Amino Acids (2012) 42:2243–2249
DOI 10.1007/s00726-011-0964-5

ORIGINAL ARTICLE

Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles

Taigang Liu · Xingbo Geng · Xiaoqi Zheng ·
Rensuo Li · Jun Wang

Received: 24 January 2011 / Accepted: 11 June 2011 / Published online: 23 June 2011
© Springer-Verlag 2011

Abstract Computational prediction of protein structural class based solely on sequence data remains a challenging problem in protein science. Existing methods differ in the protein sequence representation models and prediction engines adopted. In this study, a powerful feature extraction method, which combines position-specific score matrix (PSSM) with auto covariance (AC) transformation, is introduced. Thus, a sample protein is represented by a series of discrete components, which could partially incorporate the long-range sequence order information and evolutionary information reflected from the PSI-BLAST profile. To verify the performance of our method, jackknife cross-validation tests are performed on four widely used benchmark datasets. Comparison of our results with existing methods shows that our method provides the state-of-the-art performance for structural class prediction. A Web server that implements the proposed method is freely available at http://202.194.133.5/xinxi/AAC_PSSM_AC/index.htm.

T. Liu · R. Li
College of Information Sciences and Engineering, Shandong
Agricultural University, Taian 271018, China

X. Geng
School of Mathematical Sciences, Dalian University of
Technology, Dalian 116024, China

X. Zheng (✉) · J. Wang
Department of Mathematics, Shanghai Normal University,
Shanghai 200234, China
e-mail: xqzheng@shnu.edu.cn

X. Zheng · J. Wang
Scientific Computing Key Laboratory of Shanghai Universities,
Shanghai 200234, China

Keywords Protein structural class · PSI-BLAST profile · Auto covariance transformation · Support vector machine

Introduction

Knowledge of structural class information of a given protein plays an important role in the prediction of secondary structure, tertiary structure and function analysis from the amino acid sequence (Anand et al. 2008). Based on the visual inspection of polypeptide chain topologies in a dataset of 31 globular proteins, Levitt and Chothia (1976) first introduced the concept of structural class and categorized the protein domains of known structure into four structural classes: all-α, all-β, α/β and $\alpha + \beta$. Nowadays, the most frequently used classification of protein structural classes can be found in the structural classification of proteins (SCOP) database (Murzin et al. 1995), which further divides proteins into 11 structural classes. But currently, the four major structural classes, which cover almost 90% of all SCOP entries, are still commonly adopted by many researchers.

In the first decade of the twenty-first century, prediction of protein structural class from primary sequence data became a hot topic in current bioinformatics, and a great number of statistical learning algorithms were developed. These algorithms include neural network (Cai and Zhou 2000), support vector machine (SVM) (Cai et al. 2001; Chen et al. 2006a; Li et al. 2008; Qiu et al. 2009), fuzzy *k*-nearest neighbor (Zhang et al. 2008, Zheng et al. 2010), fuzzy clustering (Shen et al. 2005), Bayesian classification (Wang and Yuan 2000), logistic regression (Kurgan and Chen 2007; Kurgan and Homaeian 2006), rough sets (Cao et al. 2006) and classifier fusion techniques (Cai et al. 2006; Chen et al. 2006b, 2009; Feng et al. 2005;

Kedarisetti et al. 2006). Among them, SVM is the most popular and the best-performing classifier for this task (Kurgan et al. 2008a). In SVM-based predictive methods, an essential step is to transform protein sequences into fixed-length feature vectors as SVM cannot be directly applied to amino acid sequences with different lengths. Several frequently used sequence representations include amino acid composition (AAC) (Chou 1999; Nakashima et al. 1986; Zhou 1998), pseudo amino acid composition (Chou 2001; Li et al. 2009; Xiao et al. 2008; Zhang and Ding 2007), polypeptide composition (Costantini and Facchiano 2009; Luo et al. 2002; Sun and Huang 2006), functional domain composition (Chou and Cai 2004), amino acid sequence reverse encoding (Yang et al. 2009), etc. Recently, Kurgan et al. proposed to extract features from the predicted secondary structure and PSI-BLAST profile rather than directly from the amino acid sequence itself and reported that a higher prediction accuracy could be consequently achieved (Chen et al. 2008; Kurgan et al. 2008a, b; Mizianty and Kurgan 2009). Motivated by their work, some researchers further improved the prediction accuracy based solely on the predicted secondary structure information (Liu and Jia 2010; Yang et al. 2010). On the other hand, in our previous study (Liu et al. 2010), we extracted AAC and dipeptide composition from the PSI-BLAST profile and also obtained favorable prediction accuracy when the predicted secondary structure was not utilized.

