Gene 554 (2015) 241–248

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Methods Paper

Prediction of protein structural classes for low-similarity sequences using reduced PSSM and position-based secondary structural features

Junru Wang [b,1], Cong Wang [a,1], Jiajia Cao [a], Xiaoqing Liu [c], Yuhua Yao [a], Qi Dai [a,d,*]

[a] College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[b] College of Mechanical Engineering and Automation, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China
[c] College of Sciences, Hangzhou Dianzi University, Hangzhou 310018, People's Republic of China
[d] Department of Molecular and Cell Biology, University of Texas at Dallas, Richardson, TX 75080, USA

ARTICLE INFO

Article history:
Received 13 September 2014
Received in revised form 19 October 2014
Accepted 22 October 2014
Available online 24 October 2014

Keywords:
Reduced PSSM
Global correlation
Secondary structural feature
Protein structural class prediction

ABSTRACT

Many efficient methods have been proposed to advance protein structural class prediction, but there are still some challenges where additional insight or technology is needed for low-similarity sequences. In this work, we schemed out a new prediction method for low-similarity datasets using reduced PSSM and position-based secondary structural features. We evaluated the proposed method with four experiments and compared it with the available competing prediction methods. The results indicate that the proposed method achieved the best performance among the evaluated methods, with overall accuracy 3–5% higher than the existing best-performing method. This paper also found that the reduced alphabets with size 13 simplify PSSM structures efficiently while reserving its maximal information. This understanding can be used to design more powerful prediction methods for protein structural class.

1. Introduction

Protein structural classes play an important role in protein science, such as protein function prediction, protein folding rate analysis, prediction of DNA binding sites, protein fold recognition, membrane protein analysis, reduction of the conformation search space and implementation of a heuristic approach to find tertiary structure (Klein and Delisi, 1986; Chou, 2006; Levitt and Chothia, 1976; Andreeva et al., 2004; Murzin et al., 1995; Ferragina et al., 2007; Dai and Wang, 2008). With the development of sequencing technologies, the gap between sequence-known and structure-known proteins has become larger in recent years. Consequently, the burden of experimental methods to find the 3-dimensional structures would become even more unbearable. Therefore, it is necessary to develop computational methods for fast and accurate determination of protein structural classes.

Since protein structural class concept was proposed by Levitt and Chothia (Levitt and Chothia, 1976; Andreeva et al., 2004; Murzin et al., 1995), various significant efforts have been made to predict protein structural class during the past 30 years (Dai and Wang, 2008; Chen et al., 2006; Chou, 2000; Kedarisetti et al., 2006; Dai et al., 2011). Previous studies indicated that protein structural classes could be predicted from amino acid sequences (Klein and Delisi, 1986; Chou, 1999; Chou and Shen, 2007), consequently, several features of protein sequences have been proposed for protein structural class prediction, such as short polypeptide composition (Luo et al., 2002; Sun and Huang, 2006; Zhang et al., 2014), pseudo AA composition (Ding et al., 2007; Wu et al., 2011; Liao et al., 2012; Kong et al., 2014) and collocation of function domain composition (Chou and Cai, 2004).

Evolutionary profile is another widely used feature in protein structural class predication. Given a query sequence, it can be searched against a database of proteins using position-specific iterated BLAST (PSI-BLAST) (Altschul et al., 1997), from which a position specific scoring matrix (PSSM) is extracted to represent evolutionary information of protein (Stormo et al., 1982). Jones used PSI-BLAST to search a large non-redundant protein sequence dataset to obtain the position specific scoring matrix (log-odds values) and further input it to neural network for prediction (Jones, 1999). Chou and Shen proposed the pseudo-position specific scoring matrix (PsePSSM) and developed a new web-server for predicting protein subnuclear localization, Nuc-PLoc (Chou and Shen, 2007; Shen and Chou, 2007). Chen et al. extracted evolutionary information using PSI-BLAST profile-based collocation of AA pairs, and achieved 61–96% accuracy on the six datasets using

Abbreviations: AA, amino acids; AAC-PSSM, amino acid composition and position specific scoring matrix; AvgSeg, average length of segment; BP, back-propagation; DPC-PSSM, dipeptide composition and position specific scoring matrix; GP, *g* position; K-NN, k-nearest neighbor algorithm; PSI-BLAST, position-specific Iterated BLAST; PSSM, position specific scoring matrix; RedPSSM, reduced position specific scoring matrix; MaxSeg, length of longest segment; MCC, Matthew's correlation coefficient; NAvgSeg, normalized average length of segment; NCount, normalized count of segments; NMaxSeg, normalized length of longest segment; NRG, number of reduced groups; Sens, sensitivity; Spec, specificity; SVM, support vector machine.

* Corresponding author at: College of Life Sciences, Zhejiang Sci-Tech University, Hangzhou 310018, People's Republic of China.
E-mail address: daiailiu04@yahoo.com (Q. Dai).

[1] Junru Wang and Cong Wang contributed equally to this work as co-first authors.

http://dx.doi.org/10.1016/j.gene.2014.10.037
0378-1119/

support vector machine (Chen et al., 2008). Liu et al. calculated amino acid composition and dipeptide composition from PSI-BLAST profiles, in which the average scores of the amino acid residues in the protein being mutated to another amino acid type were calculated as AAC-PSSM features, and the traditional DPC of PSSM was explored as DPC-PSSM (Liu et al., 2010). Recently, Liu et al. measured the average correlation between two residues separated by a g distance in a column of position specific scoring matrix and applied it to predict protein structural class (Liu et al., 2012). Ding et al. extracted the long-range correlation information and linear correlation information from the PSSM (Ding et al., 2014).

Although promising results have been achieved using above methods, but prediction accuracy is limited especially for low-similarity datasets (Kedarisetti et al., 2006; Kurgan and Homaeian, 2006). Recently, several features associated with predicted secondary structures have been proposed (Chou and Cai, 2004; Altschul et al., 1997; Stormo et al., 1982; Jones, 1999; Shen and Chou, 2007; Chen et al., 2008; Liu et al., 2010; Liu et al., 2012; Ding et al., 2014; Kurgan and Homaeian, 2006; Kurgan et al., 2008; Zheng and Kurgan, 2008; Mizianty and Kurgan, 2009; Liu and Jia, 2010; Zhang et al., 2011; Hobohm and Sander, 1994; Zhang et al., 2013). First of all, the popular ones are widely used content of predicted secondary structural elements ($content_{SE}$), normalized count of segments (NCount), length of the longest segment (MaxSeg), normalized length of the longest segment (NMaxSeg), average length of the segment (AvgSeg), and normalized average length of the segment (NAvgSeg) (Kurgan et al., 2008). Zheng and Kurgan studied 3PATTERN of the predicted secondary structures and used them to predict protein β-turns (Zheng and Kurgan, 2008). MODAS exploring both secondary structural information and evolutionary profiles is also a widely used predication method (Mizianty and Kurgan, 2009). Recently, Liu and Jia (2010), Zhang et al. (2011) and (Zhang et al. (2013) studied the distribution of helices and strands among four structural classes. For example, Zhang et al. calculated transition probabilities of helices and strands to numerically characterize their alterations along secondary structure sequences.

