https://doi.org/10.20965/jaciii.2016.p0402

Paper:

Improving the Prediction of Protein Structural Class for Low-Similarity Sequences by Incorporating Evolutionary and Structural Information

Liang Kong*,, Lingfu Kong**, and Rong Jing****

*School of Mathematics and Information Science & Technology, Hebei Normal University of Science & Technology
Qinhuangdao, China
**School of Information Science and Engineering, Yanshan University
Qinhuangdao, China
E-mail: kongliangouc@126.com

[Received April 24, 2015; accepted December 25, 2015]

Protein structural class prediction is beneficial to study protein function, regulation and interactions. However, protein structural class prediction for low-similarity sequences (i.e., below 40% in pairwise sequence similarity) remains a challenging problem at present. In this study, a novel computational method is proposed to accurately predict protein structural class for low-similarity sequences. This method is based on support vector machine in conjunction with integrated features from evolutionary information generated with position specific iterative basic local alignment search tool (PSI-BLAST) and predicted secondary structure. Various prediction accuracies evaluated by the jackknife tests are reported on two widely-used low-similarity benchmark datasets (25PDB and 1189), reaching overall accuracies 89.3% and 87.9%, which are significantly higher than those achieved by state-of-the-art in protein structural class prediction. The experimental results suggest that our method could serve as an effective alternative to existing methods in protein structural classification, especially for low-similarity sequences.

Keywords: protein domains, secondary protein structure, protein sequence similarity, support vector machines, position specific scoring matrices

1. Introduction

Protein structural class is an important property for characterizing the over folding type of a protein and plays an important role in studying protein function, regulation and interactions [1, 2]. Since Levitt and Chothia introduced the concept of protein structural class in 1976 [3], the identification of protein structural class has become one of the hot topics in protein science [1, 4]. Based on the type, amount and arrangement of the secondary structure elements, a structurally-known protein is generally categorized into the four major classes: all-α, all-β, α/β, $\alpha+\beta$. Traditional methods manually assign the structural class to a protein by manual inspection, which is a complex and time-consuming process. Therefore, with the rapid development of the genomics and proteomics, it is urgently desirable to develop prediction methods to automatically determine structural class for the dramatically expanding newly-discovered proteins.

As a typical statistical pattern recognition problem, computational protein structural class prediction is usually performed in two steps: representation of protein sequences and selection of classification algorithms. The existing sequence representation methods and classification algorithms have been extensively reviewed [4–6]. Many previous structural class prediction methods use features directly from the sequence such as amino acid composition [7, 8], pseudo amino acid composition [9–12], polypeptide composition [13, 14], etc. These methods perform well on high-similarity datasets where similarities between protein sequences are higher than 50%, and the prediction accuracies have reached up to 90%. However, for low-similarity datasets such as the widely-used 25PDB and 1189 datasets (with sequence similarity lower than 25% and 40%, respectively), the reported accuracies range between 50-70% [5]. In order to improve prediction accuracies of low-similarity sequences, several recent methods extract protein features from sequence-derived information such as evolutionary profiles generated with PSI-BLAST [15] and predicted secondary structure. Evolutionary relationship is one of the most important information in biological analysis. Numerous previous studies have illustrated that evolutionary information is more informative than the sequence itself [16, 17]. Overall accuracies of the recently reported evolutionary information based protein structural class prediction methods have been about 75% for some low-similarity datasets [18–20]. Considering the fact that proteins with low sequence similarity but in the same structural class are likely to have high similarity in their corresponding secondary structure elements, several predicted secondary structure based features have been proposed [2, 21–25]. Novel computational predictors that utilize these features have achieved significantly improved accuracies, between 80% and 85% on several low-similarity bench-

mark datasets.

Despite some success in prediction tasks with above advanced features, a carefully engineered integrated feature model generally offers higher accuracy than those with single type of features [26, 27]. In our previous studies [25, 28, 29], we extracted predicted secondary structure based features to reflect the general contents and spatial arrangements of the secondary structure elements of a given protein. Here, in order to further improve the prediction accuracy of protein structural class for low-similarity sequences, we focus on extracting comprehensive features from PSI-BLAST profiles and combining them with several secondary structural features in this study. A total of 148 features are extracted and selected by a filtered feature selection method, and a multi-class nonlinear support vector machine (SVM) classifier is applied to predict protein structural class. The prediction performance is evaluated by jackknife test on two widely-used low-similarity datasets (25PDB and 1189). The experimental results show that the evolutionary features and secondary structural features make complementary contributions to each other. The proposed predictor with integrated feature model provides significantly improved ability to differentiate protein structural classes.

2. Materials and Methods

2.1. Datasets

Sequence similarity has a significant impact on prediction accuracy of protein structural class [5]. Datasets with sequence similarity ranging between 20%−40% tend to obtain more reliable and robust results [2]. In order to facilitate comparison with other existing methods, two widely-used low-similarity protein datasets are selected to design and assess our proposed method. The 25PDB dataset contains 1673 proteins with less than 25% sequence similarity. This dataset was introduced by [5] and extracted from 25% PDBSELECTED list [30] which includes high-resolution non-homologous proteins from the Protein Data Bank (PDB) [31]. The 1189 dataset [5] includes 1092 proteins with sequence similarity lower than 40%. Since sequences in this dataset have lower resolution than proteins in 25PDB dataset, despite higher sequence similarity, similar (or in many case, even lower) prediction accuracy has been reported for 1189 dataset compared to the 25PDB dataset [32]. Since protein structural domains always have limits on size, short sequences are unsuitable for the protein structural class prediction and also cannot be performed by PSI-BLAST. Hence, we remove those sequences with lengths less than 30 residues from the original datasets. For convenience, we still denote the revised datasets as 25PDB and 1189. The contents of these datasets are shown in **Table 1**. Here 1189 dataset is selected for optimization of the feature sets and the parameters in support vector machine, and chosen to predict the structural class of a new protein.

