile is used to extract the evolutionary information of protein, and the position-specific scoring matrix is obtained from PSI-BLAST profile. Then formula is used to transform PSSM into a fixed length feature vector. Extract the protein composition information and sequence order information from the pseudo-amino acid composition, and fuse all the extracted feature vectors. Finally, the fused feature vector is input to the support vector machine classifier to predict protein structural classes. The results were obtained by jackknife test and compared with other prediction methods on the two low similarity benchmark datasets 1189 and 640. The results show that the proposed method can predict the protein structural classes effectively.

Keywords: protein structural class prediction; position-specific score matrix; pseudo amino acid composition; support vector machine.

INTRODUCTION

Protein structural classes play a key role in protein secondary structure prediction, protein tertiary structure prediction and protein function analysis. Levitt and Chothia [1] proposed the concept of protein structural classes in 1976. They classify the protein sequence into four main classes: $\text{all-}\alpha$ class, $\text{all-}\beta$ class, $\alpha+\beta$ class, α/β class.

With the development of bioinformatics, many different methods have been proposed for protein structural classes prediction. Chen [2] proposed the SCEC method; the method incorporates evolutionary information encoded using PSI-BLAST profile based collocation of AA pairs. Liang [3] proposed a new feature extraction method MBMGAC-PSSM, which use three different auto-correlation descriptors on the position-specific score matrix obtained a 560-dimensional feature vector, the principal component analysis reduce the dimension to 175 dimensions. Raicar [4] proposed a Forward Consecutive Search scheme and used this strategy to exhaustively search for 544 physical and chemical properties, identified a subset of physicochemical properties, combined evolutionary information and syntactic information, the accuracy of protein structural classes prediction is improved on the benchmark datasets. In order to explore the potential of protein secondary structure information, Zhang [5] used the chaos game representation based on the protein secondary structure to obtain the protein sequence information and the secondary structure segment distribution information.

In this paper, a new method for protein structural class's prediction is proposed. PSI-BLAST profile and PseAAC are used to extract the features of protein sequence. PSI-BLAST profile can be used to obtain PSSM. PSSM contains abundant protein evolutionary information. By fusing protein evolutionary information and the protein sequence order information obtained a 142-dimensional feature vector, the fusion feature vector were input to the support vector machine classifier to predict. The overall accuracy of the two benchmark datasets 1189 and 640 are 70.5% and 64.2%, respectively. Finally, we compared with other prediction methods. The experimental results show that the proposed method can significantly improve the prediction accuracy of protein structural classes.

MATERIALS AND METHODS

Datasets

In order to make a fair and reasonable comparison with existing research results, two popular benchmark datasets used to evaluate the method: 1189 dataset [6], 640 dataset [2], with sequence similarity less than 40%, 25%, respectively. The 1189 dataset contains 1092 proteins, which contains 223 $\text{all}-\alpha$ class proteins, 294 $\text{all}-\beta$ class proteins, 334 α/β class proteins and 241 $\alpha+\beta$ class proteins. The 640 dataset contains 640 proteins, which contains 138 $\text{all}-\alpha$ class proteins, 154 $\text{all}-\beta$ class proteins, 177 α/β class proteins and 171 $\alpha+\beta$ class proteins.

Feature extraction

PsePSSM

The evolutionary information of proteins reflects conserved information and mutation information in the evolutionary process of amino acid residues at each position in the protein sequence, which determines the structure and function of the protein sequence. PSSM contain rich protein sequence evolutionary information, PSSM can be obtained based on PSI-BLAST profile, utilize each protein sequence as a seed to search and align homogenous sequences from NCBI's NR database using the PSI-BLAST program [7] with three iterations and a cutoff E-value 0.001.The generated PSSM can be expressed as:

$$PSSM = \begin{pmatrix} p_{1,1} & p_{1,2} & \cdots & p_{1,j} & \cdots & p_{1,20} \\ p_{2,1} & p_{2,2} & \cdots & p_{2,j} & \cdots & p_{2,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{i,1} & p_{i,2} & \cdots & p_{i,j} & \cdots & p_{i,20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ p_{L,1} & p_{L,2} & \cdots & p_{L,j} & \cdots & p_{L,20} \end{pmatrix}$$

Where L is the length of the query amino acid sequence, $P_{i,j}$ is the probability that the position of the i - th amino acid residue in the protein sequence is replaced by j - th the amino acid during evolution.