With the help of the above features, prediction accuracy was improved over 80% for several low-similarity benchmark datasets, but several critical problems still exist in their development. First, some PSSM-based methods focus on composition and average correlation between two residues in a column of the PSSM, and therefore to sometimes are unaware of their global structural correlation among the different columns. Second, the available structural features are associated with the structural elements' contents and combinations, but their position distributions along proteins are rarely used.

With the above problems in mind, we presented a scheme to predict the protein structural classes using the reduced PSSM (RedPSSM) structural properties and position-based secondary structural features. We first explored a potential way to simplify PSSM structure while reserving its maximal information. With the help of auto covariance transformation, we studied global structural properties of the RedPSSM and discussed the influence of its parameters. Based on our previous study (Dai et al., 2013), we combined the position-based structural features with the RedPSSM to predict protein structural classes using a multi-class support vector machine (SVM) (Vapnik, 2000). Through a comprehensive comparison and discussion, some novel valuable guidelines for the use of the RedPSSM structural properties and position-based secondary structural features were obtained.

The remainder of this paper was organized as follows. Section 2 presented used benchmark datasets, extraction features of reduced PSSM (RedPSSM) structural properties and position-based secondary structural features, and prediction method. Section 3 summarized the key results of the proposed method, performance comparison with the competing predictions and discussion of parameters in RedPSSM structural properties.

2. Materials and methods

2.1. Datasets

This paper selected four widely used low similarity benchmark datasets that facilitate the comparison with the available methods (Kurgan and Homaeian, 2006; Kurgan et al., 2008; Zheng and Kurgan, 2008; Mizianty and Kurgan, 2009; Liu and Jia, 2010; Zhang et al., 2011; Zhang et al., 2013; Dai et al., 2013; Vapnik, 2000; Kurgan and Chen, 2007). The first dataset is 25PDB with 25% sequence identity originally published in (Kurgan and Homaeian, 2006). It contains 1673 proteins and domains downloaded from PDB and scanned with high resolution. The secondary dataset is 640 with 25% sequence identity. It consists of 640 proteins with classification labels retrieved from SCOP database (Kurgan and Homaeian, 2006). The third dataset is FC699, in which there are 858 sequences sharing low 40% identity. The last dataset is referred to as 1189 with 40% sequence identity. It consists of 1092 3-D structural data of proteins downloaded from RCSB protein data bank with PDB IDs listed in this paper (Chen et al., 2008). Table 1 provides more detailed information on these low similarity benchmark datasets.

2.2. RedPSSM structural properties

2.2.1. Position specific scoring matrix

Position specific scoring matrix is a commonly-used evolutionary profile in protein study. For a protein sequence s with length L, its PSSM can be represented as the following equation

$$PSSM_s = \begin{bmatrix} P_{1\to1} & P_{1\to2} & \cdots & P_{1\to j} & \cdots & P_{1\to20} \\ P_{2\to1} & P_{2\to2} & \cdots & P_{2\to j} & \cdots & P_{2\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{i\to1} & P_{i\to2} & \cdots & P_{i\to j} & \cdots & P_{i\to20} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ P_{L\to1} & P_{L\to2} & \cdots & P_{L\to j} & \cdots & P_{L\to20} \end{bmatrix} \quad (1)$$

where $i \to j$ describes i-th amino acid residue of the protein sequence s being mutated to amino acid type j in the biology evolution process, $P_{i\to j}$ is the score of this mutation and L is the length of the sequence s, and the numerical codes 1, 2, 3 … 20 denote the single character of ordered 20 native amino acid types in the above equation.

High PSSM-scores of some proteins often indicate some possible biological evolutionary relationships between these sequences and their family (Jones, 1999; Shen and Chou, 2007; Chen et al., 2008; Liu et al., 2010; Liu et al., 2012; Ding et al., 2014). Here, we used PSI-BLAST program ($h = 0.001$ and $j = 3$) to search and align homogeneous sequences from SWISS-PROT database (published: May-20-2013 with 540,052 sequences). PSI BLAST will return a PSSM with 20-dimensional vector whose values are conserved mutation scores for 20 amino acids. And the PSSM elements were then scaled to the range from 0 to 1 with the help of the following sigmoid function:

$$f(x) = \frac{1}{1+e^{-x}}. \quad (2)$$

2.2.2. Reduced amino acids

Position specific scoring matrix reflects conserved mutation scores among 20 amino acids. It is well known that 20 amino acids are subtly

Table 1
The number of proteins belonging to different structural classes in the datasets.

Dataset	All-α	All-β	α/β	$\alpha+\beta$	Total
25PDB	443	443	346	441	1673
640	138	154	177	171	640
FC699	130	269	377	82	858
1189	223	294	334	241	1092

different from each other, but some of them have similar basic structures and functions. So the structure of PSSM can be simplified by clustering 20 amino acids.

Here, we used BLOSUM62 to cluster amino acids based on reasonable simplification of protein sequences (Li et al., 2003). Let X_i represents the ith group, we defined the jth amino acid in the ith group as $X_i(j)$. With the help of BLOSUM62, we defined similarity score of $X_i(j)$ in the reduced sequence Seq_s to the kth amino acid R_k in the ith group in the parent sequence Seq_0 is

$$S(X_i(j), R_k) = Blosum(X_i(j), R_k), \tag{3}$$

where $Blosum(X_i(j), R_k)$ is the substitution value for exchange of residue $X_i(j)$ with residue R_k. Then, the total similarity score of the simplified sequence Seq_s to the Seq_0 can be calculated as the sum of the scores over all groups:

$$S = \sum_{i=1}^{N} \left[\sum_{j=1}^{g_s(i)} \sum_{k=1}^{g_0(i)} m_i(k) S(X_i(j), R_k) \right] / g_s(i), \tag{4}$$

where $g_0(i)$ and $g_s(i)$ denote the size of the ith group in Seq_0 and Seq_s, $m_i(k)$ is the number of the residue R_k in Seq_0, and N is the number of groups. Clearly, S is a measure of the reservation of the information on the parent sequence. Given a group size N, we went through all amino acids' groups and calculated their similarity scores between the parent sequence Seq_0 and reduced sequence Seq_s. We finally chose the reduced alphabets with the highest similarity score. All kinds of reduced alphabets for 20 amino acids were listed in Table 2. In this paper, we selected 13 reduced amino acids to simplify the structure of PSSM, and more discussions can be found in Section 3.3.