In this study, we try to extract other more informative data solely from the PSI-BLAST profile to further improve the prediction accuracy. First, the position-specific score matrix (PSSM) generated by PSI-BLAST program (Altschul et al. 1997) is transformed into a fixed-length feature vector by auto covariance (AC) transformation. Then, these resulting vectors are input to an SVM classifier to perform the prediction. Jackknife cross-validation tests on four working datasets show that the current method presents satisfying prediction accuracies in comparison with existing methods. A Web server that implements the proposed method is freely available at http://202.194.133.5/xinxi/AAC_PSSM_AC/index.htm.

Materials and methods

Datasets

Two widely studied datasets constructed by Zhou (1998) are used to demonstrate the performance of the proposed method. The first dataset contains 277 domains and the second consists of 498 domains (denoted as *Z277* and *Z498*, respectively). Although two datasets have small size and high similarity, they were used extensively in previous prediction studies. To investigate the effect of sequence similarity on the performance of our method, we also studied two larger and low-similarity datasets: *1189* (Wang and Yuan 2000) and *25PDB* (Kedarisetti et al. 2006; Kurgan and Homaeian 2006), which include 1,092 and 1,673 protein domains with sequence similarity lower than 40 and 25%, respectively. More details about the four datasets are listed in Table 1.

Table 1 The compositions of four datasets adopted in this study

Dataset	All-α	All-β	α/β	$\alpha+\beta$	Total
Z277	70	61	81	65	277
Z498	107	126	136	129	498
1189	223	294	334	241	1,092
25PDB	443	443	346	441	1,673

Protein sequence representation

The representation of a protein sequence by a fixed-length feature vector is one of the primary tasks for most protein classification techniques. In this section, we propose a simple and powerful sequence representation model by combining PSSM and AC transformation.

To extract the evolutionary information, the profile of each protein sequence is generated by running PSI-BLAST program against the NCBI's non-redundant (NR) (ftp://ftp.ncbi.nih.gov/blast/db/nr) database with parameters h and j set to 0.001 and 3, respectively. The (i, j)th entry of the resulting matrix represents the score of the amino acid in the ith position of the query sequence being mutated to amino acid type j during the evolution process.

For convenience, let us denote

$$PSSM = (P_1, P_2, \ldots, P_{20})$$

as the PSSM of the query sequence S, where

$$P_j = (p_{1,j}, p_{2,j}, \ldots, p_{L,j})^{\mathrm{T}} \quad (j = 1, 2, \ldots, 20),$$

L is the length of the query sequence S, and T is the transpose operator.

To make the PSSM descriptor become a uniform representation, we introduce two different approaches, denoted by PSSM-AAC and PSSM-AC, respectively. In the PSSM-AAC model, we represent the query sequence S by

$$AAC(S) = (\overline{P_1}, \overline{P_2}, \ldots, \overline{P_{20}})^{\mathrm{T}},$$

where

$$\overline{P_j} = \frac{1}{L}\sum_{i=1}^{L} p_{i,j} \quad (j = 1, 2, \ldots, 20).$$

$\overline{P_j}$ is the composition of amino acid type j in the PSSM and represents the average score of the amino acid residues in the protein *S* being mutated to amino acid type *j* during the evolution process.

In the PSSM-AC model, the AC transformation is applied to each column of PSSM. As a powerful statistical tool for analyzing sequences of vectors developed by Wold et al. (1993), the AC transformation has been widely applied to the field of bioinformatics (Dong et al. 2009; Guo et al. 2006, 2008; Wu et al. 2010). Here, the AC variable measures the average correlation between two residues separated by a distance of *g* along the sequence *S*, which can be calculated by

$$AC_{j,g}(S) = \frac{1}{L-g}\sum_{i=1}^{L-g}(p_{i,j}-\overline{P_j})\times(p_{i+g,j}-\overline{P_j})$$
$$(j = 1, 2, \ldots, 20).$$

So, the number of AC variables is 20*G, where G is the maximum of g (g = 1, 2,…, G).