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

Where x is the original PSSM value?

The classifier based on machine learning needs to be input with a fixed length of feature vector. For different protein sequences, the sequence length L is different, The PSSM is a log-odds matrix of $L \times 20$, in order to transform PSSM into a fixed-length feature vector, and the following method is used to express protein P:

$$P_{psePSSM} = \left(\hat{p}_1, \hat{p}_2, \ldots \hat{p}_{20}, \psi_1^1, \psi_2^1, \cdots, \psi_{20}^1, \psi_1^2, \psi_2^2, \cdots, \psi_{20}^2, \cdots, \psi_1^m, \psi_2^m, \cdots \psi_{20}^m\right)^T$$

$$\hat{p}_j = \frac{1}{L}\sum_{i=1}^{L} p_{i,j} \qquad (j = 1,2,\cdots,20)$$

$$\psi_j^m = \frac{1}{L-m}\sum_{i=1}^{L-m}\left(p_{i,j} - p_{(i+m),j}\right)^2 , \qquad (j = 1, 2, \cdots, 20; m < L, m \neq 0)$$

Due to the maximum value of m must be less than the length of the shortest sequence in datasets, the length of the shortest sequence in the 1189 dataset is 10 and the length of the shortest sequence in the 640 dataset is 37, so the range of m is 1 to 9 and integer. The PSSM can be transformed into a $20+(20\times m)$ dimensional fixed-length feature vector by the above method.

PseAAC

Pseudo amino acid composition [8] was originally introduced to improve the prediction quality for protein subcellular localization and membrane protein type. The pseudo-amino acid composition method is a protein amino acid sequence coding method, which can extract a 20+ λ dimension feature vector from the amino acid sequence, in which the first 20 dimensions are the amino acid composition and the latter λ dimensions express the protein sequence order information. By definition, the pseudo-amino acid composition can be expressed as

$$P = [p_1, p_2, p_3, \cdots, p_{20}, p_{20+1}, \cdots p_{20+\lambda}]^T$$

$$p_u = \begin{cases} \dfrac{f_u}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 1 \le u \le 20 \\ \dfrac{\omega \tau_{u-20}}{\sum_{u=1}^{20} f_u + \omega \sum_{k=1}^{\lambda} \tau_k} & 21 \le u \le 20+\lambda \end{cases}$$

Where ω is the weighting factor ($0<\omega<1$), the value is 0.05. f_u represents the frequency at which the u-th amino acid appears in the protein. τ_k is the k-th sequence correlation factor, which can be calculated from the sequence correlation function $J_{i,i+k}$:

$$\tau_k = \frac{1}{L-k}\sum_{i=1}^{L-k} J_{i,i+k} \qquad k<L \quad J_{i,i+k} = \frac{1}{6}\left\{\left[H_1(R_{i+k})-H_1(R_i)\right]^2+\left[H_2(R_{i+k})-H_2(R_i)\right]^2+\cdots+\left[H_6(R_{i+k})-H_6(R_i)\right]^2\right\}$$

Where L is the amino acid sequence length and k is any positive integer less than L. $H_1, H_2, \cdots, H_6$ represents the hydrophobic value, the hydrophilic value, the molecular weight of the side chain, the ionization constant of α-$COOH$ and NH_3, and the isoelectric point at $25\ ^{\circ}C$. λ represents the correlation factor, λ is in the range 1 to 9 and integer. Since the concept of pseudo-amino acid components has been proposed, multiple pseudo amino acid component representation has been developed for enhancing the prediction quality of protein attributes.

Support vector machine

Support vector machine (SVM) is a machine learning algorithm based on statistical learning theory. SVM are widely used in statistical classification and regression analysis. SVM has many unique advantages in solving small sample, nonlinear, high dimensional pattern recognition and can be applied to other machine learning problems such as function fitting. The basic idea of SVM is to transform the space of the input sample into a high-dimensional space by nonlinear transformation, in this new high-dimensional space, the optimal linear class hyperplane is obtained. By defining the appropriate kernel function To achieve, the discriminant function is:

