Interdiscip Sci Comput Life Sci (2014) 6: 235–240
DOI: 10.1007/s12539-013-0205-6

Prediction of Protein Structural Classes Based on Feature Selection Technique

Hui Ding[1], Hao Lin[1,*], Wei Chen[2,*], Zi-Qiang Li[3], Feng-Biao Guo[1], Jian Huang[1], Nini Rao[1]
[1](Key Laboratory for NeuroInformation of Ministry of Education, Center of Bioinformatics, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China)
[2](Department of Physics, Center for Genomics and Computational Biology, College of Sciences, Hebei United University, Tangshan 063000, China)
[3](School of information and Engineering, Sichuan Agricultural University, Yaan 625014, China)

Received 7 September 2013 / Revised 2 October 2013 / Accepted 12 November 2013

Abstract: The prediction of protein structural classes is beneficial to understanding folding patterns, functions and interactions of proteins. In this study, we proposed a feature selection-based method to accurately predict protein structural classes. Three datasets with sequence identity lower than 25% were used to test the prediction performance of the method. Through jackknife cross-validation, we have verified that the overall accuracies of these three datasets are 92.1%, 89.7% and 84.0%, respectively. The proposed method is more efficient and accurate than other existing methods. The present study will offer an excellent alternative to other methods for predicting protein structural classes.

Key words: protein structural class, feature selection technique, support vector machine, tetrapeptide.

1 Introduction

The prediction of protein structure based on amino acid sequences is one of the most important problems in theoretical biology. Generally, the globular protein domains can be categorized into all-α, all-β, $\alpha+\beta$ and α/β according to the types and arrangements of their secondary structural elements (Levitt and Chothia, 1976). The knowledge of protein structural class can provide valuable information to reduce the scope of conformational search during energy optimization.

Since the concept of protein structural class was proposed three decades ago, it has been used as an important attribute to characterize the overall folding type of a protein or a domain. The prediction of protein structural classes has become one of the hotspots in bioinformatics and has attracted the attention from lots of structural biologists and bioinformatics scholars (Lin *et al.*, 2012; Ding *et al.*, 2012; Qin *et al.*, 2012; Chen *et al.*, 2008, 2012; Dai *et al.*, 2011; Liu *et al.*, 2010, 2011; Yang *et al.*, 2009, 2010; Liu and Jia, 2010; Mizianty and Kurgan, 2009; Li *et al.*, 2009; Costantini and Facchiano, 2009; Kurgan *et al.*, 2008a, b; Yu *et al.*, 2007; Lin and Li, 2007; Kurgan and Homaeian, 2006; Cai *et al.*, 2000; Bu *et al.*, 1999; Zhou, 1998). A large number of machine learning algorithms were developed for protein structural class prediction, such as: support vector machine (SVM) (Lin *et al.*, 2012; Ding *et al.*, 2012; Qin *et al.*, 2012; Chen *et al.*, 2012; Dai *et al.*, 2011; Liu *et al.*, 2011), Fisher's discriminant (FD) (Yang *et al.*, 2009, 2010), increment of diversity combined with quadratic discriminant (IDQD) (Lin and Li, 2007), neural networks (NNs) (Cai *et al.*, 2000), etc. For machine-learning-based methods, a key step is to transform protein sequences into feature vectors with fixed length. For the description of protein sequences, various feature extraction methods were proposed to discretely describe protein sequences. Because the structure class of a protein is closely related to its primary sequence, the amino acid compositions (AAC) were selected as the inputs of predictors (Bu *et al.*, 1999; Zhou, 1998). To catch the sequence-order information, the dipeptide compositions were added into feature vectors (Lin and Li, 2007). And polypeptide compositions were also used to enhance the predictive power of predictors (Yu *et al.*, 2007). It is well known that the physicochemical properties of 20 amino acids can influence the fold mode of proteins. Hence, the pseudo amino acid compositions (PseAAC) were developed to describe not only amino acid composition, but also the long-distance interaction of physicochemical properties among residues (Qin *et*

*Corresponding authors.
E-mail: hlin@uestc.edu.cn,
greatchen@heuu.edu.cn

al., 2012; Chen *et al.*, 2012).

