MATCH
Communications in Mathematical and in Computer Chemistry

MATCH Commun. Math. Comput. Chem. **75** (2016) 743-758

ISSN 0340 - 6253

Using Correlation Analysis and Nonnegative Matrix Factorization to Predict Protein Structural Classes via Position–Specific Scoring Matrix

Yunyun Liang[*]**, Sanyang Liu, Shengli Zhang**

School of Mathematics and Statistics, Xidian University, Xi'an 710071, P. R. China

(Received November 4, 2015)

Abstract

Prediction of protein structural classes plays an important role in protein science, such as protein function prediction, protein fold recognition and protein folding rate analysis. Currently, prediction based solely on the position-specific scoring matrix(PSSM) has played a key role in improving the prediction accuracy. Feature extraction and feature selection are two critical steps for the prediction quality. In this paper, we propose a novel method using correlation analysis on the PSSM. Then a 3600-dimensional(3600D) feature vector is constructed and the dimension is decreased to 200D by using nonnegative matrix factorization (NMF). To evaluate the proposed method, objective jackknife cross-validation tests are performed on two widely used low-similarity datasets: 1189 and 25PDB. Our method achieves the favorable performance on prediction accuracies and also outperforms the other listed PSSM-based methods. The result shows that our approach will offer a reliable tool for prediction of protein structural classes, especially for low-similarity sequences.

1 Introduction

Protein structural classes prediction problem is a typical pattern recognition problem and defined as categorizing a given protein into one of four structural classes namely all-α, all-β, α/β, and $\alpha+\beta$, which is proposed by Levitt and Chothia [1]. Knowledge of protein structural class can provide useful information to understand protein folding patterns, and

[*]Corresponding author. Tel./Fax:+86-29-88202860. E-mail: yunyunliang88@163.com

play an important role in improving the prediction quality of protein tertiary structure as well as protein function [2]. Hence, finding a fast and accurate computational approach is critical for solving this problem, especially for low-similarity sequences.

During the last two decades, various important efforts that have been made to establish a really useful statistical predictor to tackle this problem. During the last two decades, a great number of statistical learning algorithms were developed to tackle this problem. The main improvement focused on three aspects: the first aspect is feature extraction, by which the different length sequences are converted into a fixed-length vector. The methods include amino acid composition (AAC) [3–5], pseudo-amino acid composition (PseAAC) [6–8], polypeptide composition [9, 10], functional domain composition [11], PSI-BLAST profile [12–14], sequence comparison [15,16], PsePSSM [17,18] and predicted protein secondary structure [19, 20]. The second aspect is feature selection for reducing the influence of redundancy, which includes principal component analysis (PCA) [21], SVM-RFE [22], wrapper and filter [23] and so on. The final aspect is a choice of favorable classification algorithm. Currently, the algorithm contains neural network [24], support vector machine (SVM) [5, 25, 26], fuzzy clustering [27], Bayesian classification [28], rough sets [29] and k-nearest neighbor [30] and so on. Despite some of the existing methods have shown the excellent performance, the information embedded in the PSSM has not been sufficiently explored, there remains have space for further improvement.

In this paper, we propose a novel method using correlation analysis on the PSSM for feature extraction and nonnegative matrix factorization(NMF) for feature selection. According to correlation analysis, we construct a 3600D feature vector, which is too large to input into SVM. The large dimension will lead to a handicap for the computation and information redundancy. Hence, finding a suitable dimension reduction method is very important. Originally, Lee and Seung [31, 32] applied NMF to decomposed facial images and derived parts-based representation of whole images. NMF is proposed as a matrix factorization technique that produces a useful decomposition in the analysis of data. NMF decomposes the data as a product of two matrices that are constrained by nonnegative elements. This method results in a reduced representation of the original data that can be seen either as a feature extraction or a dimensionality reduction technique. In the field of bioinformatics, NMF has successfully be applied to biclustering of gene expression data [33], metagenes and molecular pattern discovery [34], improving molecular cancer class discovery [35] and improving profile-profile alignment features for fold recognition and remote homolog detection [36]. Finally, with the help of NMF, 200 features are obtained for SVM classifier. To evaluate our method, jackknife cross-validation test is employed on two widely benchmark datasets, the experimental results demonstrate that our approach is an effective classifier and achieves the competitive performance compared with the other PSSM-based methods for low-similarity sequences.