2.2.3. Structural properties of RedPSSM

With the help of the reduced amino acids, we transformed a position specific scoring matrix into a simple one (RedPSSM) and extracted its structural features. For a given protein sequence s with length L, we obtained its PSSM with the help of the PSI BLAST. We defined it as a function M: $[0, L-1] \times \sum \rightarrow \Re$, where L is the length of M, and $\sum$ is a finite alphabet. Usually, PSSM is represented by a $L \times |\sum|$ matrix. With the help of the reduced amino acids listed in Table 2, the PSSM of the parent sequence Seq_0 could be transformed into the RedPSSM of the reduced sequence Seq_s,

$$[0, L] \times \sum\nolimits_{Seq_0} \rightarrow [0, L] \times \sum\nolimits_{Seq_r}, \tag{5}$$

Table 2
Clustering of amino acids by the maximal score for different levels of reduction.

Group																				
2					CMFILVWY										AGTSNQDEHRKP					
3	CMFILVWY								AGTSP						NQDEHRK					
4	CMFWY				ILV				AGTS						NQDEHRKP					
5	WFYH				MILV				CATSP					G	NQDERK					
6	WFYH				MILV				CATS				P	G	NQDERK					
7	WFYH				MILV				CATS				P	G	NQDE				RK	
8		WFYH				MILV			CA			NTS		P	G	DE			QRK	
9		WFYH			MI		LV		CA			NTS		P	G	DE			QRK	
10		WFY		ML		IV		CA		TS		NH		P	G	DE			ERK	
11		WFY		ML		IV		CA		TS		NH		P	G	D		QE		RK
12		WFY		ML		IV		C	A	TS		NH		P	G	D		QE		RK
13		WFY		ML		IV		C	A	T	S	NH		P	G	D		QE		RK
14		WFY		ML		IV		C	A	T	S	NH		P	G	D		QE	R	K
15		WFY		ML		IV		C	A	T	S	N	H	P	G	D		QE	R	K
16	W		FY		ML		IV		C	A	T	S	N	H	P	G	D		QE	R	K
17	W		FY		ML		IV		C	A	T	S	N	H	P	G	D	Q	E	R	K
18	W		FY	M	L		IV		C	A	T	S	N	H	P	G	D	Q	E	R	K
19	W	F	Y	M	L		IV		C	A	T	S	N	H	P	G	D	Q	E	R	K
20	W	F	Y	M	L	I	V		C	A	T	S	N	H	P	G	D	Q	E	R	K

where $\sum_{Seq_0}$ and $\sum_{Seq_r}$ are finite alphabets of 20 amino acids and reduced amino acids. According to the PSSM and reduced amino acids, $[\text{RedPSSM}]_{ij}$ could be calculated as follows

$$[\text{RedPSSM}]_{ij} = \sum_{j=1}^{g_s(j)} P_{i \rightarrow j} / g_s(j), \tag{6}$$

where $g_s(j)$ denotes the size of the reduced group consisting the reduced amino acid j, and $1 \le j \le |\sum_{seq_r}|$.

Auto covariance transformation as a powerful statistical tool has been used in the prediction of protein structural classes (Liu et al., 2012). Liu et al. measured average correlation between two residues separated by a distance of g along the sequence S, calculated by

$$AC_{j,g}(S) = \frac{1}{L-g} \sum_{i=1}^{L-g} (P_{i,j} - \overline{P}_j)\left(P_{i,j+g} - \overline{P}_j\right). \tag{7}$$

$AC_{j,g}(S)$ describes reaction information between two residues separated by a distance of g in the same column. But it still focuses mostly on the local information in the same column, and therefore to sometimes is unaware of the useful global discriminatory information embedded in different columns. With this problem in mind, we measured the global correlation between two different columns by,

$$\begin{aligned} RAC_g(h, j) &= \frac{1}{L-g} \sum_{i=1}^{L-g} \left| [\text{RedPSSM}]_{i,h} - \frac{[\text{RedPSSM}]_{i,h} + [\text{RedPSSM}]_{i+g,j}}{2} \right| \\ &\quad \times \left| [\text{RedPSSM}]_{i+g,h} - \frac{[\text{RedPSSM}]_{i,h} + [\text{RedPSSM}]_{i+g,j}}{2} \right| \\ &= \frac{1}{4(L-g)} \sum_{i=1}^{L-g} \left([\text{RedPSSM}]_{i,h} - [\text{RedPSSM}]_{i+g,h} \right)^2. \end{aligned} \tag{8}$$

From the above equation, it is obvious that RAC_g contains both local correlation information between two residues in the same column and the global correlations among different columns. This paper calculated 169 RAC_g to describe RedPSSM structural properties instead of widely used 400 $AC_{j,g}(S)$, which allows prediction method to run with less memory and is faster.

2.3. Position-based secondary structure features

Using protein evolutionary profile achieves promising results in the prediction of protein structural classes, but its accuracy is limited. The research indicates that the contents and spatial arrangements of secondary structural elements are also significant factors that influence the proteins' intricate functions or structures. Consequently, various secondary structural features have been proposed to improve protein structural class prediction (Kurgan et al., 2008; Zheng and Kurgan, 2008; Mizianty and Kurgan, 2009; Liu and Jia, 2010; Zhang et al., 2011; Hobohm and Sander, 1994; Zhang et al., 2013). But we should note that these features mainly described the content distribution of secondary structural elements, and therefore may ignore their position distribution among protein secondary structures. For example, as for the secondary structure CCEEEEECCCCCHHHHHHH, we can obtain another secondary structure CCHHHHHHHEEEECCCCC when moving its last seven HHHHHHH to the third position. Take a closer look at the above two secondary structures, we found that their elements' content did not change. Therefore, the secondary structure elements' position should be considered as another deciding factor when assigning the protein structural classes (Dai et al., 2013).

We first transformed a protein secondary structure into three position sequences according to the appearance of the given secondary structural element δ. As for these position sequences, it is easy to observe that if the interval distance $Dis(\delta)$ is equal to 1, they are from

the same structure domain, otherwise they belong to two different ones. So we further analyzed the distribution of interval distances between two continuing structural elements.