$$f(x) = sign\left\{\sum_{i=1}^{n} \alpha_i y_i K(x_i, x_j) + b_0\right\}$$

Where α_i is the Lagrange multiplier, b_0 is the classification threshold, and $K(x_i, x_j)$ is the kernel function. Generally, the four kernel functions are often used for SVM, linear kernel functions, polynomial kernel functions, radial basis (RBF) kernel functions and sigmoid kernel functions. Empirical studies have shown that the RBF kernel function outperfoms the other three kernel functions. Therefore, we choose to use the RBF kernel function, which is defined as $K(x_i, x_j) = \exp\left(-\gamma\|x_i - x_j\|^2\right)$. The regularization parameter C and kernel parameter γ are optimized based on ten-fold cross-validation on 1189 dataset using a grid search strategy in the LIBSVM software. The range of values for C and γ are $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{5}]$, respectively. Various pairs of (C, γ) values are tried and the one with the best cross cross-validation accuracy is picked. This paper uses the LIBSVM software developed by Chang and Lin [9], which can be downloaded at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Prediction assessment

Independent sample test, jackknife test and self-compatible three methods are often used to test the effectiveness of the model. Jackknife has been widely used in protein structure and function because its results are unique. In this paper, the predictive model was evaluated by jackknife test and some indicators Sensitivity (Sens), Specificity (Spec), Overall Accuracy (OA) and Matthew Correlation Coefficient (MCC). These indicators are defined as follows:

$$Sens_j = \frac{TP_j}{TP_j + FN_j} = \frac{TP_j}{|C_j|}$$

$$Spec_j = \frac{TN_j}{FP_j + TN_j} = \frac{TN_j}{\sum_{k \ne j} |C_k|}$$

$$OA = \frac{\sum_j TP_j}{\sum_j \left|C_j\right|}$$

$$MCC_j = \frac{TP_j \times TN_j - FP_j \times FN_j}{\sqrt{\left(TP_j + FP_j\right)\left(TP_j + FN_j\right)\left(TN_j + FP_j\right)\left(TN_j + FN_j\right)}}$$

Where TP_j is the number of true positives, FP_j is the number of false positives, TN_j is the number of true negatives, FN_j is the number of false negatives, and $\left|C_j\right|$ is the number of j-th class proteins.

For convenience, we designed the flow chart of the proposed method, as shown in Figure 1.

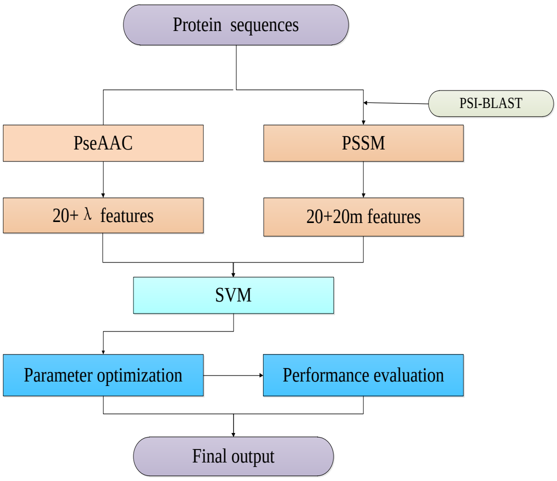

Fig-1: The flow chart of this method.

The step of protein structural classes prediction based on our method is described as follows:

1) input the amino acid sequence of protein, and the class label corresponding to the four classes proteins;
2) PSSM was obtained by PSI-BLAST profile, the protein sequence was transformed into a numerical value, and the PSSM was transformed into a $20+(20\times m)$ dimension feature vector by the formula , where $1\le m\le 9$;
3) The 20+ λ dimension feature vector is extracted using the pseudo-amino acid component , and fusing the feature vector that have been obtained in 2) ;
4) The fusion feature vector is input into the support vector machine to predict the protein structural classes ;
5) According to the predicted accuracy, determine the best parameters λ , m ;
6) According to the best parameters of the model obtained from 5), the model performance was evaluated using the given evaluation indexes Sens, Spec, OA and MCC.

RESULTS AND DISCUSSION

The selection of the optimal parameter λ

In this paper, the PsePSSM and PseAAC methods are used to extract the amino acid sequence of the protein. In the process of using the PseAAC method, an important consideration is the choice of the λ value. The λ value has a important influence on the prediction result. If λ is too small, the feature vector contain little sequence information, and if λ is too large, it will bring more redundant information. In order to make the feature vector contain more protein amino acid sequence information and carry less redundant information, we set the λ value from 1 to 9 in turn, and obtained the result by jackknife test. The overall accuracy on the two benchmark datasets 1189 and 640, as shown in Figure 2.