2 Materials and methods

2.1 Datasets

In order to test current method strictly and facilitate the comparison with the previous works, two popular benchmark datasets are adopted for evaluating the performance of our method: the 1189 dataset [28]and the 25PDB dataset [37], which include 1092 and 1673 protein domains with sequence similarity lower than 40% and 25%, respectively. More details about the two datasets are listed in Table 1.

Table 1 The compositions of three datasets adopted in this paper.

Dataset	All-α	All-β	α/β	$\alpha+\beta$	Total
1189	223	294	334	241	1092
25PDB	443	443	346	441	1673

2.2 Feature extraction

To develop a powerful predictor for the protein structural class based on position-specific scoring matrix(PSSM), one of keys is to formulate the protein samples with an effective mathematical expression that truly reflect their intrinsic correlation [38]. Here, we define auto-cross correlation on PSSM to extract features.

2.2.1 Position-specific scoring matrix

To extract the evolutionary information in protein study, we use each protein sequence as a seed to search and align homogenous sequences from NCBI's NR database(ftp://ftp.ncbi.nih.gov/blast/db/nr) using the PSI-BLAST program [39] with parameters h=0.001 and j=3. PSI-BLAST will return a position-specific scoring matrix(PSSM), the $(i,\ j)$th entry of the obtained matrix represents the score of the amino acid residue in the ith position of the protein sequence being changed to amino acid type j in the biology evolution process.

Let us denote the PSSM as

$$PSSM = (P_1, P_2, \cdots, P_j, \cdots, P_{20}), \tag{2.1}$$

where $P_j = (P_{1,j}, P_{2,j}, \cdots, P_{L,j})^T$ $(j = 1, 2, \cdots, 20)$, L represents the length of the protein sequence, 20 represents the 20 native amino acid types and T is the transpose operator. Based on the original PSSM scores, we further normalize each element using the following logistic function:

$$f(s) = 1/(1 + e^{-s}), \tag{2.2}$$

where s is the original PSSM value, this process can reduce the bias and noise contained in the original scores.

2.2.2 Auto-cross correlation analysis

A protein sequence can be viewed as a time sequence of the corresponding physicochemical properties. Here, only the evolutionary information represented in the form of PSSM is adopted as the considered properties. In this paper, each amino acid is taken as one property and the PSSM is considered as the time sequences of all properties. However, according to the PSSM descriptor, proteins with different lengths will correspond to matrices with different numbers of rows. Here auto-cross correlation analysis is introduced to transform protein sequences into a uniform representation. As a powerful statistical tool, autocorrelation descriptor has been successfully adopted by our research group for prediction of protein structural classes for low-similarity sequences [40]. The autocorrelation only measures the correlation of the same property between two amino acid residues separated by a certain distance of lag apart along a protein sequence. Whereas, for the different properties or the different columns, the correlation analysis is missing. Hence, we define PSSM based on cross correlation transformation, the equation combined with autocorrelation descriptor can be defined as

$$C^{lag}_{j1,j2} = \frac{1}{L - lag} \sum_{i=1}^{L-lag} P_{i,j1} \times P_{i+lag,j2}, (j1, j2 = 1, 2, \cdots, 20; lag < L, lag \neq 0) \tag{2.3}$$

while $j1 = j2, C^{lag}_{j1,j2}$ represents autocorrelation factor of amino acid type $j1$, $j1 \neq j2, C^{lag}_{j1,j2}$ represents cross correlation factor of two different amino acid types $j1$ and $j2$. In this way, a protein sequence is represented by a vector of $lg * 400$, lg is the maximum lag $(lag = 1, 2, \cdots, lg)$. The parameter lg must be smaller than the length of the shortest sequence in the all datasets. Here, the length of the shortest sequence for our datasets is 10, which belongs to 1189 dataset, hence $lg = 9$, the value of lag can be $1, 2, 3, \cdots, 9$. Ultimately, each protein sequence is converted into a 3600-dimensional vector by fusing PSSM and auto-cross correlation analysis.