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	442	441	344	441	1668
1189	223	292	331	240	1086

2.2. Feature Representation

2.2.1. PSI-BLAST Profile Based Features

Numerous successful applications of PSI-BLAST profile, which can be represented as a matrix called position specific scoring matrix (PSSM), illustrate that the evolutionary information is more informative than the sequence itself [16, 17, 32–40]. In this study, the PSSM is obtained by PSI-BLAST with parameters h and q set to 0.001 and 3 using every protein sequence as a seed to search and align homogenous sequences from NCBI's non-redundant (NR) protein database (ftp://ftp.ncbi.nih.gov/blast/db/nr), where parameters h and q denote the E-value threshold for inclusion in PSSM and the maximum number of iterations. The generated PSSM is an $L\times 20$ matrix $(p_{i,j})_{L\times 20}$, where L is the length of protein sequence, $p_{i,j}$ represents the conservation score of the amino acid in the ith position of the protein sequence being mutated to amino acid type j during the evolution process. Here the entries of PSSM are scaled to the range from 0 to 1 using the following sigmoid function:

$$f(x)=\frac{1}{1+e^{-x}}, \qquad \ldots\ldots\ldots\ldots (1)$$

where x is the original PSSM value.

Amino acid composition is a wildly-used classical feature model. However, the main deficiency of amino acid composition is ignoring the important sequence order information. To partially overcome this deficiency, dipeptide composition feature model and its variants are proposed. Huang et al. [41] proposed a spaced bipeptide coding method to better describe the local interactions among neighboring amino acids in a protein sequence. The spaced bipeptide coding is to detect the appearance frequency of any two-alphabet pair in interleaving neighboring amino acids of a protein sequence. To partially reflect the local sequence order effect, Liu et al. [18] extended traditional dipeptide composition from the protein amino acid sequence to the PSSM. Furthermore, Ding et al. [27] proposed a pseudo dipeptide composition feature model based on the PSSM to compute the deviation of scores of neighboring amino acid pairs to reflect the local sequence order information. Inspired by their works, we integrate the concepts of long-range correlation and dipeptide composition from PSSM into a unified feature model PSSMF to reflect sequence order information and evolutionary difference information between amino acid pairs. Firstly, evolutionary difference formula between amino acid pairs (represented by the corresponding columns in PSSM) along the protein sequence is defined

as follows:

$$X_{s,t}(i,g) = (p_{i,s} - \bar{p}_{s,t}(i,g))^2 + (p_{i+g,t} - \bar{p}_{s,t}(i,g))^2, \quad (2)$$

where $s,t = 1,2,\ldots,20$; g is a distance factor which determines the degree of separation (spatially) between two amino acids along the protein sequence; $i = 1,2,\ldots,L-g$; $\bar{p}_{s,t}(i,g)) = (p_{i,s} + p_{i+g,t})/2$ represents the average evolutionary score between two amino acids. According to Eq. (2), the average evolutionary difference between amino acid pairs is defined as:

$$\begin{aligned} \mathrm{PSSMF}_{s,t}(g) &= \frac{1}{L-g}\sum_{i=1}^{L-g} X_{s,t}(i,g) \\ &= \frac{1}{L-g}\sum_{i=1}^{L-g} \frac{(p_{i,s} - p_{i+g,t})^2}{2}. \end{aligned} \quad (3)$$

Given a distance factor g, $\mathrm{PSSMF}(g)$ is defined as a 400-dimensional vector:

$$\mathrm{PSSMF}(g) = (\mathrm{PSSMF}_{1,1}(g),\ldots,\mathrm{PSSMF}_{20,20}(g)). \quad (4)$$

Suppose G is the maximum of g ($g = 0,1,2,\ldots,G$), the feature model PSSMF for a protein sequence is a $400 \times (G+1)$-dimensional vector which is constructed as:

$$\mathrm{PSSMF} = \mathrm{PSSMF}(0) \oplus \ldots \oplus \mathrm{PSSMF}(G), \quad (5)$$

where $\oplus$ is the operator of concatenation.

2.2.2. Predicted Secondary Structure Based Features

To effectively extract structural information, a protein amino acid sequence is transformed into the corresponding sequence of secondary structure elements (helix (H), strand (E), coil (C)). In this study, we predict secondary structure using the recently proposed SPINE-X [42, 43] which has better performance than previous widely-used PSIPRED [16]. Besides the predicted secondary structure sequence, SPINE-X also outputs an $L \times 3$ matrix $(s_{i,j})_{L\times 3}$ (denoted by SPINE-M) consisting of the normalized probability of contribution of a given amino acid based on its position along the protein sequence to build one of the three secondary structure elements.

Similar to PSSMF, we propose a structural feature model SPINEF1 based on SPINE-M to partially reflect local sequence order information. Given a distance factor g, the average sequence order correlation factor between two secondary structure elements is defined as:

$$\mathrm{SPINEF1}_{s,t}(g) = \frac{1}{L-g}\sum_{i=1}^{L-g} \frac{(s_{i,s} - s_{i+g,t})^2}{2}. \quad (6)$$

For every value of g, a different 9-dimensional vector is generated for the same protein sequence, which is represented by $\mathrm{SPINEF1}(g)$. Suppose G is the maximum of g ($g = 0,1,2,\ldots,G$), all the $\mathrm{SPINEF1}(g)$ are concatenated to form a feature vector SPINEF1 of $9 \times (G+1)$ dimensions.