Although these features have achieved encouraging results for high-identity (≤40% identity) datasets, most of them were not applicable for the datasets with low identity (≤25% identity). To overcome this limitation, some informative parameters were proposed for the low-identity datasets. Since the PSI-BLAST program can find distant relatives of a protein and is much more sensitive in picking up distant evolutionary relationships, position-specific score matrix (PSSM) extracted from the PSI-BLAST profile were proposed for prediction (Liu *et al.*, 2010, 2011). However, the predicted accuracy was only slightly improved. Because the types and arrangements of secondary structural elements of four structural classes are different (Levitt and Chothia, 1976), the predicted secondary structural features (PSSF) were used to improve prediction performance for low-similarity sequences (Ding *et al.*, 2012; Dai *et al.*, 2011; Yang *et al.*, 2010; Liu and Jia, 2010; Mizianty and Kurgan, 2009; Kurgan *et al.*, 2008a). And the prediction results were dramatically improved. However, the results are largely dependent on the results derived from secondary structure prediction. Furthermore, the accuracies are still far from satisfactory. The way to extract representational parameters of proteins is the most important problem for protein structural class prediction.

As the structural unit of alpha helix, tetrapeptide plays a crucial role in the formation of the regular structure for the hydrogen bonds in the helix connecting the ith residue with the (i+4)th residue (Qi *et al.*, 2012). Rackovsky has estimated that 60-70% of tetrapeptides encode specific structures (Rackovsky, 1993). These tetrapeptides can be regarded as the protein folding code and has been applied in protein secondary structure prediction through feature selection technique (Feng and Luo, 2008). Therefore, in the paper, we introduced a novel method based on the tetrapeptide signals selected by binomial distribution to predict protein structural classes. To test the method and facilitate the comparison with previous studies, three benchmark datasets were selected to evaluate the prediction performance of the method. The high prediction accuracy indicates that the tetrapeptide signals are important in protein structural class prediction.

2 Materials and methods

2.1 Database

Three low-identity benchmark datasets are selected to test the proposed method. The first dataset is called 25PDB which contains 443 all-α proteins, 443 all-β proteins, 346 α/β proteins and 441 $\alpha+\beta$ proteins (Yang *et al.*, 2010; Kurgan and Homaeian, 2006). The second dataset is called 640 which contains 138 all-α proteins, 154 all-β proteins, 177 α/β proteins and 171 $\alpha+\beta$ proteins (Yang *et al.*, 2010; Chen *et al.*, 2008). The sequence identity of the above two datasets is less than 25%. The third dataset is called ACS dataset which contains 124 all1-α proteins, 112 all-β proteins and 163 mixed $\alpha\beta$ proteins (Lin *et al.*, 2012).

2.2 Tetrapeptide signals

By sliding the window of four residues with the step of one residue along a protein sequence, we obtain 160 000 kinds of tetrapeptides and convert them into the following vector:

$$F = [f_1, f_2, \cdots, f_i, \cdots, f_{160000}]^T \tag{1}$$

where, symbol **T** denotes the transposition of the vector; f_i is the frequency of the ith tetrapeptide and expressed as:

$$f_i = n_i / \sum\nolimits_{j=1}^{160000} n_j = n_i/(L-3) \tag{2}$$

where, n_i and L denote the number of the ith tetrapeptide and the length of the protein, respectively.

The vector F is a high dimensional vector. The prediction of protein structural classes with high dimensional features will result in three problems. Firstly, over-fitting results in low generalization ability and overestimation of prediction model. Secondly, information redundancy or noise results in poor prediction accuracy and erroneous description of intrinsic properties. Thirdly, dimension disaster results in a handicap for the computation or the increase in computational time. The binomial distribution describes the outcome of n independent trials in an experiment. Thus, we use it to optimize feature set (Feng and Luo, 2008).

It is supposed that the ith tetrapeptide occurs n_{ij} times in structural class j and N_i times in benchmark dataset. If its occurrence in structural class j is a stochastic event, the confidence level of n_{ij} or more times of the ith tetrapeptide occurrence in structural class j (CL_{ij}) can be calculated as follows:

$$CL_{ij} = 1 - \sum_{n=n_{ij}}^{N_i} \frac{N_i!}{n!(N_i-n)!} p_j^n (1-p_j)^{N_i-n} \tag{3}$$

where, p_j is the prior probability of tetrapeptides of structural class j in benchmark dataset and can be defined as:

$$p_j = \sum\nolimits_{i=1}^{160000} n_{ij} \Big/ \sum\nolimits_{i=1}^{160000} N_i \tag{4}$$

where, $\sum_{i=1}^{160000} N_i$ and $\sum_{i=1}^{160000} n_{ij}$ are the total occurrence times of all the tetrapeptides in the dataset and structural class j, respectively.