Through the above analysis, we know that in the PSSM-AAC model, *AAC*(*S*) can only describe the amino acid composition in the PSSM of the query sequence *S*. So if *AAC*(*S*) was used to represent the protein sample, all the sequence-order information would be lost. On the contrary, in the PSSM-AC model, the sequence-order effect can be partially reflected. In view of this, to incorporate evolutionary information and sequence-order information, we propose a hybrid representation model (AAC-PSSM-AC) by fusing PSSM-AAC and PSSM-AC. As a result, each protein sequence is characterized by a (20 + 20*G)-dimensional vector.

Support vector machine

SVM, introduced by Vapnik (1995), is a machine learning technique based on statistical learning theory. As SVM is a popular algorithm and widely applied in the biological sequence analysis, it is not described in this study. More details about SVM can be found in some machine learning publications (Cortes and Vapnik 1995; Vapnik 1998). In our work, the publicly available LIBSVM software (Chang and Lin 2001) is used to implement the SVM classifier. The software toolbox can be freely downloaded from http://www.csie.ntu.edu.tw/~cjlin/libsvm. Here, radial basis function is chosen as the kernel function. Two parameters, the regularization parameter C and the kernel width parameter γ are optimized based on tenfold cross-validation using a grid search strategy in the LIBSVM software. The final classifier uses $C = 2.0$ and $\gamma = 0.0078$.

Results and discussion

To evaluate the present method comprehensively, we first check the effect of the parameter G on the performance of the AAC-PSSM-AC model, then three sequence representation models proposed in this study are discussed, and finally we compare the AAC-PSSM-AC model with existing methods. All experiments are performed using the jackknife cross-validation test and report the overall accuracy, as well as the accuracy for each structural class.

Effect of the parameter G

Theoretically speaking, the maximum value of parameter G is the length of the shortest sequence in the dataset minus one, which is 30 for the *Z277* dataset and 9 for the *1189* dataset. So the value of G can be 1, 2, …, or 30 for the *Z277* dataset. However, preliminary test results indicated that when $G > 10$, the corresponding accuracy dropped down (data not shown). To simplify the problem, we also focus on the optimal region of G = 1, 2, …, and 9. Here, the overall accuracies for different values of G on the *Z277* and *1189* datasets are shown in Fig. 1. As can be seen from the figure, the optimal value of G for the *Z277* dataset is 6, corresponding to a peak with an overall accuracy of 91.0%. For the *1189* dataset, the accuracy first increases to a maximum value at $G = 3$, then does not vary significantly with the increase of G, and finally achieves the best value of 75.4% at $G = 7$. To make the proposed descriptor become a uniform representation, the value of G is set to 6 in the rest of this study.

Prediction performance of three sequence representation models

In this section, we evaluate the performance of three sequence representation models proposed in this study, i.e., PSSM-AAC, PSSM-AC and AAC-PSSM-AC. All

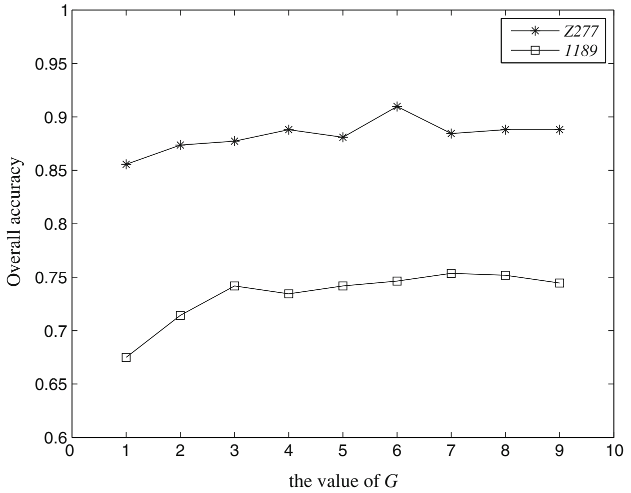

Fig. 1 This graph shows how different values of G affect the overall accuracies of AAC-PSSM-AC model on the *Z277* and *1189* datasets