In addition to extract structural features from SPINE-M, we also introduce a comprehensive set of 11 predicted secondary structure based features. The details of these features are given as follows:

(1) The contents and second order composition moments of the helix and strand are formulated as:

$$p(x) = \frac{N(x)}{L},\ x \in \{\mathrm{H},\mathrm{E}\}, \quad (7)$$

$$\mathrm{CMV}(x) = \frac{\sum_{k=1}^{N(x)} n_{x_k}}{L(L-1)},\ x \in \{\mathrm{H},\mathrm{E}\}, \quad (8)$$

where $N(x)$ is the number of secondary structural elements; n_{x_k} is the k-th corresponding secondary structural element's order (or position) along the protein sequence. For example, given a secondary structure sequence CCEEEECCCHHEEHH, the length of protein sequence is $L = 15$ and the number of strand is $N(\mathrm{E}) = 6$. The 6 strands order is 3, 4, 5, 6, 12, 13, respectively. According to Eq. (8), the second order composition moment of strand can be computed as:

$$\mathrm{CMV}(\mathrm{E}) = \frac{3+4+5+6+12+13}{15\times(15-1)} = 0.2048. \quad (9)$$

(2) As the objects of structural classification are globular proteins, the size (length) of helix and strand segments is one of the deciding factors when it comes to the assignment of the structural class. In order to utilize this information, normalized maximal lengths of helix and strand segments are proposed as follows:

$$\mathrm{NMaxSeg}(x) = \frac{\mathrm{MaxSeg}(x)}{L},\ x \in \{\mathrm{H},\mathrm{E}\}, \quad (10)$$

where $\mathrm{MaxSeg}(x)$ is the lengths of the longest α-helices (β-strands).

(3) While proteins in the α/β and $\alpha+\beta$ classes contain both α-helices and β-strands, they are usually segregated in the α/β class but are usually interspersed in the $\alpha+\beta$ class. In proteins of the α/β class, α-helices and β-strands alternate more frequently than in proteins of the $\alpha+\beta$ class. The preferred way to represent the spatial arrangements of the secondary structures for structural class prediction is to utilize 3D protein structure. However, since the input is only flat secondary structure sequence, quantifying collocation of helix and strand segments in the predicted secondary structure sequence would be an effective way to approximate this information. Hence, we construct a simplified segment sequence from the predicted secondary structure sequence in the following steps: (1) every H, E and C segment is respectively replaced by the individual letter H, E and C, (2) all of the letters C are removed. Based on the segment sequence, we count the number of helix-coil-helix motifs (two α-helices separated by a coil segment), strand-coil-strand motifs (two β-strands separated by a coil segment), helix-coil-strand motifs (α-helices and β-strands separated by a coil segment) and strand-coil-helix motifs (β-strands and α-helices separated by a coil segment). Then, the normalized alternating frequency of α-helices and β-strands, the helices bundle probability and the sheets probability are

respectively defined as:

$$\mathrm{NAltn} = \frac{N'(\mathrm{HE}) + N'(\mathrm{EH})}{L}, \quad \ldots\ldots\ldots (11)$$

$$p'(xx) = \frac{N'(xx)}{L},\ xx \in \{\mathrm{HH}, \mathrm{EE}\}, \quad \ldots\ldots (12)$$

where $N'(xx)$ is the number of substring HE, EH, HH or EE in the segment sequence. In addition, to reflect the level of separation for α-helices and β-strands, the normalized maximal distance between the adjacent α-helices and β-strands as well as β-strands and α-helices in the predicted secondary structure sequence are defined as:

$$\mathrm{NMaxD}(xx) = \frac{\mathrm{MaxD}(xx)}{L},\ xx \in \{\mathrm{HE}, \mathrm{EH}\}, \quad . (13)$$

where HE denotes segment from α-helices to the adjacent β-strands, and EH denotes segment from β-strands to the adjacent α-helices in the predicted secondary structure sequence. From the above description, a 11-dimensional structural feature vector can be constructed and formally denoted by SPINEF2 for the rest of this study.

2.3. Feature Selection

Among the above three feature models, the number of features in PSSMF and SPINEF1 vary with the maximum value of distant factor g (denoted by G). In this study, the value of parameter G is set to 9, and the total number of features in PSSMF and SPINEF1 are 4000 and 90, respectively. Due to the large number of features, irrelevant and redundant information will be inevitable which can result in less effective prediction. Feature selection is the process of identifying and removing as much irrelevant and redundant features as possible. This will enable a more efficient prediction model, and helps speed up the computational analysis time. Many feature selection methods have been used in a wide range of bioinformatics studies [44–46]. In this study, a correlation-based feature subset selection method (CFS) [47, 48] is adopted. CFS is a filtering method which identifies a small set of nonredundant features that are highly correlated with the outcome while having low correlation among themselves. Specifically, CFS uses a correlation based heuristic to evaluate the worth of features:

$$Merit_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}, \quad \ldots\ldots\ldots (14)$$

where $Merit_S$ is the heuristic "merit" of a feature subset S containing k features, $\overline{r_{cf}}$ measures the average dependence between k features and the class label, and $\overline{r_{ff}}$ measures the average dependence among k features in S. The above average dependences $\overline{r_{cf}}$ and $\overline{r_{ff}}$ are computed by information gain. Obviously, the value of $Merit_S$ increases when the selected features are highly informative about the outcome, but decreases when there is a high correlation among those features. CFS implemented hill-climbing optimization as in the best first search with five levels of backtracking, which iteratively expands the feature subset S starting from an empty set

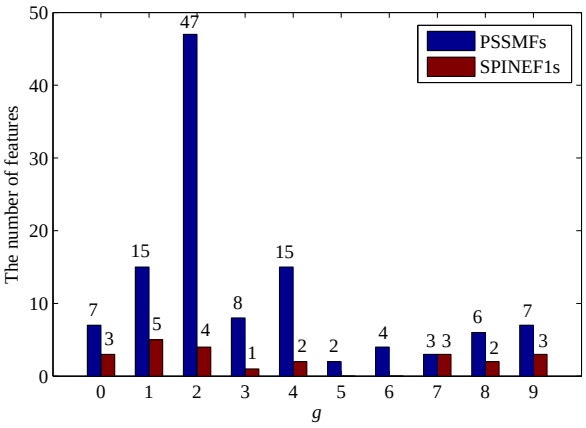


Fig. 1. The number of selected features using CFS on 1189 dataset (g ranges from 0 to 9).

to identify a better S based on the merit value among all the possible expansions at each step until there are five consecutive nonimproving expansions. Here, we perform feature selection based on 1189 dataset. As a result, 114 and 23 features are selected among the original PSSMF and SPINEF1. The selected feature numbers with varying distance factor g are shown in **Fig.1**. For convenience, the above two feature subsets are denoted by PSSMFs and SPINEF1s, respectively. Finally, given a protein sequence, a 148-dimensional feature vector (PSSMFs+SPINEF1s+SPINEF2) is constructed and then used to predict protein structural class.