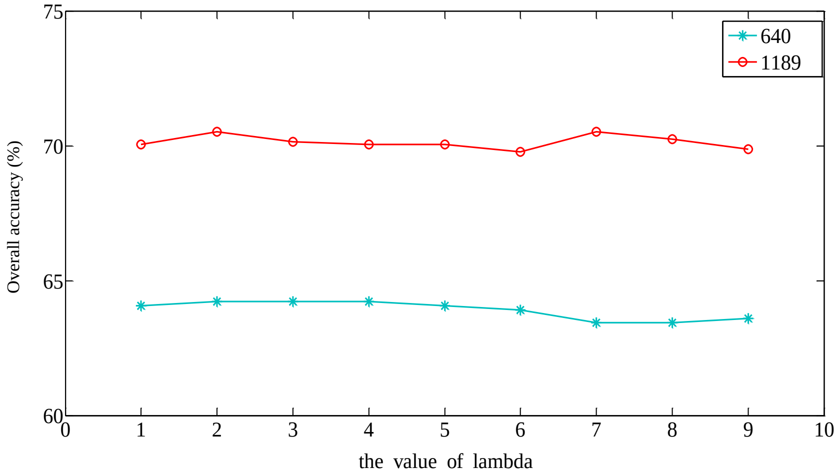

Fig-2: The overall accuracy of different values of λ for our method on the 1189 and 640 datasets.

As shown in Fig.2, We observe that the λ value has a minor effect on the overall prediction accuracy, which indicates that the model proposed in this paper is reliable and robust. Finding the appropriate λ value not only depends on the specific problem, but also by the nature of the dataset itself. The optimal λ is 2 due to the accuracy of the two datasets.

The selection of the optimal parameter m

The pseudo position-specific score matrix extracts the evolutionary information of the amino acid sequence. The parameter m of PsePSSM represents the distance between two amino acid residues in the protein sequence. Because of $m < L$, the value of m is in the range of 1 to 9. we set the m value from 1 to 9, and the results are verified by jackknife. The overall accuracy on the two benchmark datasets 1189 and 640, as shown in Figure 3.

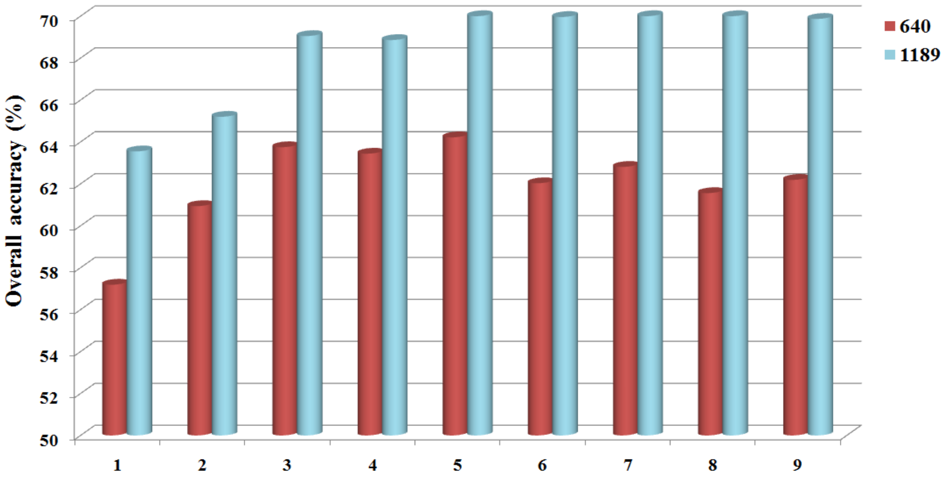

Fig-3: The overall accuracy of different values m of for our method on the 1189 and 640 datasets.