Given $Dis(\delta)$ and a positive integer n, $p(Dis(\delta)=n)$ is the probability that $Dis(\delta)$ takes the value n. When collecting all pairs $(Dis(\delta)=n, P(Dis(\delta)=n))$, we obtained the probability distribution of the $Dis(\delta)$ of the given secondary structure. We then calculated numerical characteristics semi-mean $Semi-E_{(k)}(\delta)$ and semi-variance $Semi-D_{(k)}(\delta)$ defined by:

$$Semi{-}E_{(k)}(\delta)=\sum_{Dis(\delta)=1}^{k} Dis(\delta)\times P(Dis(\delta)), \tag{9}$$

$$Semi{-}D_{(k)}(\delta)=\sum_{Dis(\delta)=1}^{k}(Dis(\delta))^2\times P(Dis(\delta))-\left[\sum_{Dis(\delta)=1}^{k} Dis(\delta)\times P(Dis(\delta))\right]^2. \tag{10}$$

We have analyzed the influence of parameter k, and found that $k=5$ shows similar performance to $k=\max(Dis(\delta))$ (Dai et al., 2013). Therefore, this paper defined the position-based secondary structural feature as the ratio of the standard $Semi-D_{(5)}$ to $Semi-E_{(5)}$

$$PSSF(\delta)=\frac{Semi{-}E_{(5)}(\delta)}{\sqrt{Semi{-}D_{(5)}(\delta)}}. \tag{11}$$

$PSSF(\delta)$ reflects variability of the position distribution for the element δ in relation to the mean of its population. Here, we calculated $PSSF(C)$, $PSSF(E)$ and $PSSF(H)$ as position-based secondary structure features to predict protein structural classes.

2.4. Classification algorithm

Support vector machine (SVM) is a well-known large margin classifier based on statistical learning theory, in which an optimal separating hyper-plane is found to separate two classes. As for the binary SVM, its decision function is

$$f(x)=\sum_{i=1}^{N}\alpha_i y_i K(x_i,x)+b, \tag{12}$$

where b is a constant, C is a cost parameter controlling the trade-off between allowing training errors and forcing rigid margins, $y_i\in\{-1,+1\}$, x_i is the support vector, $0\le\alpha_i\le C$, and $K(x_i, x)$ is the kernel function. This paper adopted Vapnik's support vector machine to predict protein structural classes. Because there are more than two structural classes, we chose the multi-class SVM using"one-against-others"strategy. Given a test protein with unknown class, we calculated the proposed feature vector and mapped it into the feature space. SVM will then find an optimized linear division to solve this multi-class problem. Lastly, SVM will assign a predicted label to the test protein. More detailed information on prediction scheme can be found in Vapnik's book (Vapnik, 2000).

Here, we selected Gaussian kernel function of the SVM because of its superiority for solving nonlinear problem (Cai et al., 2002; Yuan et al., 2005). A simple grid search strategy was further used to select parameters C and gamma with the highest overall prediction. It was designed based on 10-fold cross-validation for each dataset, and the values of C and gamma were taken from 2^{-10} to 2^{10}.

2.5. Performance evaluation

Sub-sampling test, independent dataset test and jackknife test are three widely used cross-validation methods to evaluate classifier's capability. The jackknife test always yields a unique outcome, which facilitates examining the quality of various predictors. Hence, we chose jackknife test to evaluate the performance of the proposed method and introduced sensitivity (Sens), specificity (Spec) and Matthew's correlation coefficient (MCC) as standard performance measures as well as the accuracy for each class and overall accuracy. These standard performance measures are defined as follows:

$$Accuracy_j=\frac{TP_j}{|C_j|}, \tag{13}$$

$$Overall\,accuracy=\frac{\sum_j TP_j}{\sum_j |C_j|}, \tag{14}$$

$$Sens=\frac{TP}{TP+FN}, \tag{15}$$

$$Spec=\frac{TN}{FP+TN}, \tag{16}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{17}$$

where TP_j is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FN is the number of false negatives, and $|C_j|$ is the number of proteins in each structural class C_j (all-α, all-β, α/β and $\alpha+\beta$ classes).

3. Results and discussion

3.1. Performance of proposed prediction method

Four benchmark datasets 25PDB, 640, FC699 and 1189 were used with low sequence identity 25%, 25%, 40% and 40%, respectively. A simple grid search strategy over C and gamma values was employed based on 10-fold cross-validation for each dataset, where C and gamma were allowed to take the values only between 2^{-10} and 2^{10}. Table 3 summarizes sensitivity (Sens), specificity (Spec) and Matthew's correlation coefficient (MCC) of the proposed method, and the accuracy for class C_j and overall accuracy are shown in Table 4.

Table 3 shows that all-α class prediction achieves the best performance among four structural classes, its sensitivities, specificities and Matthew's correlation coefficient are higher than 93%. But lower prediction is associated with $\alpha+\beta$ class. From Table 4, we found that the overall accuracies of the proposed method are above 87% for four datasets. The overall accuracy of all-α class is significantly higher than other classes with accuracies over 93%, which is followed by that of

Table 3
Sensitivity (Sens), specificity (Spec) and Matthew's correlation coefficient (MCC) of proposed method on four datasets.

Dataset	Class	Sens (%)	Spec (%)	MCC (%)
25PDB	All-α	98.87	99.11	97.56
	All-β	89.62	96.75	86.74
	α/β	85.55	96.08	81.46
	$\alpha+\beta$	78.91	92.61	71.63
640	All-α	93.68	98.80	94.93
	All-β	83.77	97.74	84.28
	α/β	91.53	93.95	83.69
	$\alpha+\beta$	79.53	92.54	72.07
FC699	All-α	97.69	100.0	98.64
	All-β	97.40	98.98	96.48
	α/β	97.08	96.88	93.86
	$\alpha+\beta$	79.27	97.81	77.08
1189	All-α	98.65	98.5	95.60
	All-β	89.12	98.50	89.64
	α/β	89.22	92.48	80.37
	$\alpha+\beta$	73.44	93.77	68.36

Table 4
Prediction accuracies (variances in the brackets) of the proposed method for four datasets and comparison with other reported results.