As CL_{ij} tends to 1, the occurrence of the ith tetrapeptide in structural class j for n_{ij} times should not be a random event. If there are m tetrapeptides whose CLs are larger than a given cutoff CL_o, the frequencies of these tetrapeptides are selected as the optimized features and expressed as:

$$F_m = [f_1, f_2, \cdots, f_i, \cdots, f_m]^T \tag{5}$$

If CL_o is set to zero, all the 160 000 tetrapeptides are selected. If $CL_o > 1$, no tetrapeptide is selected. According to this process, high-dimensional data can be projected into low-dimensional space. The parameter m or CL_o may be selected through cross-validation.

2.3 The definition of tendency to secondary structure

Based on the Chou-Fasman's conformational parameters of secondary structure (Prevelige Jr and Fasman, 1989), the tendency of an arbitrary tetrapeptide j is defined as:

$$Tendency_j = \sum\nolimits_{i=1}^{4} f_{\alpha}^{ij} - \sum\nolimits_{i=1}^{4} f_{\beta}^{ij} \tag{6}$$

where, f_{α}^{ij} and f_{β}^{ij} are respectively the helix propensity and sheet propensity for the ith residue in the jth tetrapeptide. The larger the $Tendency_j$ is, the easier the helix formation with the jth tetrapeptide, and vice versa.

2.4 Support vector machine

SVM is a popular and wonderful supervised machine learning technique. The basic idea of SVM is to map the data of samples into a high dimensional Hilbert space and to seek a separating hyperplane in this space. Here, the free software tool box LibSVM (Fan *et al.*, 2005) was adopted to implement SVM. The radial basis function (RBF) was selected as the kernel function. The one-versus-one (OVO) strategy is used for multiclass classification. The regularization parameter C and kernel parameter γ were optimized through grid search with cross-validation.

2.5 Performance evaluation

As a kind of statistical test method, the cross-validation test has been widely applied in bioinformatics prediction. Since the jackknife cross-validation can yield unique results and has been employed to estimate the performance of other methods of protein structural class prediction (Lin *et al.*, 2012; Ding *et al.*, 2012; Qin *et al.*, 2012; Chen *et al.*, 2012; Dai *et al.*, 2011; Liu *et al.*, 2011), we adopt it to evaluate the performance of our method. The two parameters of sensitivity (Sn) and overall accuracy (Oa) are used to describe the prediction accuracy of our method.

$$Sn_j = TP_j/(TP_j + FN_j) \tag{7}$$

$$Oa = \sum\nolimits_{j=1}^{N} TP_j/N \tag{8}$$

where, TP_j and FN_j respectively denote true positives and false positives of the jth structural class; N is the number of samples.

3 Results and discussion

3.1 Accuracy

In general, the tetrapeptides with a high confidence level will give more reliable information for classification. However, the number of these tetrapeptides is too small to afford enough information, leading to poor prediction accuracy. In contrast, the tetrapeptide set with low confidence level contains too many components, which reduces the cluster-tolerant capacity and lower the cross-validation accuracy. Therefore, it is necessary to adopt the optimized tetrapeptides to perform prediction. Through adjusting the cutoff of confidence level CL_o from 1 to 0, we obtained a series of tetrapeptide sets. After inputting these feature sets into SVM and examining their prediction accuracies, we picked out the optimized feature sets with the highest accuracy. In order to reduce time and improve efficiency, the CL_o, the regularization parameter C and the kernel parameter were optimized through the five-fold cross-validation.

For the three benchmark datasets, 25PDB, 640 and ACS datasets, the processes of feature selection are the same, but the optimized confidence levels CL_o are different. When the optimized CL_os were set to 85.0%, 87.2% and 88.9%, the highest accuracies were obtained for 25PDB, 640 datasets and ACS datasets, respectively. The results were recorded in Table 1. For comparison, the results of other existing methods were also provided in Table 1. As shown in Table 1, our method obtained the overall accuracies of 92.1%, 89.7% and 84.0% for 25PDB, 640 and ACS datasets, respectively. Several studies indicated that it was difficult to discriminate α/β class from $\alpha+\beta$ class (Ding *et al.*, 2012; Yang and Chen, 2010; Kurgan *et al.*, 2008a). However, the prediction accuracy of α/β class and $\alpha+\beta$ class obtained by the proposed method is higher than 89% (Table 1). These results are better than previous results, suggesting that the proposed method is superior to other methods.