2.4. Classification Algorithm Construction

Support vector machine (SVM) [49], a particular learning system based on Vapnik's statical learning theory, has been widely used to deal with various important biological problems [50–53]. There are four kinds of kernel functions, i.e. linear function, polynomial function, sigmoid function and radial basis function (RBF), are commonly used to perform prediction. Here, the publicly available software package LIBSVM [54] with RBF is adopted. The best combination of penalty parameter C and kernel parameter γ are selected by 10-fold cross-validation with a simple but effective grid search strategy. The parameters C and γ are searched exponentially in the ranges of $[2^{-5}, 2^{15}]$ and $[2^{-15}, 2^{5}]$, respectively, with a step size of 2^1 to probe the highest classification rate.

2.5. Performance Measures

In statistical prediction, independent dataset test, subsampling test and jackknife test are often used to examine a predictor for its effectiveness in practical application [55]. Among the three test methods, the jackknife test is deemed the least arbitrary that can always yield a unique result for a given benchmark dataset [6]. Hence the jackknife test is employed to examine the performance of our method. For comprehensive evaluation, the overall accuracy (the number of correct predictions divided by the

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

Dataset	Class	Sens(%)	Spec(%)	MCC
25PDB	All-α	97.3	98.7	0.957
	All-β	91.2	96.9	0.881
	α/β	87.2	97.3	0.852
	$\alpha+\beta$	81.2	92.8	0.737
	OA	89.3		
1189	All-α	94.6	98.0	0.919
	All-β	91.4	98.6	0.915
	α/β	88.2	95.1	0.835
	$\alpha+\beta$	77.1	92.2	0.682
	OA	87.9		

total number of test sequences, denoted by OA), the individual sensitivity (or accuracy, denoted by Sens), the individual specificity (Spec) and Matthew's correlation coefficient (MCC) over each of the four structural classes are reported. These parameters are detailed as follows [18]:

$$\mathrm{Sens}_j = \frac{\mathrm{TP}_j}{\mathrm{TP}_j + \mathrm{FN}_j} = \frac{\mathrm{TP}_j}{|\mathrm{C}_j|}, \quad \ldots\ldots\ldots (15)$$

$$\mathrm{Spec}_j = \frac{\mathrm{TN}_j}{\mathrm{FP}_j + \mathrm{TN}_j} = \frac{\mathrm{TN}_j}{\sum_{k\neq j}|\mathrm{C}_k|}, \quad \ldots\ldots (16)$$

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

where TN_j, TP_j, FN_j, FP_j and $|\mathrm{C}_j|$ are the number of true negatives, true positives, false negatives, false positives and proteins in the structural class C_j, respectively. The MCC value ranges between -1 and 1 with 0 denoting random prediction and higher absolute values denoting more accurate predictions.

3. Results and Discussion

3.1. Prediction Performance of Our Method

We report the results of jackknife tests performed on 25PDB and 1189 datasets in **Table 2**. As can be seen, the overall accuracies of the two datasets are all above 87%. Comparing the prediction accuracies of four structural classes with each other, the Sens, Spec and MCC values of the all-α class are always highest. It indicates that the prediction for the all-α class is most reliable. The main reason for good performance for the all-α class is that these sequences are helix rich and helical structures are the easiest to predicted, i.e., a helix is formed by a single, continuous sequence segment and is characterized by highly repetitive structure [21]. Meanwhile, the results of the all-β and α/β classes are also satisfactory with the accuracies nearly 90%. However, the prediction accuracies of the $\alpha+\beta$ class are inferior to those of other three classes, suggesting that difficulty existed in recognizing the anti-parallel β-sheets [25]. This trend is universal for all protein structural class prediction methods, although the corresponding accuracies are lower.

As mentioned earlier, three groups of features (PSSMFs, SPINEF1s and SPINEF2) are extracted and selected to represent a protein. In order to further investigate how these feature subsets contribute to the prediction performance, we compare the accuracies among all the possible combinations of feature subsets, and the results are listed in **Table 3**. It can be seen that when the feature subsets are used individually, the overall accuracies and accuracies of four structural classes obtained by predicted secondary structure based features (SPINEF1s and SPINEF2) are higher than evolutionary features (PSSMFs), and those of SPINEF2 are often the highest. As more features are involved in the prediction, the prediction accuracies are shown to increase steadily. For instance, when 25PDB dataset is tested, the overall accuracy and accuracies of four structural classes with PSSMFs are 76.6%, 91.0%, 80.7%, 75.0% and 59.2%, respectively. With addition of SPINEF1s, these accuracies increase by 10.6%, 5.6%, 11.1%, 9.0% and 16.5%. If SPINEF2 is further added, the overall accuracy increases by 2.1% up to 89.3%. Therefore, we may conclude that the three groups of features which characterize a protein from different aspects can make complementary contributions to each other, and combining the PSI-BLAST profile based features and predicted secondary structure features is an effective method to improve the prediction accuracy of protein structural class.

3.2. Comparison with Other Prediction Methods

In this section, we compare our method with the recently reported competing protein structural class prediction methods on the same datasets. Here we rationally classify the compared methods into three groups: (1) AADP-PSSM [18], AATP [19] and AAC-PSSM-AC [20] are recently reported prediction methods based on PSI-BLAST profile; (2) SCPRED [21], RKS-PPSC [2], Liu and Jia [22], Zhang et al. [23], Ding et al. [24] and Zhang et al. [25] are prediction methods based on the predicted secondary structure; (3) MODAS [26] and PSSS-PSSM [27] are prediction methods with integrated features from the PSI-BLAST profile and predicted protein secondary structure. For convenience, the above three groups of methods are respectively denoted by M1, M2 and M3.