Table 2 Comparison of PSSM-AAC, PSSM-AC and AAC-PSSM-AC models

Dataset	Model	Prediction accuracy (%)				
		All-α	All-β	α/β	α + β	Overall
Z277	PSSM-AAC	84.3	86.9	90.1	75.4	84.5
	PSSM-AC	91.4	91.8	97.5	78.5	90.3
	AAC-PSSM-AC	88.6	95.1	97.5	81.5	91.0
1189	PSSM-AAC	76.2	80.3	79.9	32.8	68.9
	PSSM-AC	79.8	87.1	79.3	44.4	73.8
	AAC-PSSM-AC	80.7	86.4	81.4	45.2	74.6

experiments were performed on the *Z277* and *1189* datasets and the results are shown in Table 2.

Referring to Table 2, the overall accuracy of the PSSM-AC model is 90.3% on the *Z277* dataset, which is 5.8% higher than that of the PSSM-AAC model and slightly lower than that of the AAC-PSSM-AC model. This well proves the importance of the local sequence-order effects and suggests that the sequence-order effects can be quite effectively reflected by the AC transformation. For the *1189* dataset, AAC-PSSM-AC model also achieves the best overall accuracy of 74.6%, which is 5.7 and 0.8% higher than that of the PSSM-AAC and PSSM-AC models. It is revealed that the incorporation of evolutionary information and sequence-order information do help to improve the prediction of protein structural class.

Performance comparison with existing methods

The proposed AAC-PSSM-AC model is first tested on the *Z277* and *Z498* datasets and compared with other recently reported prediction methods on the same datasets. The results of the jackknife tests are shown in Tables 3 and 4.

Table 3 Comparison of different methods by the jackknife test for the *Z277* dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	α + β	Overall
Neural network (Cai and Zhou 2000)	68.6	85.2	86.4	56.9	74.7
Component coupled (Zhou 1998)	84.3	82.0	81.5	67.7	79.1
SVM (Cai et al. 2001)	74.3	82.0	87.7	72.3	79.4
Rough sets (Cao et al. 2006)	77.1	77.0	93.8	66.2	79.4
Information-theoretical approach (Zheng et al. 2010)	87.1	80.3	93.8	67.7	83.0
LogitBoost (Feng et al. 2005)	81.4	88.5	92.6	72.3	84.1
IGA-SVM (Li et al. 2008)	84.3	88.5	92.6	70.7	84.5
CWT-PCA-SVM (Li et al. 2009)	85.7	90.2	87.7	80.1	85.9
IB1 (Chen et al. 2008)	89.7	88.1	92.2	80.0	87.7
SVM fusion (Chen et al. 2006b)	85.7	90.2	93.8	80.0	87.7
AAC-PSSM-AC	88.6	95.1	97.5	81.5	91.0

Table 4 Comparison of different methods by the jackknife test for the *Z498* dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	α + β	Overall
Neural network (Cai and Zhou 2000)	86.0	96.0	88.2	86.0	89.2
Component coupled (Zhou 1998)	93.5	88.9	90.4	84.5	89.2
Rough sets (Cao et al. 2006)	87.9	91.3	97.1	86.0	90.8
SVM fusion (Chen et al. 2006b)	99.1	96.0	80.9	91.5	91.4
SVM (Cai et al. 2001)	88.8	95.2	96.3	91.5	93.2
Information-theoretical approach (Zheng et al. 2010)	95.3	93.7	97.8	88.3	93.8
IGA-SVM (Li et al. 2008)	96.3	93.6	97.8	89.2	94.2
LogitBoost (Feng et al. 2005)	92.6	96.0	97.1	93.0	94.8
CWT-PCA-SVM (Li et al. 2009)	94.4	96.8	97.0	92.3	95.2
IB1 (Chen et al. 2008)	95.0	95.8	97.8	94.2	95.7
AAC-PSSM-AC	94.4	96.8	97.8	93.8	95.8