Dataset	Method	Prediction accuracy (%)				
		All-α	All-β	α/β	α + β	Overall
25PDB	AADP-PSSM (Liu et al., 2010)	69.1	83.7	85.6	35.7	70.7
	AAC-PSSM-AC (Liu et al., 2012)	85.3	81.7	73.7	55.3	74.1
	SCPRED (Kurgan et al., 2008)	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
	Zhang et al. 2011 (Liu and Jia, 2010)	95.0	85.6	81.5	73.2	83.9
	RKS-PPSC (Yang et al., 2010)	92.8	83.3	**85.8**	70.1	82.9
	Ding et al. (2012)	95.0	81.3	83.2	77.6	84.3
	Xia et al. (2012)	92.6	72.5	71.7	71.0	77.2
	Zhang et al. (2013)	95.7	80.8	82.4	75.5	83.7
	Ding et al. (2014)	91.7	80.8	79.8	64.0	79.0
	Zhang et al. (2014)	94.4	83.3	83.5	73.2	83.6
	This paper	**98.9**	**89.6**	85.6	**78.9**	**88.4 (0.08)**
640	SCEC (Chen et al., 2008)	73.9	61.0	81.9	33.9	62.3
	SCPRED (Kurgan et al., 2008)	90.6	81.8	85.9	66.7	80.8
	RKS-PPSC (Yang et al., 2010)	89.1	85.1	88.1	71.4	83.1
	Ding et al. (2014)	92.8	88.3	85.9	66.1	82.7
	Zhang et al. (2014)	92.0	81.8	87.6	74.3	83.6
	Kong et al. (2014)	94.2	80.5	87.6	77.2	84.5
	This paper	93.7	**83.8**	**91.5**	**79.5**	**87.5 (0.07)**
FC699	SCPRED (Kurgan et al., 2008)	–	–	–	–	87.5
	11 features (Liu and Jia, 2010)	**97.7**	88.0	89.1	**84.2**	89.6
	Kong et al. (2014)	96.2	90.7	96.3	69.5	92.0
	This paper	**97.7**	**97.4**	**97.1**	79.3	**95.6 (0.04)**
1189	AADP-PSSM (Liu et al., 2010)	69.1	83.7	85.6	35.7	70.7
	AAC-PSSM-AC (Liu et al., 2012)	80.7	86.4	81.4	45.2	74.6
	SCPRED (Kurgan et al., 2008)	89.1	86.7	**89.6**	53.8	80.6
	MODAS (Mizianty and Kurgan, 2009)	92.3	87.1	87.9	65.4	83.5
	RKS-PPSC (Yang et al., 2010)	89.2	86.7	82.6	65.6	81.3
	Zhang et al. (2013)	92.4	84.4	84.4	73.4	83.6
	Ding et al. (2014)	89.2	88.8	85.6	58.5	81.2
	Zhang et al. (2014)	91.5	86.7	82.0	66.4	81.8
	Kong et al. (2014)	91.9	84.4	85.3	72.2	83.5
	This paper	**98.7**	**89.1**	89.2	**73.4**	**87.6 (0.07)**

Bold values are the best prediction results in each experiments.

all-β and α/β classes. When going through all the results, it is not difficult to note that the average overall accuracy of $\alpha + \beta$ class for four datasets is 77.8%, which is 19.5% lower than that of all-α class. Both T 3 and 4 indicate that it is more difficult to predict $\alpha + \beta$ class because there is non-negligible overlap in this class.

3.2. Performance comparison with the competing predictions

This paper further compared the proposed method with the available competing methods. Here, the accuracy of each class and overall accuracy were chosen as evaluation indexes to evaluate all the prediction methods, and their results were summarized in Table 4.

The proposed method was first compared with AADP-PSSM (Liu et al., 2010), AAC-PSSM-AC (Liu et al., 2012) and Ding's method (Ding et al., 2014) based on the position specific scoring matrix. Among all the experiments, the proposed method achieved the best performance, with accuracy above 5.4–9.4% better than the next competing Ding's method (Ding et al., 2014).

We further compared the proposed method with recently published methods. In the 25PDB experiment, the competing methods consist of SCPRED (Kurgan et al., 2008), MODAS (Mizianty and Kurgan, 2009), Zhang et al. (2011), RKS-PPSC (Yang et al., 2010), Ding et al. (2012), Xia et al. (2012), Zhang et al. (2013) and Zhang et al. (2014). Table 4 shows that the proposed method achieved the best performance with an overall accuracy of 88.4%, which is 4.4% higher than Ding's method (Ding et al., 2012). In the 640 dataset, we compared the proposed method with SCEC (Chen et al., 2008), SCPRED (Kurgan et al., 2008), RKS-PPSC (Yang et al., 2010), Zhang et al. (2014) and Kong et al. (2014). Our method achieved an overall accuracy of 87.5%, which is 3–4% higher than the next-competing methods (Zhang et al., 2014; Kong et al., 2014). As for FC699 experiment, the competing methods include SCPRED (Kurgan et al., 2008), 11 features (Liu and Jia, 2010) and Kong et al. (2014). The comparison indicates that the proposed method significantly outperforms the other methods with an overall accuracy of 95.6%. In the 1189 dataset, the compared methods are SCPRED (Kurgan et al., 2008), MODAS (Mizianty and Kurgan, 2009), RKS-PPSC (Yang et al., 2010), Zhang et al. (2013), Zhang et al. (2014) and Kong et al. (2014). From Table 4, we also found that the proposed method achieved the best performance among all the competing methods. It is the only prediction method with an overall accuracy over 87%, with 4.1% higher than the next-competing methods (Kong et al., 2014).

From Table 4, we also noted that the prediction accuracy of α/β class has been improved. To be specific, the accuracies for α/β class are 78.9%, 79.5%, 79.3% and 73.4 for 25PDB, 1189, 640 and FC699 datasets respectively, which are 5.7%, 2.3%, 9.8% and 1.2% higher than the next-competing methods (Zhang et al., 2014; Kong et al., 2014). All in all, the above comparison shows that the proposed method outperforms the available PSSM-based and PSSM-free prediction methods, which indicates that the RedPSSM structural properties and position-based structural features reflect some critical information related to protein structural class. This understanding can be then used to develop more powerful methods for protein structural class prediction.

3.3. Influence of parameters in RedPSSM structural properties

A feature of the proposed method is the structural properties of RedPSSM, which describes globe correlation among different columns and simplifies PSSM structures with the reduced amino acids. However, it should be noted that the RedPSSM structural properties are associated