The representation of protein sequence is a key step for protein structural class prediction. For high-identity dataset, primary sequence information, such as amino acid composition, dipeptide composition, physiochemical properties of amino acids and PSI-BLAST profiles, can achieve high accuracy. However, for low-identity datasets, these parameters are not available and the predicted accuracies are generally less than 75% (Table 1). Thus, the predicted secondary structural feature derived from software PSIPRED (McGuffin *et al.*, 2000) is proposed to represent proteins in previous studies

Table 1 Comparison of different methods on two benchmark datasets

Dataset	Method	Sensitivity (%)				Overall accuracy (%)
		All-α	All-β	α/β	$\alpha+\beta$	
25PDB	WSVM (PseAAC) (Chen *et al.* 2012)	59.1	60.0	58.1	50.1	56.8
	LR (physicochemical properties) (Kurgan and Homaeian 2006)	69.1	61.6	60.1	38.3	57.1
	KNN (specific peptide frequencies) (Costantini and Facchiano 2009)	60.6	60.7	67.9	44.3	58.6
	SVM (PseAAC+Over represented k-mers) (Qin *et al.* 2012)	83.3	50.8	59.8	44.4	59.6
	linear LR (physicochemical properties) (Kurgan and Chen 2007)	77.4	66.4	61.3	45.4	62.7
	FD (chaos game representation) (Yang *et al.* 2009)	64.3	65.0	65.0	61.7	64.0
	SVM (PseAAC) (Li *et al.* 2009)	76.5	67.3	66.8	45.8	64.0
	SVM (PSSM) (Liu *et al.* 2010)	83.3	78.1	76.3	54.4	72.9
	SVM (PSSM) (Liu *et al.* 2011)	85.3	81.7	73.7	55.3	74.1
	DT (PSSF)(Kurgan *et al.* 2008b)	90.7	80.1	68.8	64.1	76.4
	SCPRED, SVM (PSSF) (Kurgan *et al.* 2008a)	92.6	80.1	74.0	71.0	79.7
	MODAS, SVM (PSSF) (Mizianty and Kurgan 2009)	92.3	83.7	81.2	68.3	81.4
	FD (PSSF) (Yang and Chen 2010)	92.8	83.3	85.8	70.1	82.9
	SVM (PSSF) (Liu and Jia, 2010)	92.6	81.3	81.5	76.0	82.9
	SVM (PSSF) (Dai *et al.* 2011)	90.1	84.7	79.5	77.6	83.1
	SVM (PSSF) (Ding *et al.* 2012)	95.0	81.3	83.2	77.6	84.3
	SVM (optimized tetrapeptides) (This paper)	97.5	90.5	89.3	90.2	92.1
640	SVM (PSSM) (Chen *et al.* 2008)	73.9	61.0	81.9	33.9	62.3
	SCPRED, SVM (PSSF) (Kurgan *et al.* 2008a)	90.6	81.8	85.9	66.7	80.8
	FD (PSSF) (Yang *et al.* 2010)	89.1	85.1	88.1	71.4	83.1
	SVM (PSSF) (Ding *et al.* 2012)	94.9	76.6	89.3	74.3	83.4
	SVM (PSSF) (Dai *et al.* 2011)	91.3	87.7	88.7	81.3	87.0
	SVM (optimized tetrapeptides) (This paper)	94.2	78.6	93.2	92.4	89.7
ACS	SVM (Dipeptides + AAC) (Lin *et al.* 2012)	49.2	35.7	66.9		52.6
	SVM (PseAAC) (Lin *et al.* 2012)	68.5	59.8	65.6		64.9
	SVM (averaged chemical shift) (Lin *et al.* 2012)	92.7	77.7	91.4		88.0
	SVM (optimized tetrapeptides) (This paper)	91.9	75.0	84.0		84.0