2.3 Nonnegative matrix factorization

A 3600-dimensional feature vector is too large to input into classifier. The large dimension can lead to three problems: over-fitting, information noise and a handicap for the computation. Hence, dimensionality reduction or feature selection plays an important role in classification task.

NMF is a matrix factorization algorithm originally proposed by Lee et al. [31] to analysis of facial images. This technique can be applied to the analysis of multidimensional features data in order to reduce the dimensionality.

A formal description of nonnegative matrix factorization can be described as follows [31]:

$$V \approx WH, \tag{2.4}$$

where $V = \mathbb{R}^{m\times n}$ is a positive data matrix with m variables and n objects, $W = \mathbb{R}^{m\times r}$ is the reduced r basis vectors for factors, and $H = \mathbb{R}^{r\times n}$ contains the coefficients of the linear combinations of the basis vectors, which are also known as encoding vectors. $(m+n)r < mn$, all matrices V, W, H are nonnegative, and the columns of W are normalized(sum up to 1). As we have known, the main difference between NMF and other classical factorization methods relies in the non-negativity constraints imposed on both the basis vectors W and the encoding vectors H.

The objective function, based on the square of the Euclidean distance between V and WH, can be defined using the following function, which we need to minimize [32]:

$$D(V\|WH) = \|V - WH\|^2 = \sum_{ij}(V_{ij} - (WH)_{ij})^2 \tag{2.5}$$

with the constraints $W, H \geq 0$.

The derived algorithm is as follows:

1. Initialize W and H with positive random numbers.
2. Compute the new basis matrix W by the update rules:

$$W_{ia} \leftarrow W_{ia}\frac{(VH^T)_{ia}}{(WHH^T)_{ia}} \tag{2.6}$$

3. The columns of W are normalized.

4. Compute the new coefficient matrix H by the update rules:

$$H_{a\mu} \leftarrow H_{a\mu} \frac{(W^T V)_{a\mu}}{(W^T W H)_{a\mu}} \tag{2.7}$$

5. Determine whether meeting the terminating condition, if yes, it will stop computing, output the base matrix W and the coefficient matrix H, or return to the step 2. Repeat this process until termination condition is satisfied.

2.4 Support vector machine

Support vector machine (SVM) [41] is not only a kind of machine learning algorithm based on statistical learning theory for binary classification problems but is also superior in practical applications. As a supervised machine learning technology, it has been successfully used in bioinformatics [5,23,25,42,43,53] by transforming the input vector into a high-dimension Hilbert space to find a separating hyperplane in this space. In this paper, we adopt one against all strategy for solving a multi-class problem by converting it into a series of two problems. For example, for a K-class problem, there are K two-class subclassifiers needed to be constructed by one against all method. The ith subclassifier is trained by considering all the proteins in the ith class as positive samples and all other classes as negative. Generally, four basic kernel functions used by SVM, i.e. linear function, polynomial function, sigmoid function and radial basis function (RBF). Here, we choose the RBF as SVM's kernel due to its superiority for solving nonlinear problem [42,43,50,54], which is defined as K$(x, x^{'})$=exp$(-\gamma\|x - x^{'}\|^2)$. The kernel parameter γ and the regularization parameter C are optimized based on the 1189 dataset by fifteen-fold cross validation using a grid search strategy in the LIBSVM package [44], which is written by Lin's lab and can be freely downloaded from http://www.csie.ntu.edu.tw/∼cjlin/libsvm, where C is allowed to take a value only between 2^{-5} to 2^{15} and γ only between 2^{-15} to 2^5.

2.5 Performance evaluation

Generally speaking, the following three cross-validation methods in statistical prediction are often used to examine the quality of a predictor and its effectiveness in practical application: independent dataset test, sub-sampling test or K-fold crossover and jackknife test. Among these three methods, the jackknife test can exclude the "memory" effect.

Also, it does not have the arbitrariness problem at all due to its ability of yielding a unique result for a given dataset [45]. Hence, we adopt jackknife test to evaluate the accuracy in this paper. During the process of the jackknife test, all the samples in the benchmark dataset is singled out one by one and tested by the predictor, trained with the remaining samples.