The comparison results are shown in **Table 4**. From **Table 4**, we can find that the prediction accuracies obtained by methods M2 are about 10% higher than those of methods M1, and the top two overall and individual accuracies are commonly from methods M3 and our method. As for the 25PDB dataset, our method outperforms all other compared methods. Specifically, the overall accuracy and accuracies of four structural classes are respectively 2.7%, 0.7%, 4.1%, 1.4% and 2.3% higher than previous best-performing results. Moreover, our method is the only method which improves the $\alpha+\beta$ class accuracy up to 80%. Referring to the 1189 dataset, there are only two

Table 3. Performance comparison of different feature subsets on the 25PDB and 1189 datasets.

Dataset	Features	Accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	Overall
25PDB	PSSMFs	91.0	80.7	75.0	59.2	76.6
	SPINEF1s	92.1	83.2	75.9	71.7	81.0
	SPINEF2	94.1	83.0	81.1	73.5	83.0
	PSSMFs+SPINEF1s	96.6	91.8	84.0	75.7	87.2
	PSSMFs+SPINEF2	97.5	88.4	86.6	78.7	87.9
	SPINEF1s+SPINEF2	93.2	84.4	82.6	77.6	84.5
	PSSMFs+SPINEF1s+SPINEF2	97.3	91.2	87.2	81.2	89.3
1189	PSSMFs	85.2	86.0	82.5	47.9	76.3
	SPINEF1s	90.1	88.4	79.2	60.0	79.7
	SPINEF2	92.8	85.3	84.3	73.8	84.0
	PSSMFs+SPINEF1s	94.6	91.4	88.2	67.9	85.9
	PSSMFs+SPINEF2	93.7	89.0	88.8	70.8	85.9
	SPINEF1s+SPINEF2	89.7	91.1	85.8	72.9	85.2
	PSSMFs+SPINEF1s+SPINEF2	94.6	91.4	88.2	77.1	87.9

Table 4. Performance comparison of different methods on the 25PDB and 1189 datasets.

Dataset	Method	Accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	Overall
25PDB	AADP-PSSM [18]	83.3	78.1	76.3	54.4	72.9
	AATP [19]	81.9	74.7	75.1	55.8	71.7
	AAC-PSSM-AC [20]	85.3	81.7	73.7	55.3	74.1
	SCPRED [21]	92.6	80.1	74.0	71.0	79.7
	RKS-PPSC [2]	92.8	83.3	85.8	70.1	82.9
	Liu and Jia [22]	92.6	81.3	81.5	76.0	82.9
	Zhang et al. [23]	95.0	85.6	81.5	73.2	83.9
	Ding et al. [24]	95.0	81.3	83.2	77.6	84.3
	Zhang et al. [25]	95.7	80.8	82.4	75.5	83.7
	MODAS [26]	92.3	83.7	81.2	68.3	81.4
	PSSS-PSSM [27]	96.6	87.1	83.0	78.9	86.6
	Our study	**97.3**	**91.2**	**87.2**	**81.2**	**89.3**
1189	AADP-PSSM [18]	69.1	83.7	85.6	35.7	70.7
	AATP [19]	72.7	85.4	82.9	42.7	72.6
	AAC-PSSM-AC [20]	80.7	86.4	81.4	45.2	74.6
	SCPRED [21]	89.1	86.7	**89.6**	53.8	80.6
	RKS-PPSC [2]	89.2	86.7	82.6	65.6	81.3
	Zhang et al. [23]	92.4	87.4	82.0	71.0	83.2
	Ding et al. [24]	93.7	84.0	83.5	66.4	82.0
	Zhang et al. [25]	92.4	84.4	84.4	73.4	83.6
	MODAS [26]	92.3	87.1	87.9	65.4	83.5
	PSSS-PSSM [27]	94.2	88.4	85.3	71.8	85.0
	Our study	**94.6**	**91.4**	88.2	**77.1**	**87.9**

methods that provide the overall accuracy over 85%. One is our method, and the other is PSSS-PSSM. However, our result is 2.9% higher than PSSS-PSSM. In this study, PSSMFs are designed to reflect sequence order information and evolutionary difference information between amino acid pairs. Overall accuracies obtained by PSSMFs are 76.6% and 76.3% on 25PDB and 1189 datasets (see **Table 3**), which are obviously higher than those of the similar methods M1. Likewise, it can be seen from **Tables 3** and **4** that the proposed predicted secondary structure based features (SPINEF1s+SPINEF2) also obtains competitive prediction accuracies when compared to the similar methods M2. Particularly, the overall accuracy of 1189 dataset is 85.2%, which is 1.6% higher than the next best one proposed by Zhang et al. [25]. Therefore, we attribute the high prediction accuracy achieved by our method to the carefully designed integrated feature model which effectively characterizes a protein from different aspects.

Among the predictions of four structural classes, the predictions of the α/β and $\alpha+\beta$ classes are relatively difficult. In order to further demonstrate the effectiveness

Table 5. The accuracies of differentiating between the α/β and $\alpha+\beta$ classes.

Dataset	Method	Accuracy(%)		
		α/β	$\alpha+\beta$	Overall
25PDB	SCPRED [21]	76.0	83.2	80.1
	RKS-PPSC [2]	86.4	82.8	84.4
	PSSS-PSSM [27]	84.1	88.4	86.5
	Our study	**89.0**	**92.7**	**91.1**
1189	SCPRED [21]	88.6	63.1	77.9
	RKS-PPSC [2]	83.8	**81.3**	82.8
	PSSS-PSSM [27]	87.4	77.2	83.1
	Our study	**90.3**	80.4	**86.2**

of the proposed method in differentiating between the α/β and $\alpha+\beta$ classes, another experiment is performed and the results are listed in **Table 5**. Similar to RKS-PPSC, we generate a subset for each benchmark dataset by removing all the proteins in the all-α and all-β classes to avoid any potential outside effects. Then we predict the accuracies of the α/β and $\alpha+\beta$ classes on these reduced subsets instead of the whole dataset. To differentiate with 25PDB and 1189 datasets, 25PDBs and 1189s are used to denote the corresponding subsets. As can be seen from **Table 5**, our method outperforms all the compared methods on both datasets. The increments of overall accuracies are 4.6% and 3.1%, respectively. In addition, the MCC values of the α/β and $\alpha+\beta$ classes are also computed. As for 25PDB dataset,the MCC values of the two classes are all 0.819. As for 1189 dataset, the MCC values of the two classes are all 0.715. The above results clearly shows that our method is essential to achieve good prediction performance for differentiating between the α/β and $\alpha+\beta$ classes.