(Ding *et al.*, 2012; Dai *et al.*, 2011; Yang *et al.*, 2010; Liu and Jia, 2010; Mizianty and Kurgan, 2009). The results demonstrated that this feature did improve the performance of predictive models and that the accuracies were usually higher than 80% (Ding *et al.*, 2012; Dai *et al.*, 2011; Yang *et al.*, 2010; Liu and Jia, 2010; Mizianty and Kurgan, 2009). However, the accuracy of PSIPRED is only about 80%. If the secondary structure of a protein chain is not correctly predicted, it will lead to the incorrect prediction of the structural class of this protein. Thus, without other information, the models based on the secondary structural information can not achieve higher accuracies. Moreover, 60%-70% of tetrapeptides encode the specific secondary structure (Rackovsky, 1993). Therefore, we directly extracted tetrapeptide information from primary sequence to represent protein samples. In this way, we can avoid potential misleading information from secondary structure prediction programs. And our results demonstrated that the proposed feature was a powerful feature for protein structural class prediction.

Because the occurrence frequency of a tetrapeptide in random sequence is very low (1/160 000), particular tetrapeptides within a protein may contribute to the particular functional role of the protein, other than the result of random selection. Tripeptides could be used in prediction. However, tripeptides appear about 20 times more than tetrapeptides. Hence, they would bring more noise into prediction. For example, the optimized tripeptides produced by the binomial distribution can only achieve the overall accuracies of 69.4% for 25PDB dataset and 81.7% for 640 dataset. These results further suggest that the tetrapeptides are suitable for the protein structural class prediction. Shafiullah and Al-Mamun (Shafiullah and Al-Mamun, 2010) have also used tetrapeptides to predict protein structural classes. However, they didn't give a strict statistical analysis to optimize tetrapeptides. Moreover, they didn't demonstrate the performance of their method by comparing with other published methods. Finally, they didn't provide any analysis to explain why the tetrapeptides were selected.

3.2 Tendency to secondary structure

It is also indicated that the selected tetrapeptides play important roles in the protein structural class prediction. Thus, we calculated the tendencies of forming α helix or β sheet for all the over represented tetrapeptides in different structural classes according to the Chou-Fasman's conformational parameters (Kurgan and Chen, 2007). The tendency of an arbitrary tetrapeptide is defined as the difference between the sum of helix propensities and the sum of sheet propensities of four residues in the tetrapeptide. The averaged tendencies for different classes were recorded in Table 2. The selected tetrapeptides in all-α class have the maximum averaged tendency and the selected tetrapeptides in all-β class have the minimum one. The tendency of tetrapeptides to form alpha helices is meaningful and the reason is the ith to (i-4)th hydrogen bonds connections. The beta-structural form can also be recognized as longer polypeptides (the phi and psi of an isolated amino acid or a dipeptide can not be classified as an orderly form) (Meus *et al.*, 2006). This is the reason why the tetrapeptide is selected as the unit for protein structural class prediction. In fact, to determine the evolutionary relationship between proteins, the protein structural domains can be classified at lower levels such as fold, superfamily, family and domain levels based on similarities of their structures and amino acid sequences. Thus, the feature selection method proposed in this paper can also be used for classifying protein structures at lower levels in the hierarchy.

Table 2 The averaged tendencies for selected tetrapeptides in different structural classes

dataset	All-α	All-β	α/β	$\alpha+\beta$
25PDB	0.4133	−0.1488	0.1201	0.1414
640	0.4687	−0.1077	0.2247	0.1492
ACS	0.7330	−0.0492	0.5354	

4 Conclusion

In this paper, we developed a novel method based on selected tetrapeptides to predict protein structural classes. The statistical significance of each tetrapeptide was given by the binomial distribution. Based on the statistical significance, the feature set was optimized. High prediction accuracy obtained in the two benchmark datasets with low identity indicates that the method is superior to other existing methods for protein structural classification. By calculating tendencies of forming α helix or β sheet for all the over represented tetrapeptides, we provided the reason why the tetrapeptide was selected as the unit for protein structural class prediction.

Acknowledgements

This work was supported by Grants from The National Nature Scientific Foundation of China (No. 61202256, 61100092, 61301260) and the Fundamental Research Funds for the Central Universities (ZYGX2012J113, ZYGX2013J102).

Conflict of Interest

The authors have declared that no competing interests exist.

References

[1] Bu, W.S., Feng, Z.P., Zhang, Z., Zhang, C.T. 1999. Prediction of protein (domain) structural classes based on amino-acid index. Eur J Biochem 266, 1043-1049.