To evaluate the performance of our method comprehensively, we report six standard performance measures, including Sensitivity (Sens), Specificity (Spec), F-measure, Matthew's correlation coefficient (MCC), Area Under ROC Curve (AUC) and Overall accuracy (OA). F-measure is a more robust metric avoid to overestimate the performance of some metrics, which is the harmonic mean of recall and precision. MCC represents the correlation coefficients between the observed and the predicted class. Its value ranges from +1 (indicating best prediction model) to -1 (indicating worst prediction model). AUC is the area calculated under receiver operating characteristic (ROC) curve plotted by FP rate vs TP rate. Its value ranges from 0 to 1. These measures are defined as follows

$$Recall \ \ or \ \ Sens = \frac{TP}{TP+FN} \tag{2.8}$$

$$Spec = \frac{TN}{FP+TN} \tag{2.9}$$

$$Precision = \frac{TP}{TP+FP} \tag{2.10}$$

$$F = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.11}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{2.12}$$

$$AUC = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right) \tag{2.13}$$

$$OA = \frac{TP+TN}{TP+FN+FP+TN} \tag{2.14}$$

where TP represents the number of true positives, FP represents the number of false positives, TN represents the number of true negatives and FN represents the number of false negatives, n represents the number of classes, respectively.

3 Results and discussion

3.1 Selection of the factorization rank *r*

An important consideration in the application of the classical NMF model, is the selection of the number of factors used to better represent the data. Generally, as a rule of thumb, this value is generally chosen so that $(m+n)r < mn$. Nevertheless, this estimation is not informative enough to make a proper decision. Finding an appropriate value of r depends on the concrete problem and it is mostly influenced by the nature of the dataset itself.

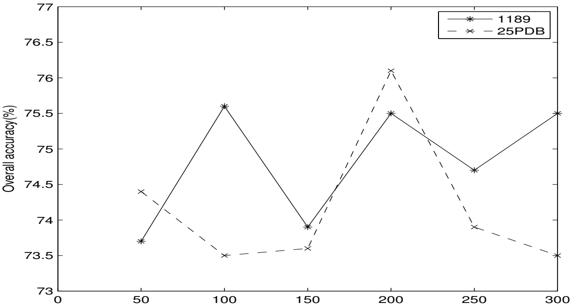

Figure 1: The choice of factorization rank r.

In this paper, we chose the factor for 50, 100, 150, 200, 250, 300, and obtain the overall accuracy for 1189 and 25PDB, respectively. As shown in Fig. 1, the optimal r is 200 due to the accuracies of the two datasets. So a 3600D feature vector is reduced to 200D by NMF. The general framework of the proposed method is shown in Fig. 2.

3.2 Prediction performance of our method

According to the NMF algorithm, we obtain a 200D feature vector, then the 200 features are input into SVM using one against all strategy. The RBF kernel function, the grid-search approach and fifteen-fold cross-validation for 1189 dataset are used to find the best parameters of $C = 362.0387$ and $\gamma = 32$ for SVM. To verify the performance of our method, rigorous jackknife cross-validation tests are performed on two widely used low-similarity datasets(1189, 25PDB). The experiment results are shown in Table 2.

As listed in Table 2, we report the Sens, Spec, F-measure, MCC and AUC for each structural class, as well as the OA. Relying solely on PSSM for feature extraction, we

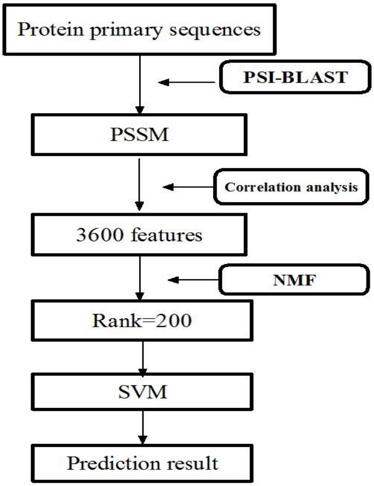

Figure 2: The general framework of the proposed method.