3.3. Comparison with Different Classification Algorithms

To evaluate the prediction performance of different classification algorithms, we consider other five classification algorithms which are based on complementary model types: Naive Bayes, linear logistic regression, k-Nearest Neighbor with $k=3$, linear discriminant analysis and decision tree. The selection is motivated by their prior successful applications in the context of the structural class predictions, i.e., Naive Bayes based classifier was used in [56], logistic regression in [5, 35], nearest neighbor in [12, 57], linear discriminant analysis in [2] and decision tree [58, 59]. All experiments are performed using jackknife test, and the overall accuracies as well as the accuracies for each structural classes are listed in **Table 6**. It can be seen that linear logistic regression, linear discriminant analysis and SVM obviously outperform other three classification algorithms, and SVM are shown to perform best among all the classification algorithms. Although k-nearest neighbor algorithms obtains the highest α/β class accuracies on both datasets, the overall accuracies and the $\alpha+\beta$ class accuracies are much lower. These experimental results indicates that the SVM is more suitable for protein structural class prediction, which is consistent with the successful prior application of this classification algorithm.

3.4. Discussion on the Relationship Between Feature Patterns of PSSMF or SPINEF1 and Protein Structural Class

In order to describe the long-range correlation from PSSM and SPINE-M, different distant factors g are adopted to extract PSSMF or SPINEF1 features. Here each feature groups with a g can be considered as a feature pattern. In this section, we discuss the the relationship between the feature patterns and protein structural class. Based on the protein structural class C_j (all-α, all-β, α/β and $\alpha+\beta$) to which the proteins belong, we first divide the datasets into two subsets, one subset consists of proteins from structural class C_j as positive samples, the other contains proteins which not from structural class C_j as negative samples. Then, we predict the protein structural class C_j as binary classification using different feature pattern PSSMF(g)(SPINEF1(g)) with varying distance factor g, and the results are listed in **Tables 7** and **8**. Here we use the MCC value since this measure, in contrast to accuracy, takes into account the unbalanced nature of the datasets. It can be seen that the PSSMF features corresponding to $g=2$ perform best to predict all the four structure classes for 25PDB dataset. As for 1189 dataset, PSSMF features corresponding to $g=2$ achieve the highest MCC value for the all-α class, $g=1$ features achieve the highest MCC value for the all-β class, and $g=3$ features perform best for the α/β and $\alpha+\beta$ classes. Hence, it can be concluded that the PSSMF features corresponding to smaller distance factors are more effective to predict the protein structure class. When the SPINEF1 features are tested, the trend is somewhat different. The SPINEF1 features corresponding to $g=6,7,8,9$ always perform well to predict the structure class for two datasets. Hence, the SPINEF1 features corresponding to bigger distance factors are more effective to predict the protein structure class. As for the profile-based protein features are less explicit than the sequence-based protein features, further investigations about the relationship between the feature patterns of PSSMF or SPINEF1 and some specific sequence characteristics will constitute an interesting subject for our future work.

4. Conclusions

Prediction of protein structural class for low-similarity sequences is a challenging problem. This study proposes a computational method that aims to employ both PSI-BLAST profile based features and predicted secondary structure based features to improve the protein structural class prediction accuracy. Based on comprehensive experimental comparison with the state-of-the-art struc-

Table 6. Performance comparison of different classification algorithms on the 25PDB and 1189 datasets.

Dataset	Method	Accuracy(%)				
		All-α	All-β	α/β	$\alpha+\beta$	Overall
25PDB	*k*-nearest neighbor	93.4	91.4	**94.5**	46.3	80.6
	Naive Bayes	94.1	88.7	84.6	69.8	84.3
	Decision tree	90.5	85.9	77.3	60.5	78.7
	Linear logistic regression	95.0	**91.8**	87.2	79.4	88.4
	Linear discriminant analysis	94.0	87.5	88.7	**82.5**	88.2
	SVM	**97.3**	91.2	87.2	81.2	**89.3**
1189	*k*-nearest neighbor	90.6	90.8	**94.9**	43.3	81.5
	Naive Bayes	**95.1**	88.7	82.5	65.0	82.9
	Decision tree	87.4	87.7	79.8	60.0	79.1
	Linear logistic regression	91.9	**93.2**	88.5	70.8	86.6
	Linear discriminant analysis	89.2	88.7	87.9	**77.1**	86.0
	SVM	94.6	91.4	88.2	**77.1**	**87.9**

Table 7. The prediction quality of PSSMF features on the 25PDB and 1189 datasets (g ranges from 0 to 9).

Dataset	Class	g =0	g =1	g =2	g =3	g =4	g =5	g =6	g =7	g =8	g =9
25PDB	All-α	0.744	0.752	0.791	0.761	0.76	0.723	0.723	0.722	0.724	0.716
	All-β	0.651	0.671	0.7	0.666	0.652	0.669	0.643	0.645	0.631	0.622
	α/β	0.67	0.687	0.691	0.691	0.683	0.657	0.643	0.639	0.623	0.642
	$\alpha+\beta$	0.425	0.433	0.461	0.439	0.397	0.443	0.386	0.396	0.42	0.398
1189	All-α	0.744	0.761	0.791	0.732	0.724	0.726	0.705	0.703	0.694	0.664
	All-β	0.713	0.77	0.753	0.709	0.746	0.718	0.722	0.725	0.709	0.701
	α/β	0.64	0.678	0.688	0.695	0.670	0.648	0.645	0.656	0.626	0.619
	$\alpha+\beta$	0.393	0.417	0.408	0.423	0.421	0.4	0.352	0.354	0.343	0.319

Table 8. The prediction quality of SPINEF1 features on the 25PDB and 1189 datasets (g ranges from 0 to 9).