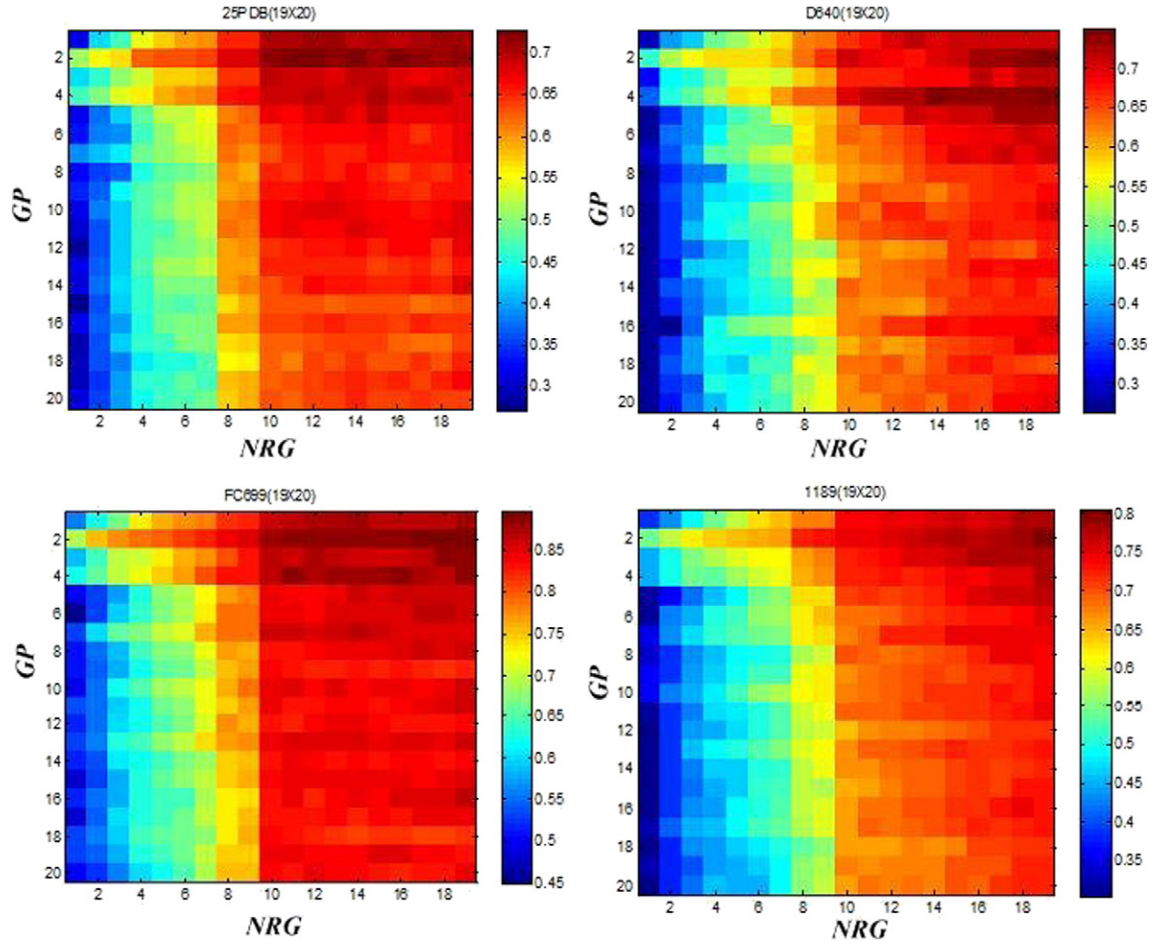

Fig. 1. Overall accuracies of all experiments with each combination of parameters. The number of reduced groupings for 20 amino acids (NRG) and *g* positions (GP) using the jackknife cross-validation test for four datasets, and the color from blue to red denotes the overall accuracy from low to high.

with the number of reduced groups for 20 amino acids (*NRG*) and *g* positions (*GP*). For a better understanding of these parameters, we calculated the structural properties of RedPSSM with *NRG* from 2 to 20 (listed in Table 2) and *GP* from 1 to 20. All experiments were performed with each combination of parameter values using the jackknife cross-validation test, and overall accuracy was chosen to represent the score in this prediction. Fig. 1 is the overall accuracies of all experiments with each combination of parameter values using the jackknife cross-validation test for four datasets, and the color from blue to red denotes the overall accuracy from low to high.

As would be expected, the prediction accuracies with different parameters *NRG* and *GP* show two clear trends. One is that overall accuracy changes for 25PDB, D640, FC699 and 1189 are similar in spite of their different performances, and the other is that overall prediction accuracy increases as *NRG* increases, but it decreases as *g* positions (*GP*) increase.

To further study *NRG*, twenty experiments were performed with given $GP = 2$ and *NRG* from 2 to 20, and overall accuracy was represented in Fig. 2. Fig. 2 shows that there is a considerable increase of the overall accuracy which occurred with *NRG* from 2 to 13. When the *NRG* is over 13, the overall accuracy remains steady. The overall accuracies for datasets 25PDB, 1189, FC699 and 640 are 72.80%, 75.55%, 88.69% and 68.59% respectively, and there is not a great deal of difference compared with the best combination of parameters *NRG* and *GP*.

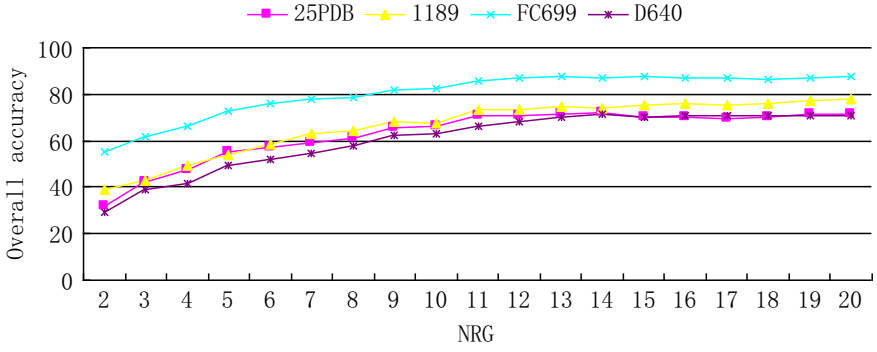

Fig. 2. Overall accuracies of all experiments with GP = 1 and number of reduced groupings (NRG) from 2 to 20 using the jackknife cross-validation test for four datasets.

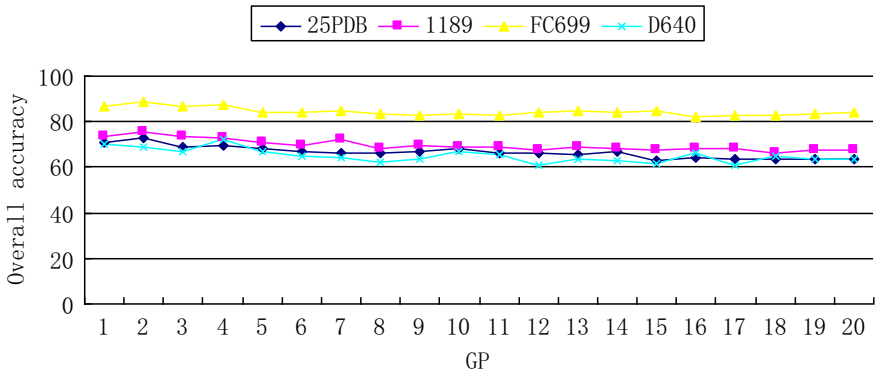

Fig. 3. Overall accuracies of all experiments with NRG = 13 and *g* position (GP) from 1 to 20 using the jackknife cross-validation test for four datasets.

To optimize the parameter *GP*, we also performed 20 experiments with given *NRG* = 13 and *GP* from 1 to 20 for each dataset, and the results were shown in Fig. 3. From Fig. 3, it is apparent that the best performance of prediction achieved as *g* positions (*GP*) is equal to 2. Therefore, this paper selected *NRG* = 13 and *GP* = 2 to construct the structural properties of the RedPSSMs.