Table 5 Performance comparison of different methods on the *1189* dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha + \beta$	Overall
Bayes classifier (Wang and Yuan 2000)	54.8	57.1	75.2	22.2	53.8
Logistic regression (Kurgan and Homaeian 2006)	57.0	62.9	64.7	25.3	53.9
SVM[a] (Anand et al. 2008)	–	–	–	–	54.7
FKNN classifier (Zhang et al. 2008)	48.9	59.5	81.7	26.6	56.9
StackingC ensemble (Kedarisetti et al. 2006)	–	–	–	–	58.9
WSVM (Qiu et al. 2009)	–	–	–	–	59.2
Specific tri-peptides (Costantini and Facchiano 2009)	–	–	–	–	59.9
IB1 (Chen et al. 2008)	65.3	67.7	79.9	40.7	64.7
AAD-CGR (Yang et al. 2009)	62.3	67.7	66.5	63.1	65.2
SVM (Chen et al. 2008)	75.8	75.2	82.6	31.8	67.6
AADP-PSSM (Liu et al. 2010)	69.1	83.7	85.6	35.7	70.7
SCPRED (Kurgan et al. 2008a)	89.1	86.7	89.6	53.8	80.6
RKS-PPSC (Yang et al. 2010)	89.2	86.7	82.6	65.6	81.3
MODAS (Mizianty and Kurgan 2009)	92.3	87.1	87.9	65.4	83.5
AAC-PSSM-AC	80.7	86.4	81.4	45.2	74.6

[a] The result is evaluated using five runs of tenfold cross-validation test

Table 6 Performance comparison of different methods on the *25PDB* dataset

Method	Prediction accuracy (%)				
	All-α	All-β	α/β	$\alpha + \beta$	Overall
Logistic regression (Kurgan and Homaeian 2006)	69.1	61.6	60.1	38.3	57.1
Specific tri-peptides (Costantini and Facchiano 2009)	60.6	60.7	67.9	44.3	58.6
StackingC ensemble (Kedarisetti et al. 2006)	–	–	–	–	59.9
LLSC-PRED (Kurgan and Chen 2007)	75.2	67.5	62.1	44.0	62.2
SVM (Kurgan and Chen 2007)	77.4	66.4	61.3	45.4	62.7
AAD-CGR (Yang et al. 2009)	64.3	65.0	65.0	61.7	64.0
CWT-PCA-SVM (Li et al. 2009)	76.5	67.3	66.8	45.8	64.0
AADP-PSSM (Liu et al. 2010)	83.3	78.1	76.3	54.4	72.9
SCPRED (Kurgan et al. 2008a)	92.6	80.1	74.0	71.0	79.7
MODAS (Mizianty and Kurgan 2009)	92.3	83.7	81.2	68.3	81.4
RKS-PPSC (Yang et al. 2010)	92.8	83.3	85.8	70.1	82.9
SVM (Liu and Jia 2010)	92.6	81.3	81.5	76.0	82.9
AAC-PSSM-AC	85.3	81.7	73.7	55.3	74.1

For the *Z277* dataset, the overall accuracy of the current approach is 91.0%, which is higher than those of the methods listed in Table 3 (from 3.3 to 16.3%). In addition, our method achieves the best performances among the four structural classes except for the all-α class. Meanwhile, our method also performs better than some complex classifiers such as LogitBoost (Feng et al. 2005) and SVM fusion (Chen et al. 2006b). As shown in Table 4, our method also achieves the best performance among these methods on the *Z498* dataset, with the overall accuracy of 95.8%. It is worth noting that IB1 (Chen et al. 2008) algorithm, which also extracts sequence features from the PSI-BLAST profile to represent the query protein, shows a comparable accuracy to the present method. This demonstrates that the PSI-BLAST profile provides a better source of information for the prediction of protein structural class.

As reported by some researchers, protein sequence similarity within the training and testing datasets has a significant effect on the prediction performance of protein structural class, i.e., accuracy will be overestimated when using high-similarity datasets. Thus, to test our method strictly and facilitate the comparison, two low-similarity datasets are also studied separately. The results by jackknife tests are listed in Tables 5 and 6.