Table 2 The prediction quality of our model on the 1189 and 25PDB datasets

Dataset	Structural class	Sens(%)	Spec(%)	F-measure	MCC	AUC
1189	All-α	77.1	94.1	0.77	0.71	0.86
	All-β	86.7	93.5	0.85	0.79	0.90
	α/β	79.9	90.1	0.79	0.70	0.85
	$\alpha+\beta$	53.9	89.4	0.56	0.45	0.72
	OA	75.5				
25PDB	All-α	86.5	94.2	0.85	0.80	0.90
	All-β	80.8	92.4	0.80	0.73	0.87
	α/β	77.2	93.5	0.76	0.70	0.85
	$\alpha+\beta$	60.1	88.0	0.62	0.49	0.74
	OA	76.1				

achieve up to 75.5% and 76.1% overall accuracies for 1189 and 25PDB datasets, respectively. For 1189 and 25PDB datasets, comparing the four structural classes to each other, the values of Sens, Spec, F-measure, MCC and AUC in the all-α class, all-β class and α/β class are obviously and separately superior to those of $\alpha+\beta$ class. The fact indicates that there are still many challenges in the future study to improve the prediction accuracy of $\alpha+\beta$ class.

3.3 Performance comparison between NMF and PCA

PCA is a statistical technique that is widely used in face recognition and image compression. It is useful when the number of variables is large and there is some redundancy in the data. The main advantage of the PCA is that it reduces the dimensionality but often does not lose much information. To investigate the superiority of NMF for our exacted features, we compare the accuracy obtained by NMF with that obtained by PCA using the same dimension of 200 for 1189 and 25PDB datasets, the results are shown in Fig. 3 and fully demonstrate that the NMF is more suitable and successful for our proposed method.

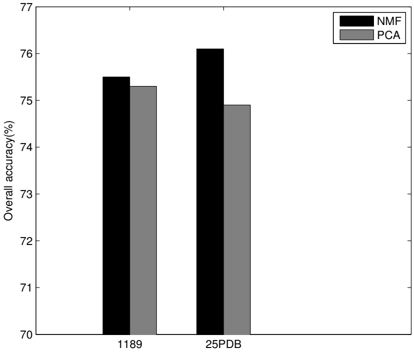

Figure 3: Performance comparison of NMF and PCA.

3.4 Performance comparison with other methods

We compare our results with previous results on the same datasets of 1189 and 25PDB. We select the accuracy of each class and overall accuracy as evaluation indexes that are listed in Table 3. The compared methods include the famous methods SCPRED [46] and MODAS [47], SCPRED mainly based on the information extracted from the predicted protein secondary structure sequence, MODAS combines evolutionary profiles and predicted secondary structure. Hence, SCPRED and MODAS are listed in Table 3 only as two reference methods. AAD-CGR [48] is other famous method and proposed to analyze amino acids sequence by recurrence quantification analysis based on chaos game representation. SCEC [49]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 [50], AADP-PSSM [51], AAC-PSSM-AC [14], AATP [52], PsePSSM [53], and MEDP [54], which are recently reported protein structural classes prediction methods based on the evolutionary information represented in the form of PSSM. RPSSM and PsePSSM are submodels from PSSS-PSSM [50]and

PSSS-PsePSSM [53], respectively.

Table 3 Performance comparison of different methods on three datasets.

Dataset	Method	Prediction	accuracy(%)			
		All-α	All-β	α/β	$\alpha+\beta$	OA(%)
1189	SCPRED [46]	89.1	86.7	89.6	53.8	80.6
	MODAS [47]	92.3	87.1	87.9	65.4	83.5
	AAD-CGR [48]	62.3	67.7	66.5	63.1	65.2
	RPSSM [50]	67.7	75.2	74.6	17.4	60.2
	AADP-PSSM [51]	69.1	83.7	85.6	35.7	70.7
	AATP [52]	72.7	85.4	82.9	42.7	72.6
	MEDP [54]	85.2	84.0	84.3	45.2	75.8
	PsePSSM [53]	82.0	82.3	84.1	44.0	74.4
	AAC-PSSM-AC [14]	80.7	86.4	81.4	45.2	74.6
	This paper	**77.1**	**86.7**	**79.9**	**53.9**	**75.5**
25PDB	SCPRED [46]	92.6	80.1	74.0	71.0	79.7
	MODAS [47]	92.3	83.7	81.2	68.3	81.4
	AAD-CGR [48]	64.3	65.0	65.0	61.7	64.0
	SCEC [49]	75.8	75.2	82.6	31.8	67.6
	RPSSM [50]	75.6	70.2	52.0	43.3	60.8
	AADP-PSSM [51]	83.3	78.1	76.3	54.4	72.9
	AATP [52]	81.9	74.7	75.1	55.8	71.7
	MEDP [54]	87.8	78.3	76.0	57.4	74.8
	AAC-PSSM-AC [14]	85.3	81.7	73.7	55.3	74.1
	PsePSSM [53]	86.2	78.8	75.7	57.6	75.5
	This paper	**86.5**	**80.8**	**77.2**	**60.1**	**76.1**