[2] Cai, Y.D., Li, Y.X., Chou, K.C. 2000. Using neural networks for prediction of domain structural classes. Biochim Biophys Acta 3, 1-2.

[3] Chen, C., Shen, Z.B., Zou, X.Y. 2012. Dual-layer Wavelet SVM for Predicting Protein Structural Class via the General Form of Chou's Pseudo Amino Acid Composition. Protein Pept Lett 19, 422-429.

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

[5] Costantini, S., Facchiano, A.M. 2009. Prediction of the protein structural class by specific peptide frequencies. Biochimie 91, 226-229.

[6] Dai, Q., Wu, L., Li, L. 2011.Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features. J Comput Chem 32, 3393-3398.

[7] Ding, S., Zhang, S., Li. Y., Wang, T. 2012. A novel protein structural classes prediction method based on predicted secondary structure. Biochimie 94, 1166-1171.

[8] Fan, R.E., Chen, P.H., Lin, C.J. 2005. Working set selection using the second order information for training SVM. J Mach Learn Res 6, 1889-1918.

[9] Feng, Y., Luo, L. 2008. Use of tetrapeptide signals for protein secondary-structure prediction. Amino Acids 35, 607-614.

[10] Kurgan, L., Chen, K. 2007. Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun 357, 453-460.

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

[12] Kurgan, L., Homaeian, L. 2006. Prediction of structural classes for protein sequences and domains—Impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy. Pattern Recog 39, 2323-2343.

[13] Kurgan, L., Zhang, T., Zhang, H., Shen, S., Ruan, J. 2008b. Secondary structure-based assignment of the protein structural classes. Amino Acids 35, 551-564.

[14] Levitt, M., Chothia, C. 1976. Structural patterns in globular proteins. Nature 261, 552-558.

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

[16] Lin, H., Ding, C., Song, Q., Yang, P., Ding, H., Deng, K.J, Chen, W. 2012. The prediction of protein structural class using averaged chemical shifts. J Biomol Struct Dyn 29, 643-648.

[17] Lin, H., Li, Q.Z. 2007. Using Pseudo Amino Acid Composition to Predict Protein Structural Class: Approached by Incorporating 400 Dipeptide Components. J Comput Chem 28, 1463-1466.

[18] Liu, T., Geng, X., Zheng, X., Li, R., Wang, J. 2011. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243-2249.

[19] Liu, T., Jia, C. 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J Theor Biol 267, 272-275.

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

[21] McGuffin, L.J., Bryson, K., Jones, D.T. 2000. The PSIPRED protein structure prediction server. Bioinformatics 16, 404-405.

[22] Meus, J., Brylinski, M., Piwowar, M., *et al.* 2006. A tabular approach to the sequence-to-structure relation in proteins (tetrapeptide representation) for de novo protein design. Med Sci Monit 12, BR208-214.

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

[24] Prevelige Jr, P., Fasman, G.D. 1989. Chou-Fasman prediction of the secondary structure of proteins, in Prediction of Protein structure and the principles of protein conformation, G.D. Fasman, ed., Plenum Press, New York, pp. 391-416.

[25] Qi, Y., Liang, H., Han, X., Lai, L. 2012. Sequence Preference of α-Helix N-Terminal Tetrapeptide. Protein Pept Lett 345-352.

[26] Qin, Y.F., Wang, C.H., Yu, X.Q., Zhu, J., Liu, T.G., Zheng, X.Q. 2012. Predicting protein structural class by incorporating patterns of over-represented k-mers into the general form of Chou's PseAAC. Protein Pept Lett 19, 388-397.

[27] Rackovsky, S. 1993. On the nature of protein folding code. Proc Natl Acad Sci USA 90, 644-648.

[28] Shafiullah, G.M., Al-Mamun, H.A. 2010. Protein strucutral class prediction using support vector machine. 6th International Conference on Electrical and Computer Engineering 179-182.

[29] Yang, J.Y., Peng, Z.L., Chen, X. 2010. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics 11, S9.

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

[31] Yu, T., Sun, Z.B., Sang, J.P., Huang, S.Y., Zou, X.W. 2007. Structural class tendency of polypeptide: A new conception in predicting protein structural class. Physica A 386, 581-589.

[32] Zhou, G.P. 1998. An intriguing controversy over protein structural class prediction. J Protein Chem 17, 729-738.