The above analysis provides us a basic understanding of the parameters in RedPSSM structural properties, from which novel valuable guidelines for the use of RedPSSM structural properties were obtained. It is interesting to note that *g* position (*GP* = *2*) in RedPSSM structural properties is not random, but closely related to the definition of protein structural classes. As is known, proteins can be classified into four structural classes dominated by different α-helices and β-strands. The alpha helix is a right hand-coiled or spiral conformation (helix) in which every backbone N–H group donates a hydrogen bond to the backbone C═O group of the amino acid four residues earlier (i + 4 to i hydrogen bonding), and a beta strand is a stretch of polypeptide chain typically 3 to 10 amino acids long with backbone in an almost fully extended conformation. Therefore, the average interval position in alpha helices and beta strands is also close to 2. That is to say, the proposed RedPSSM structural properties are consistent with the α-helices' and β-strands' conformation.

3.4. Comparison of accuracies between different classification algorithms

Support vector machine (SVM) was employed as a classifier in this work. To compare different classifiers' performance, k-nearest neighbor algorithm (K-NN) and back-propagation (BP) neural network were also adopted for protein structural class prediction. Here, we used K-NN with K = 1 and BP-NN with 9 neurons in the hidden layer, which were implemented in Matlab toolbox. All experiments were performed based on the reduced PSSM and position-based secondary structural features using jackknife test, and the overall classification accuracies as well as the accuracies for each structural class were listed in Fig. 4.

From Fig. 4, it is easy to note that the SVM classifier achieved the best performance among the three classifiers. Specifically, the average overall prediction accuracy is 89.8% for 25PDB, 1189, 640 and FC699 datasets compared with 83.0% of the K-NN and 71.4% of the BP-NN.

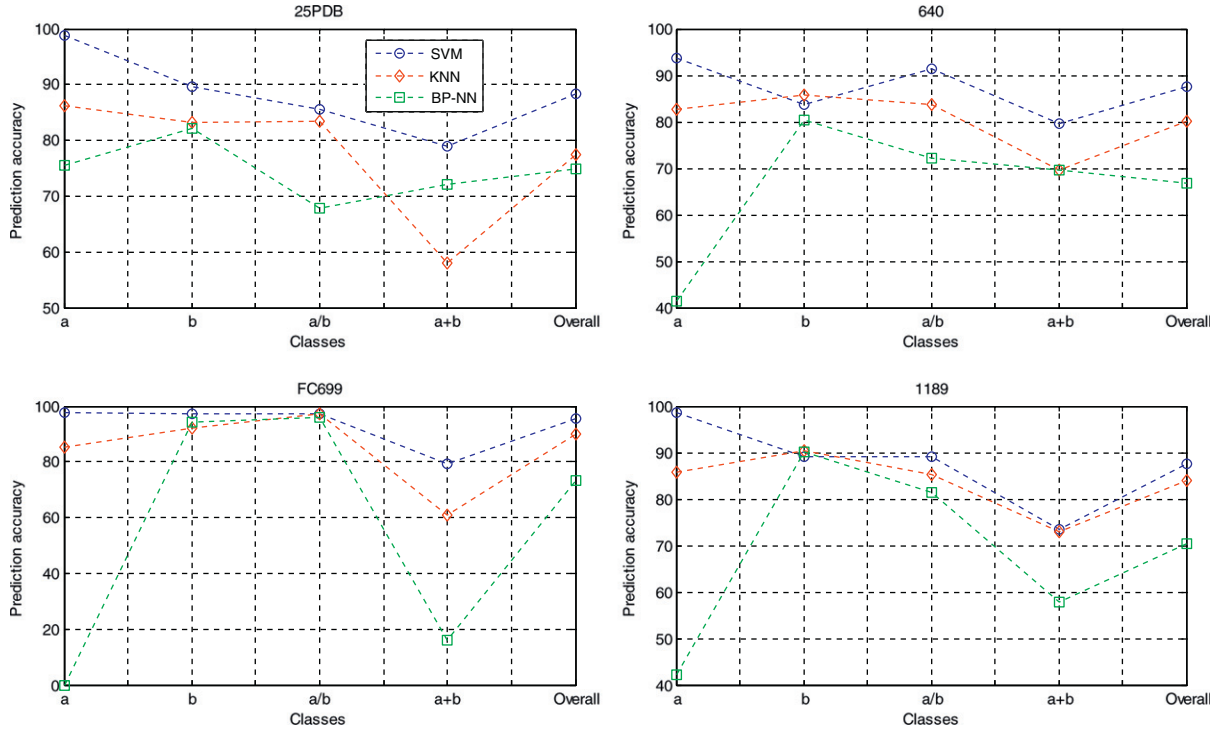

Fig. 4. Performance comparison of different classification algorithms on the four datasets.

This result may indicate that the SVM is a more powerful classifier for protein structural class prediction.

4. Conclusion

Assignment protein structural class gives some useful information on overall folding type study, especially for proteins with low sequence similarity. This paper proposed a novel scheme to predict protein structural classes, which explores the structural properties of the RedPSSM and position-based structural features. To do so, we first reduced 20 amino acids into several groups with reserving maximal information of proteins and used them to simplify the structure of the position specific scoring matrix. The experiment results indicate that the reduced alphabets with size 13 simplify PSSM structures efficiently while reserving its maximal information. With the help of auto covariance transformation, we measured the global structural properties of RedPSSM among different columns. For a better understanding of g positions, the performance optimization of this parameter was performed and the best performance of prediction was achieved with g positions 2. Instead of focusing on content-based structural features, we combined the reduced PSSM with our proposed position-based structural features for protein structural class prediction. We also evaluated the proposed method with four experiments and compared it with the available competing prediction methods. The results show that the proposed method achieved the best performance among the evaluated methods, with overall accuracy 3–5% higher than the existing best-performing method, which indicates that the RedPSSM structural properties and position-based structural features reflect some critical information related to protein structural class. This understanding can be then used to develop more powerful methods for protein structural class prediction.

Acknowledgments

We thank the referees for many valuable comments that have improved this manuscript. This work is supported by the National Natural Science Foundation of China (61170316, 61370015 and 61272312), and research grants (LY14F020046) from the Zhejiang Provincial Natural Science Foundation of China, and the 521 Talent Cultivation Plan of Zhejiang Sci-Tech University (11610032521301).