Dataset	Class	g =0	g =1	g =2	g =3	g =4	g =5	g =6	g =7	g =8	g =9
25PDB	All-α	0.864	0.887	0.873	0.873	0.887	0.886	0.893	0.885	0.887	0.881
	All-β	0.788	0.788	0.792	0.791	0.797	0.806	0.81	0.817	0.818	0.824
	α/β	0.538	0.594	0.594	0.593	0.641	0.644	0.667	0.688	0.693	0.67
	$\alpha+\beta$	0.498	0.508	0.516	0.527	0.545	0.566	0.575	0.579	0.576	0.571
1189	All-α	0.797	0.839	0.832	0.827	0.836	0.85	0.851	0.851	0.87	0.856
	All-β	0.807	0.826	0.837	0.839	0.831	0.831	0.834	0.831	0.846	0.863
	α/β	0.608	0.621	0.611	0.63	0.623	0.659	0.68	0.638	0.671	0.658
	$\alpha+\beta$	0.333	0.403	0.37	0.423	0.428	0.447	0.482	0.448	0.47	0.471

ture class prediction methods on two widely-used low-similarity benchmark datasets, the proposed method is shown to be an effective computational tool for protein structural class prediction on low-similarity protein sequences. The outstanding performance of the proposed method can be attributed to the effective usage of the integrated features as well as well-trained SVM. Since user-friendly and publicly accessible web-servers represent the future direction for developing practically more useful models, simulated methods, or predictors [53, 60], we shall make efforts in our future work to provide a web-server for the method presented in this paper.

Acknowledgements

The program file can be obtained by e-mail from the corresponding author. The authors thank the anonymous referees for many valuable suggestions that have improved this manuscript. We express our thanks to Dr. Abdollah Dehzangi for his kindly help. This work is supported by the National Natural Science Foundation of China (Grant No.61305113), the Youth Foundation of Hebei Educational Committee (Grant No.QN2015131), Doctoral Foundation of Hebei Normal University of Science and Technology (Grant No.2013YB008 and 2011YB006).

References:

[1] K. C. Chou, "Structural bioinformatics and its impact to biomedical science," Curr. Med. Chem, Vol.11, pp. 2105-2134, 2004.