From Table 3, among the six PSSM-based methods, our method achieves the highest overall prediction accuracies with improvement of 0.9-15.3% for 1189 dataset except MEDP method, and the accuracy of our method is only lower 0.3% than that of MEDP. Referring to all-β class and $\alpha+\beta$ class, our method achieves the highest results. Although the accuracies of all-α and α/β classes are not the highest, our method still obtains the satisfactory results. For 25PDB dataset, our method achieves the highest overall prediction accuracies of the six PSSM-based methods with improvement of 0.6-15.3%. For all-α, α/β and $\alpha + \beta$ class, our method achieves the highest results, although the accuracy of all-β class is lower 0.9% than that of AAC-PSSM-AC. The results sufficiently show that our proposed method successfully extracts the information hidden in the PSSM. In this paper, the most contribution of our proposed method is the improvement of the accuracy using correlation analysis and nonnegative matrix factorization.

4 Conclusions

In this work, we construct a 3600D feature vector by defining auto-cross correlation transformation on the PSSM. Then we reduce dimension of inputting vector, improve calculating efficiency and extract significant classify information by nonnegative matrix factorization(NMF). NMF is also firstly applied in protein structural classes prediction successfully. The SVM classifier and the jackknife test are employed to predict and evaluate the method on two benchmark datasets: 1189 and 25PDB datasets, with sequence similarity lower than 40% and 25%, respectively. The experiments indicate that our approach is convenient, effective and excellent in improving the overall predicting accuracy of protein structural classes prediction. We shall make efforts in our future study to provide a public accessible web-server for our proposed method. Researchers can request for the codes of this task from the corresponding author.

Acknowledgements: The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript. This work was supported by the National Natural Science Foundation of China (No. 61373174), the Fundamental Research Funds for the Central Universities (No. JB140703), and the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2015JQ1010).

References

[1] M. Levitt, C. Chothia, Structural patterns in globular proteins, *Nature* **261** (1976) 552–557.

[2] K. C. Chou, Progress in protein structural class prediction and its impact to bioinformatics and proteomics, *Curr. Protein Pept. Sci.* **6** (2005) 423–436.

[3] G. P. Zhou, An intriguing controversy over protein structural class prediction, *J. Protein Chem.* **17** (1998) 729–738.

[4] K. C. Chou, A key driving force in determination of protein structural classes, *Biochem. Biophys. Res. Commun.* **264** (1999) 216–224.

[5] Y. D. Cai , X. J. Liu, X. B. Xu, K. C. Chou, Prediction of protein structural classes by support vector machines, *Comput. Chem.* **26** (2002) 293–296.

[6] T. L. Zhang, Y. S. Ding, Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes, *Amino Acids* **33** (2007) 623–629.

[7] X. Xiao, S. H. Shao, Z. D. Huang, K.C. Chou, Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor, *J. Comput. Chem.* **27** (2006) 478–482.

[8] T. L. Zhang, Y. S. Ding, K. C. Chou, Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern, *J. Theor. Biol.* **250** (2008) 186–193.

[9] R. Y. Luo, Z. P. Feng, J. K. Liu, Prediction of protein structural class by amino acid and polypeptide composition, *Eur. J. Biochem.* **269** (2002) 4219–4225.

[10] X. D. Sun, R. B. Huang, Prediction of protein structural classes using support vector machines, *Amino Acids* **30** (2006) 469–475.