In the case of other conditions of this model, by changing the m value, the results have a large impact. m Values are constantly changing, the accuracy are constantly changing. As shown in Fig.3, it can be seen that the histogram is the highest when $m = 5$ on the 640 dataset. On the 1189 dataset, with the m value increases, the height of the histogram increases in turn, and when the value of m is 5, 6, 7, 8, 9, the height of the histogram changes slightly, the change can be ignored. We choose $m = 5$ as the best parameter for this model. Finally, a 142-dimensional feature vector is used to represent the protein sample.

Prediction performances of our method

In this section, we report the results of jackknife test performed on the two benchmark datasets in Table 1. The results include the performance evaluation indicators Sens, Spec and MCC for each structural class, and the overall accuracy OA.

Table-1: Prediction performances of our method on two benchmark datasets

Dataset	Structure class	Sens (%)	Spec(%)	MCC (%)	OA (%)
1189	all-α	74.89	91.36	66.15	70.51
	all-β	80.95	90.48	71.43	
	α/β	80.84	78.37	56.98	
	$\alpha+\beta$	39.42	89.76	33.06	
640	all-α	70.29	91.81	64.05	64.22
	all-β	68.18	87.43	56.14	
	α/β	80.23	76.86	54.47	
	$\alpha+\beta$	39.18	81.90	22.33	

As can be seen from Table 1, the overall accuracy obtained on the two benchmark datasets 1189 and 640 is 70.51% and 64.22%, respectively. Compared the prediction accuracy of four protein structural classes, all - α class has a maximum Spec value of 91.81%, and the MCC and Sens values are also high, indicating that all - α class prediction are the most reliable. At the same time, β class and α/β class protein predictions are also satisfactory, the performance indicators have reached more than 54% on the two datasets. In contrast，the prediction of $\alpha+\beta$ class is lower than the other three classes, such as Sens and MCC are less than 40%.There are similar trend in all protein structural classes prediction methods. This may be due to anti-parallel β -sheets are more difficult to identify, and it can not be ignored with other categories overlap. The fact that there is still a lot of challenges in future research to improve the predictive accuracy of $\alpha+\beta$ class of proteins.

Comparison of accuracy between different classification algorithms

In this paper, we choose the support vector machine as a classifier. In order to show the superiority of the support vector machine in the prediction of protein structural classes, we adopt other three different classifiers K nearest neighbor algorithm (KNN), naive Bayesian classifier (Bayes), linear discriminant analysis classifier (LDA).The fusion 142-dimensional feature vector and jackknife were used to predict the protein structural classes in the same datasets. The overall accuracy and the accuracy of each class are shown in Table 2.

Table-2: Comparison of accuracy between different classification algorithms.

Dataset	Classifier	Prediction accuracy (%)				
		all - α	all - β	$\alpha+\beta$	α/β	Overall
1189	KNN	51.12	58.50	30.71	75.75	56.14
	Bayes	65.02	60.54	22.82	73.65	57.14
	LDA	65.02	73.13	34.02	74.85	63.37
	SVM	74.89	80.95	39.42	80.84	70.51
640	KNN	40.58	38.31	29.24	71.75	45.63
	Bayes	61.59	53.90	23.39	74.57	53.13
	LDA	59.42	64.94	38.01	67.80	57.34
	SVM	70.28	68.18	39.18	80.23	64.22

Table 2 shows that the overall accuracy obtained with support vector machines is 6.88-18.59% higher than that of other three classifiers, which indicates that support vector machines are more suitable for protein structural classes prediction based on PsePSSM and PseAAC.

Comparison with other prediction methods

In this section, in order to objectively evaluate the validity of the proposed method, we compared the results of this paper with the other five methods on the same datasets. Select the accuracy of each class and the overall accuracy as a comparison indicator, as shown in Table 3. Comparison method Markov-SVM [10] is a new feature extraction approach based on relative polypeptide composition, AAD-CGR [11] is proposed to analyze amino acids sequence by recurrence quantification analysis based on chaos game representation. SCEC [2] incorporates evolutionary information encoded using PSI-BLAST profile-based collocation of AA pairs. The compared methods also include other competitive PSSM-based methods such as RPSSM [12]. IB1 [2] method is nearest neighbour classifier used in Chen's article.