Referring to Table 5, for the *1189* dataset, our method gets an overall accuracy of 74.6%, which is higher than those of the other methods (from 3.9 to 20.8%) except for SCPRED, RKS-PPSC and MODAS. The three methods, which additionally use predicted secondary structure information as their input, provide the overall accuracy of over 80%. This demonstrates that the predicted secondary structure provides a better source of information for the prediction of protein structural class. However, our method also obtains favorable prediction accuracy when the predicted secondary structure is not utilized. As shown in Table 6, results on the *25PDB* dataset are consistent with the results on the *1189* dataset. Our method, together with the SCPRED, MODAS, RKS-PPSC and SVM (Liu and Jia 2010) methods, performs better than other methods, with an overall accuracy of 74.1%. SVM (Liu and Jia 2010) method also uses additional input information, such as the predicted secondary structure. In detail, for the $\alpha + \beta$ class, the accuracy of AAD-CGR (Yang et al. 2009) algorithm is a little better than ours. However, for three other classes, our prediction accuracies are superior to the other methods. In summary, our method shows substantial improvements for the two low-similarity datasets. This indicates that our method is very promising and may at least play an important complementary role to existing methods.

Conclusions

In this study, we applied SVM and PSSM to predict protein structural class. By incorporating evolutionary information and sequence order effects, a powerful hybrid representation model (AAC-PSSM-AC) is proposed to convert the PSSMs into fixed-length feature vectors, which are input to SVM to perform the prediction. Jackknife cross-validation tests are performed on four working datasets to evaluate the performance of our method. According to the experimental results, our method substantially outperforms most of the existing methods and may provide a cost-alternative to predict protein structural class.

Acknowledgments This work was partially supported by the National Natural Science Foundation of China (No. 10731040), Shanghai Leading Academic Discipline Project (No. S30405), Innovation Program of Shanghai Municipal Education Commission (No. 09zz134), and the Science and Technology Project of Provincial Education Department of Shandong of China (No. J08LI09).

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402

Anand A, Pugalenthi G, Suganthan PN (2008) Predicting protein structural class by SVM with class-wise optimized features and decision probabilities. J Theor Biol 253(2):375–380

Cai YD, Feng KY, Lu WC, Chou KC (2006) Using LogitBoost classifier to predict protein structural classes. J Theor Biol 238(1):172–176

Cai YD, Liu XJ, Xu X, Zhou GP (2001) Support vector machines for predicting protein structural class. BMC Bioinformatics 2:3

Cai YD, Zhou GP (2000) Prediction of protein structural classes by neural network. Biochimie 82(8):783–785

Cao YF, Liu S, Zhang LD, Qin J, Wang J, Tang KX (2006) Prediction of protein structural class with Rough Sets. BMC Bioinformatics 7:20

Chang CC, Lin CJ (2001) LIBSVM: a library for support vector machines

Chen C, Tian YX, Zou XY, Cai PX, Mo JY (2006a) Using pseudo-amino acid composition and support vector machine to predict protein structural class. J Theor Biol 243(3):444–448

Chen C, Zhou X, Tian Y, Zou X, Cai P (2006b) Predicting protein structural class with pseudo-amino acid composition and support vector machine fusion network. Anal Biochem 357(1):116–121

Chen K, Kurgan LA, Ruan JS (2008) Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J Comput Chem 29(10):1596–1604

Chen L, Lu L, Feng K, Li W, Song J, Zheng L, Yuan Y, Zeng Z, Lu W, Cai Y (2009) Multiple classifier integration for the prediction of protein structural classes. J Comput Chem 30(14):2248–2254

Chou KC (1999) A key driving force in determination of protein structural classes. Biochem Biophys Res Commun 264(1):216–224

Chou KC (2001) Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins 43(3):246–255

Chou KC, Cai YD (2004) Predicting protein structural class by functional domain composition. Biochem Biophys Res Commun 321(4):1007–1009

Cortes C, Vapnik V (1995) Support-vector networks. Mach Learn 20(3):273–297

Costantini S, Facchiano AM (2009) Prediction of the protein structural class by specific peptide frequencies. Biochimie 91(2):226–229