References

Klein, P., Delisi, C., 1986. Prediction of protein structural class from the amino-acid sequence. Biopolymers 25, 1659–1672.
Chou, K.C., 2006. Structural bioinformatics and its impact to biomedical science and drug discovery. Front. Med. Chem. 3, 455–502.
Levitt, M., Chothia, C., 1976. Structural patterns in globular proteins. Nature 261, 552–558.
Andreeva, A., Howorth, D., Brenner, S.E., Hubbard, T.J., Chothia, C., Murzin, A.G., 2004. SCOP database in 2004: refinements integrate structure and sequence family data. Nucleic Acids Res. 32, D226–D229.
Murzin, A.G., Brenner, S.E., Hubbard, T., Chothia, C., 1995. SCOP: a structural classification of protein database for the investigation of sequence and structures. J. Mol. Biol. 247, 536–540.
Ferragina, P., Giancarlo, R., Greco, V., Manzini, G., Valiente, G., 2007. Compression-based classification of biological sequence and structures via the universal similarity metric: experimental assessment. BMC Bioinforma. 8, 252.
Dai, Q., Wang, T.M., 2008. Comparison study on k-word statistical measures for protein: from sequence to 'sequence space'. BMC Bioinforma. 9, 394.
Chen, C., Tian, Y., Zou, X., Cai, P., Mo, J., 2006. Using pseudo-amino acid composition and support vector machine to predict protein structural class. J. Theor. Biol. 243, 444–448.
Chou, K., 2000. Review: prediction of protein structural classes and subcellular locations. Curr. Protein Pept. Sci. 1, 171–208.
Kedarisetti, K.D., Kurgan, L.A., Dick, S., 2006. Classifier ensembles for protein structural class prediction with varying homology. Biochem. Biophys. Res. Commun. 348, 981–988.
Dai, Q., Wu, L., Li, L.H., 2011. Improving protein structural class prediction using novel combined sequence information and predicted secondary structural features. J. Comput. Chem. 32, 3393.
Chou, K.C., 1999. A key driving force in determination of protein structural classes. Biochem. Biophys. Res. Commun. 264, 216–224.
Chou, K.C., Shen, H.B., 2007. Recent progress in protein subcellular location prediction. Anal. Biochem. 370, 1–16.
Luo, R.Y., Feng, Z.P., Liu, J.K., 2002. Prediction of protein structural class by amino acid and polypeptide composition. Eur. J. Biochem. 269, 4219–4225.
Sun, X.D., Huang, R.B., 2006. Prediction of protein structural classes using support vector machines. Amino Acids 30, 469–475.
Zhang, S.L., Liang, Y.Y., Yuan, X.G., 2014. Improving the prediction accuracy of protein structural class: approached with alternating word frequency and normalized Lempel–Ziv complexity. J. Theor. Biol. 341, 71–77.
Ding, Y.S., Zhang, T.L., Chou, K.C., 2007. Prediction of protein structure classes with pseudo amino acid composition and fuzzy support vector machine network. Protein Pept. Lett. 14, 811–815.
Wu, L., Dai, Q., Han, B., Zhu, L., Li, L.H., 2011. Combining Sequence Information and Predicted Secondary Structural Feature to Predict Protein Structural Classes. ICBBE, pp. 1–4.
Liao, B., Xiang, Q., Li, D., 2012. Incorporating secondary features into the general form of chou's PseAAC for predicting protein structural class. Protein Pept. Lett. 19, 1133–1138.
Kong, L., Zhang, L.C., Lv, J.F., 2014. Accurate prediction of protein structural classes by incorporating predicted secondary structure information into the general form of Chou's pseudo amino acid composition. J. Theor. Biol. 344, 12–18.
Chou, K.C., Cai, Y.D., 2004. Prediction of protein subcellular locations by GO-FunD-PseAA predictor. Biochem. Biophys. Res. Commun. 321, 1007–1009.
Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402.
Stormo, G.D., Schneider, T.D., Gold, L., Ehrenfeucht, A., 1982. Use of the 'Perceptron' algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res. 10, 2997–3011.
Jones, D.T., 1999. Protein secondary structure prediction based on position-specific scoring matrices. J. Mol. Biol. 292, 195–202.
Shen, H.B., Chou, K.C., 2007. Nuc-PLoc: a new web-server for predicting protein subnuclear localization by fusing PseAA composition and PsePSSM. Protein Eng. Des. Sel. 20, 561–567.
Chen, K., Kurgan, L.A., Ruan, J.S., 2008. Prediction of protein structural class using novel evolutionary collocation-based sequence representation. J. Comput. Chem. 29, 1596–1604.
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.
Liu, T., Geng, X., Zheng, X., Li, R., Wang, J., 2012. Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles. Amino Acids 42, 2243–2249.
Ding, S.Y., Yan, S.J., Qi, S.H., Li, Y., Yao, Y.H., 2014. A protein structural classes prediction method based on PSI-BLAST profile. J. Theor. Biol. 353, 19–23.
Kurgan, L.A., 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, 2323–2343.
Kurgan, L., Cios, K., Chen, K., 2008. SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences. BMC Bioinforma. 9, 226–240.
Zheng, C., Kurgan, L., 2008. Prediction of beta-turns at over 80% accuracy based on an ensemble of predicted secondary structures and multiple alignments. BMC Bioinforma. 9, 430.
Mizianty, M.J., Kurgan, L., 2009. Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences. BMC Bioinforma. 10, 414.
Liu, T., Jia, C.Z., 2010. A high-accuracy protein structural class prediction algorithm using predicted secondary structural information. J. Theor. Biol. 267, 272–275.
Zhang, S.L., Ding, S.Y., Wang, T.M., 2011. High-accuracy prediction of protein structural class for low-similarity sequence based on predicted secondary structure. Biochimie 93, 710–714.
Hobohm, U., Sander, C., 1994. Enlarged representative set of protein structures. Protein Sci. 3, 522–524.
Zhang, L.C., Zhao, X.Q., Kong, L., 2013. A protein structural class prediction method based on novel features. Biochimie 95, 1741–1744.
Dai, Q., Li, Y., Liu, X.Q., Yao, Y.H., Cao, Y.J., He, P.A., 2013. Comparison study on statistical features of predicted secondary structures for protein structural class prediction: from content to position. BMC Bioinforma. 14, 152.
Vapnik, V., 2000. The Nature of Statistical Learning Theory. Springer Verlag.
Kurgan, L., Chen, K., 2007. Prediction of protein structural class for the twilight zone sequences. Biochem. Biophys. Res. Commun. 357, 453–460.
Li, T., Fan, K., Wang, J., Wang, W., 2003. Reduction of protein sequence complexity by residue grouping. Protein Eng. 1, 323–330.
Cai, Y., Liu, X., Xu, X., Chou, K., 2002. Prediction of protein structural classes by support vector machines. Comput. Chem. 26, 293–296.
Yuan, Z., Bailey, T.L., Teasdale, R.D., 2005. Prediction of protein B-factor profiles. Proteins 58, 905–912.
Yang, J.Y., Peng, Z.L., Chen, X., 2010. Prediction of protein structural classes for low-homology sequences based on predicted secondary structure. BMC Bioinforma. 11, S9.
Ding, S.Y., Zhang, S.L., Li, Y., Wang, T.M., 2012. A novel protein structural classes prediction method based on predict secondary structure. Biochimie 94, 1166–1171.
Xia, X.Y., Ge, M., Wang, Z.X., Pan, X.M., 2012. Accurate prediction of protein structural class. PLoS ONE 7, 37653.