Table-3: Performance comparison of different methods on two benchmark datasets

		Prediction accuracy(%)				
Dataset	Method	all-α	all-β	α/β	$\alpha+\beta$	Overall
1189	RPSSM	67.7	75.2	74.6	17.4	60.2
	Markov-SVM	53.8	79.3	68.3	32.0	60.3
	IB1	65.3	67.7	79.9	40.7	64.7
	AAD-CGR	62.3	67.7	66.5	63.1	65.2
	SCEC	75.8	75.2	82.6	31.8	67.6
	Our method	74.9	81.0	80.9	39.4	70.5
640	SCEC[26]	73.9	61.0	81.9	33.9	62.3
	Our method	70.3	68.2	80.2	39.2	64.2

As can be seen from Table 3, this method achieved the highest overall accuracy 70.5% on the 1189 dataset, which is 2.9-10.3% higher than other methods. And the accuracy of all - β class improved by 0.7% compared with previous best-performance results. For 1189 dataset, the accuracy of all-β class and $\alpha+\beta$ class is 5.8%, 7.6%higher than that of SCEC method, respectively. For 640 dataset, the accuracy of all-β class and $\alpha+\beta$ class is 1.9%, 7.2% and 5.3% higher than that of SCEC method, respectively. The method has significant improvement for difficult to $\alpha+\beta$ class. For the α/β class of proteins, our method also gets favorable accuracy, although it is not the highest. The true structure of the protein is much more complex than our theoretical model, and this paper does not use the secondary structure information, and these facts will be considered to improve the accuracy of our future work.

CONCLUSIONS

Protein structural class's prediction is a very important and challenging problem. In this paper, we proposed a new method for predicting protein structural classes. We used PSI-BLAST profile to get the position-specific score matrix, transformed it into a fixed-length feature vector and fused PseAAC information. Support vector machine as a classifier. PSSM contains abundant protein evolutionary information. PseAAC avoid the loss of sequence order information of protein sequences. Support vector machine classification algorithm can deal with high-dimensional data, to avoid over-fitting and effective removal of non-support vector. The overall accuracy of the two datasets are 70.5% and 64.2%, respectively, compared with the multiple prediction methods on two low similarity benchmark datasets 1189 and 640. The experimental results show that the proposed method can effectively improve the prediction accuracy of protein structural classes and is expected to be used for the prediction of other properties of protein.

REFERENCE

1. Chothia C, Michael L. Structural patterns in globular proteins. Nature. 1976 Jun 17; 261:552-8.
2. Chen KE, Kurgan LA, Ruan J. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. Journal of computational chemistry. 2008 Jul 30; 29(10):1596-604.
3. Liang YY, Liu SY, Zhang SL. Prediction of protein structural class based on different autocorrelation descriptors of position-specific scoring matrix. MATCH: Communications in Mathematical and in Computer Chemistry. 2015 Jan 1; 73(3):765-84.
4. Raicar G, Saini H, Dehzangi A, Lal S, Sharma A. Improving protein fold recognition and structural class prediction accuracies using physicochemical properties of amino acids. Journal of theoretical biology. 2016 Aug 7; 402:117-28.
5. Zhang L, Kong L, Han X, Lv J. Structural class prediction of protein using novel feature extraction method from chaos game representation of predicted secondary structure. Journal of theoretical biology. 2016 Jul 7; 400:1-0.
6. Wang ZX, Yuan Z. How good is prediction of protein structural class by the component-coupled method?. Proteins: Structure, Function, and Bioinformatics. 2000 Feb 1; 38(2):165-75.
7. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic acids

research. 1997 Sep 1; 25(17):3389-402.

8. Shen HB, Chou KC. PseAAC: a flexible web server for generating various kinds of protein pseudo amino acid composition. Analytical biochemistry. 2008 Feb 15; 373(2):386-8.
9. Chang CC, Lin CJ. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011 Apr 1; 2(3):27.
10. Qin YF, Wang CH, Yu XQ, Zhu J, Liu TG, Zheng XQ. Predicting protein structural class by incorporating patterns of over-represented k-mers into the general form of Chou's PseAAC. Protein and peptide letters. 2012 Feb 1; 19(4):388-97.
11. Yang JY, Peng ZL, Yu ZG, Zhang RJ, Anh V, Wang D. Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. Journal of Theoretical Biology. 2009 Apr 21; 257(4):618-26.
12. Ding S, Li Y, Shi Z, Yan S. A protein structural class’s prediction method based on predicted secondary structure and PSI-BLAST profile. Biochimie. 2014 Feb 28; 97:60-5.