Dong QW, Zhou SG, Guan JH (2009) A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation. Bioinformatics 25(20):2655–2662

Feng KY, Cai YD, Chou KC (2005) Boosting classifier for predicting protein domain structural class. Biochem Biophys Res Commun 334(1):213–217

Guo Y, Li M, Lu M, Wen Z, Huang Z (2006) Predicting G-protein coupled receptors-G-protein coupling specificity based on autocross-covariance transform. Proteins 65(1):55–60

Guo YZ, Yu LZ, Wen ZN, Li ML (2008) Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. Nucleic Acids Res 36(9):3025–3030

Kedarisetti KD, Kurgan L, Dick S (2006) Classifier ensembles for protein structural class prediction with varying homology. Biochem Biophys Res Commun 348(3):981–988

Kurgan L, Chen K (2007) Prediction of protein structural class for the twilight zone sequences. Biochem Biophys Res Commun 357(2):453–460

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

Kurgan LA, 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 Recogn 39(12):2323–2343

Kurgan LA, Zhang T, Zhang H, Shen SY, Ruan JS (2008b) Secondary structure-based assignment of the protein structural classes. Amino Acids 35(3):551–564

Levitt M, Chothia C (1976) Structural Patterns in Globular Proteins. Nature 261(5561):552–558

Li ZC, Zhou XB, Dai Z, Zou XY (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(2):415–425

Li ZC, Zhou XB, Lin YR, Zou XY (2008) Prediction of protein structure class by coupling improved genetic algorithm and support vector machine. Amino Acids 35(3):581–590

Liu T, Jia C (2010) A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J Theor Biol 267(3):272–275

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(10):1330–1334

Luo RY, Feng ZP, Liu JK (2002) Prediction of protein structural class by amino acid and polypeptide composition. Eur J Biochem 269(17):4219–4225

Mizianty MJ, Kurgan L (2009) Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinformatics 10:414

Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536–540

Nakashima H, Nishikawa K, Ooi T (1986) The folding type of a protein is relevant to the amino acid composition. J Biochem 99(1):153–162

Qiu JD, Luo SH, Huang JH, Liang RP (2009) Using support vector machines for prediction of protein structural classes based on discrete wavelet transform. J Comput Chem 30(8):1344–1350

Shen HB, Yang J, Liu XJ, Chou KC (2005) Using supervised fuzzy clustering to predict protein structural classes. Biochem Biophys Res Commun 334(2):577–581

Sun XD, Huang RB (2006) Prediction of protein structural classes using support vector machines. Amino Acids 30(4):469–475

Vapnik V (1995) The nature of statistical learning theory. Springer, New York

Vapnik V (1998) Statistical learning theory. Wiley, New York

Wang ZX, Yuan Z (2000) How good is prediction of protein structural class by the component-coupled method? Proteins 38(2):165–175

Wold S, Jonsson J, Sjostrom M, Sandberg M, Rannar S (1993) DNA and peptide sequences and chemical processes multivariately modeled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277(2):239–253

Wu J, Li M, Yu L, Wang C (2010) An ensemble classifier of support vector machines used to predict protein structural classes by fusing auto covariance and pseudo-amino acid composition. Protein J 29(1):62–67

Xiao X, Lin WZ, Chou KC (2008) Using grey dynamic modeling and pseudo amino acid composition to predict protein structural classes. J Comput Chem 29(12):2018–2024

Yang JY, Peng ZL, Chen X (2010) Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinformatics 11 Suppl 1:S9

Yang JY, Peng ZL, Yu ZG, Zhang RJ, Anh V, Wang DS (2009) Prediction of protein structural classes by recurrence quantification analysis based on chaos game representation. J Theor Biol 257(4):618–626

Zhang TL, Ding YS (2007) Using pseudo amino acid composition and binary-tree support vector machines to predict protein structural classes. Amino Acids 33(4):623–629

Zhang TL, Ding YS, Chou KC (2008) Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern. J Theor Biol 250(1):186–193

Zheng X, Li C, Wang J (2010) An information-theoretic approach to the prediction of protein structural class. J Comput Chem 31(6):1201–1206

Zhou GP (1998) An intriguing controversy over protein structural class prediction. J Protein Chem 17(8):729–738