[11] K. C. Chou, Y. D. Cai, Predicting protein structural class by functional domain composition, *Biochem. Biophys. Res. Commun.* **321** (2004) 1007–1009.

[12] T. G. Liu, X. Q. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, *Biochimie* **92** (2010) 1330–1334.

[13] Y. H. Yao, Z. X. Shi, Q. Dai, Apoptosis protein subcellular location prediction based on position-specific scoring matrix, *J. Comput. Theor. Nanos.* **10** (2014) 1–6.

[14] T. G. Liu, X. B. Geng, X. Q. Zheng, R. S. Li, J . Wang, Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles, *Amino Acids* **42** (2012) 2243–2249.

[15] X. Q. Liu, Q. Dai, L. H. Li, Z. R. He, An efficient binomial model–based measure for sequence comparison and its application, *J. Biomol. Struct. Dyn.* **28** (2011) 833–843.

[16] Y. H. Yao, S. J. Yan, J. N. Han, Q. Dai, P. A. He, A novel descriptor of protein sequences and its application, *J. Theor. Biol.* **347** (2014) 109–117.

[17] H. B. Shen, K. C. Chou, NUC-PLOC: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM, *Protein Eng. Des. Sel.* **20** (2007) 561–567.

[18] G. L. Fan, Q. Z. Li, Predicting submitochondria locations by combining different descriptors into the general form of Chou's Pseudo amino acid composition, *Amino Acids* **43** (2012) 545–555.

[19] S. L. Zhang, S. Y. Ding, T. M. Wang, High–accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure, *Biochimie* **93** (2011) 710–714.

[20] Q. Dai, Y. Li, X. Q. Liu, Y. H. Yao, Y. G. Cao, P. G. He, Comparison study on statistical features of predicted secondary structures for protein structural class prediction: From content to position, *BMC Bioinformatics* **14** (2013) #152.

[21] Z. C. Li, X. B. Zhou, Z. Dai, X. Y. Zou, Prediction of protein structural classes by Chou's pseudo amino acid composition: approached using continuous wavelet transform and principal component analysis, *Amino Acids* **37** (2009) 415–425.

[22] L. Li, X. Cui, S. Yu, Y. Zhang, Z. Luo, H. Yang, Y. Zhou, X. Zheng, PSSP-RFE: Accurate prediction of protein structural class by recursive feature extraction from PSI-BLAST profile, physical–chemical property and function annotations, *PloS One* **9** (2014) #e92863.

[23] S. Y. Ding, S. J. Yan, S. H. Qi, Y. Li, Y. H. Yao, A protein structural classes prediction method based on PSI-BLAST profile, *J. Theor. Biol.* **353** (2014) 19–23.

[24] D. Cai, G. P. Zhou, Prediction of protein structural classes by neural network, *Biochimie* **82** (2000) 783–785.

[25] C. Chen, Y. X. Tian, X. Y. Zou, P. X. Cai, J. Y. Mo, Using pseudo–amino acid composition and support vector machine to predict protein structural class, *J. Theor. Biol.* **243** (2006) 444–448.

[26] Z. X. Shi, Q. Dai, P. A. He, Y. H. Yao, B. Liao, Subcellular localization prediction of apoptosis proteins based on the data mining for amino acid index database, *IEEE 7th International Conference on Systems Biology (ISB)* (2013) 43–48.

[27] H. B. Shen, J. Yang, X. J. Liu, K. C. Chou, Using supervised fuzzy clustering to predict protein structural classes, *Biochem. Biophys. Res. Commun.* **334** (2005) 577–581.

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

[29] Y. F. Cao, S. Liu, L. D. Zhang, J. Qin, J. Wang, K. X. Tang, Prediction of protein structural class with rough sets, *BMC Bioinform.* **7** (2006) #20.

[30] T. L. Zhang, K. C. Chou, prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern, *J. Theor. Biol.* **250** (2008) 186–193.

[31] D. D. Lee, H. S. Seung, Learning the parts of objects by nonnegative matrix factorization, *Nature* **401** (1999) 788–791.

[32] D. D. Lee, H. S. Seung, Algorithms for non–negative matrix factorization, in: T. K. Leen, T. G. Dietterich, V. Tresp (Eds.), *Advance in Neural Information Processing Systems*, MIT Press, 2001, pp 556–562.

[33] P. Carmona–Sacs, R. D. Paseual–Marqui, F. Tirado, J. M. Carazo, A. Pascual–Montano, Biclustering of gene expression data by non-smooth non-negative matrix factorization, *BMC Bioinformatics* **7** (2006) #78.

[34] J. P. Brunet, P. Tamayo, T. R. Golub, J. P. Mesirov, Metagenes and molecular pattern discovery using matrix factorization, *PNAS* **101** (2004) 4164–4169.

[35] Y. Gao, G. Church, Improving molecular cancer class discovery through sparse non-negative matrix factorization, *Bioinformatics* **21** (2005) 3970–3975.

[36] I. Y. Jung, J. Y. Lee, S. Y. Lee, D. Kim, Application of nonnegative matrix factorization to improve profile–profile alignment features for fold recognition and remote homolog detection, *BMC Bioinformatics* **9** (2008) #298.

[37] L. A. Kurgan, L. Homaeian, Prediction of structural classes for protein sequences and domains – Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy, *Pattern Recogn.* **39** (2006) 2323–2343.

[38] K. C. Chou, Some remarks on protein attribute prediction and pseudo amino acid composition, *J. Theor. Biol.* **273** (2011) 236–247.

[39] S. F. Altschul, T. L . Schaffer, A .A . Madden, J . Zhang, Z . Zhang, W. Miller, D. J. Lipman, Gapped BlAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* **25** (1997) 3389–3402.

[40] Y. Y. Liang, S. Y. Liu, S. L. Zhang, Prediction of protein structural class based on different autocorrelation descriptors of position–specific scoring matrix, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 765–784.

[41] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer, New York, 1995.

[42] S. L. Zhang, Y. Y. Liang, Z. G. Bai, A novel reduced triplet composition based method to predict apoptosis protein subcellular localization, *MATCH Commun. Math. Comput. Chem.* **73** (2015) 559–571.

[43] J. Y. Yang, X. Chen, Improving taxonomy–based protein fold recognition by using global and local features, *Proteins* **79** (2011) 2053–2064.

[44] C. C. Chang, C. J. Lin, LIBSVM: a library for support vector machines, *ACM Trans. Intell. Syst. Technol.* **2** (2011) #27.

[45] K. C. Chou, H. B. Shen, Recent progress in protein subcellular location prediction, *Anal Biochem.* **370** (2007) 1–16.

[46] L. Kurgan, K. Cios, K. Chen, SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences, *BMC Bioinformatics* **9** (2008) #226.

[47] M. J. Mizianty, L. Kurgan, Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences, *BMC Bioinformatics* **10** (2009) #414.

[48] J. Y. Yang, Z. L. Peng, Z. G. Yu, R. J. Zhang, V. Anh, D. S. Wang, Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation, *J. Theor. Biol.* **257** (2009) 618–626.

[49] K. Chen, L. A. Kurgan, J. S. Ruan, Prediction of protein structural class using novel evolutionary collocation-based sequence representation, *J. Comput. Chem.* **29** (2008) 1596–1604.

[50] S. Y. Ding, Y. Li, Z. X. Shi, S. J. Yan, A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile, *Biochimie* **97** (2014) 60–65.

[51] T. G. Liu, X. Q. Zheng, J. Wang, Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile, *Biochimie* **92** (2010) 1330–1334.

[52] S. L. Zhang, Y. Feng, X. G. Yuan, Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM, *J. Biomol. Struct. Dyn.* **29** (2012) 634–642.

[53] S. L. Zhang, Accurate prediction of protein structural classes by incorporating PSSS and PSSM into Chou's general PseAAC, *Chemometr. Intell. Lab.* **142** (2015) 28–35.

[54] L. C. Zhang, X. Q. Zhao, L. Kong, Predict protein structural class for low-similarity sequences by evolutionary difference information into the general form of Chou's pseudo amino acid composition, *J. Theor. Biol.* **355** (2014) 105–110.