[2] J. Yang, Z. Peng, and X. Chen, "Prediction of protein structural classes for low-homology sequences based on predicted secondary structure," BMC Bioinforma., Vol 11, pp. S9, 2010.
[3] M. Levitt and C. Chothia, "Structural patterns in globular proteins," Nature, Vol.261, pp. 552-558, 1976.
[4] K. C. Chou, "Progress in protein structural class prediction and its impact to bioinformatics and proteomics," Curr. Protein Pept. Sci., Vol.6, pp. 423-436, 2005.
[5] L. A. Kurgan and L. Homaeian, "Prediction of structural classes for protein sequences and domains-impact of prediction algorithms, sequence representation and homology, and test procedures on accuracy," Pattern Recognit., Vol.39, pp. 2323-2343, 2006.
[6] K. C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition (50th Anniversary Year Review)," J. Theor. Biol., Vol.273, pp. 236-247, 2011.
[7] K. C. Chou, "A novel approach to predicting protein structural classes in a (20-1)-D amino acid composition space," Proteins, Vol.21, pp. 319-344, 1995.
[8] K. C. Chou, "A key driving force in determination of protein structural classes," Biochem. Biophys. Res. Commun., Vol.264, pp. 216-224, 1999.
[9] K. C. Chou, "Prediction of protein cellular attributes using pseudo amino acid composition," Proteins, Vol.43, pp. 246-255, 2001.
[10] X. Xiao, S. H. Shao, Z. D. Huang, and K. C. Chou, "Using pseudo amino acid composition to predict protein structural classes: approached with complexity measure factor," J.Comput. Chem., Vol.27, pp. 478-482, 2006.
[11] H. Lin and Q. Z. Li, "Using pseudo amino acid composition to predict protein structural class: approached by incorporating 400 dipeptide components," J. Comput. Chem., Vol.28, pp. 1463-1466, 2007.
[12] T. L. Zhang, Y. S. Ding, and K. C. Chou, "Prediction protein structural classes with pseudo-amino acid composition: approximate entropy and hydrophobicity pattern," J. Theor. Biol., Vol.250, pp. 186-193, 2008.
[13] R. Y. Luo, Z. P. Feng, and J. K. Liu, "Prediction of protein structural class by amino acid and polypeptide composition," Eur. J. Biochem., Vol.269, pp. 4219-4225, 2002.
[14] X. D. Sun and R. B. Huang, "Prediction of protein structural classes using support vector machines," Amino Acids, Vol.30, pp. 469-475, 2006.
[15] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped BLAST and PSI-BLAST: A new generation of protein database search programs," Nucleic Acids Res., Vol.25, pp. 3389-3402, 1997.
[16] D. T. Jones, "Protein secondary structure prediction based on position-specific scoring matrices," J. Mol. Biol., Vol.292, pp. 195-202, 1999.
[17] H. Kim and H. Park, "Prediction of protein relative solvent accessibility with support vector machines and long-range interaction 3D local descriptor," Proteins, Vol.54, pp. 557-562, 2004.
[18] T. G. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile," Biochimie, Vol.92, pp. 1330-1334, 2010.
[19] S. L. Zhang, Y. Feng, and X. G. Yuan, "Using principal component analysis and support vector machine to predict protein structural class for low-similarity sequences via PSSM," J. Biomol. Struct. Dyn. Vol.29, pp. 634-642, 2012.
[20] T. Liu, X. Geng, X. Zheng, R. Li, and J. Wang, "Accurate prediction of protein structural class using auto covariance transformation of PSI-BLAST profiles," Amino Acids, Vol.42, pp. 2243-2249, 2012.
[21] L. A. Kurgan, K. Cios, and K. Chen, "SCPRED: accurate prediction of protein structural class for sequences of twilight-zone similarity with predicting sequences," BMC Bioinforma., Vol.9, pp. 226, 2008.
[22] T. Liu and C. Jia, "A high-accuracy protein structural class prediction algorithm using predicted secondary structural information," J. Theor. Biol., Vol.267, pp. 272-275, 2010.
[23] S. Zhang, S. Ding, and T. Wang, "High-accuracy prediction of protein structural class for low-similarity sequences based on predicted secondary structure," Biochimie, Vol.93, pp. 710-714, 2011.
[24] S. Ding, S. Zhang, Y. Li, and T. Wang, "A novel protein structural classes prediction method based on predicted secondary structure," Biochimie, Vol.94, pp. 1166-1171, 2012.
[25] L. Zhang, X. Zhao, and L. Kong, "A protein structural class prediction method based on novel features," Biochimie, Vol.95, pp. 1741-1744, 2013.
[26] M. J. Mizianty and L. Kurgan, "Modular prediction of protein structural classes from sequences of twilight-zone identity with predicting sequences," BMC Bioinforma., Vol.10, pp. 414, 2009.
[27] S. Ding, Y. Li, Z. Shi, and S. Yan, "A protein structural classes prediction method based on predicted secondary structure and PSI-BLAST profile," Biochimie, Vol.97, pp. 60-65, 2014.
[28] L. Kong, L. Zhang, and J. Lv, "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., Vol.344, pp. 12-18, 2014.
[29] L. Kong and L. Zhang, "Novel structure-driven features for accurate prediction of protein structural class," Genomics, Vol.103, No.4, pp. 292-297, 2014.
[30] U. Hobohm and C. Sander, "Enlarged representative set of protein structures," Protein Sci., Vol.3, pp. 522-524, 1994.
[31] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank," Nucleic Acids Res., Vol.28, pp. 235-242, 2000.
[32] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Proposing a highly accurate protein structural class predictor using segmentation-based features," BMC Genomics, Vol.15, pp. S2, 2014.
[33] H. Saini, G. Raicar, A. Sharma, S. Lal, A. Dehzangi, J. Lyons, K. Paliwal, S. Imoto, and S. Miyano, "Probabilistic expression of spatially varied amino acid dimers into general form of Chou's pseudo amino acid composition for protein fold recognition," J. Theor. Biol., Vol.380, pp. 291-298, 2015.
[34] A. Dehzangi, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "A mixture of physicochemical and evolutionaryCbased feature extraction approaches for protein fold recognition," Int. J. Data Min. Bioinform., Vol.11, pp. 115-138, 2015.
[35] K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "Improving protein fold recognition using the amalgamation of evolutionary-based and structural based information," BMC Bioinform., Vol.15, pp. S12, 2014.
[36] J. Lyons, N. Biswas, A. Sharma, A. Dehzangi, and K. Paliwal, "Protein fold recognition by alignment of amino acid residues using kernelized dynamic time warping," J. Theor. Biol., Vol.354, pp. 137-145, 2014.
[37] K. Paliwal, A. Sharma, J. Lyons, and A. Dehzangi, "A tri-gram based feature extraction technique using linear probabilities of position specific scoring matrix for protein fold recognition," IEEE Trans. Nanobioscience, Vol.13, pp. 44-50, 2014.
[38] A. Sharma, A. Dehzangi, J. Lyons, S. Imoto, S. Miyano, K. Nakai, and A. Patil, "Evaluation of sequence features from intrinsically disordered regions for the estimation of protein function," PLoS One, Vol.9, pp. e89890, 2014.
[39] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "A segmentation-based method to extract structural and evolutionary features for protein fold recognition," IEEE Trans. on Computational Biology and Bioinformatics, Vol.11, pp. 510-519, 2014.
[40] A. Sharma, J. Lyons, A. Dehzangi, and K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," J. Theor. Biol., Vol.320, pp. 41-46, 2013.
[41] C. D. Huang and C. T. Lin, "Hierarchical learning architecture with automatic feature selection for multiclass protein fold classification," IEEE Trans. on Nanobioscience, Vol.2, pp. 221-232, 2003.
[42] E. Faraggi, T. Zhang, Y. Yang, L. Kurgan, and Y. Zhou, "SPINE-X: improving protein secondary structure prediction by multistep learning coupled with prediction of solvent accessible surface area and backbone torsion angles," J. Comput. Chem., Vol.33, pp. 259-267, 2012.
[43] R. Heffeman, K. Paliwal, J. Lyons, A. Dehzangi, A. Sharma, J. Wang, A. Sattar, Y. Yang, and Y. Zhou, "Improving prediction of secondary structure, local backbone angles, and solvent accessible surface area of proteins by iterative deep learning," Sci. Rep., Vol.5, pp. 11476, 2015.
[44] Y. Saeys, I. Inza, and P. Larranaga, "A review of feature selection techniques in bioinformatics," Bioinformatics, Vol.23, pp. 2507-2517, 2007.
[45] A. Sharma, "A top-r feature selection algorithm for microarray gene expression data," IEEE Trans. on Computational Biology and Bioinformatics, Vol.9, pp. 754-764, 2012.
[46] A. Sharma, S. Imoto, S. Miyano, and V. Sharma, "Null space based feature selection method for gene expression data," Int. J. of Machine Learning and Cybernetics, Vol.3, pp. 269-276, 2012.
[47] M. A. Hall, "Correlation-based feature selection for machine learning," Ph.D. Thesis, The University of Waikato, pp. 51-74, 1999.
[48] A. Ahmadi Adl, A. Nowzari-Dalini, B. Xue, V.N. Uversky, and X. Qian, "Accurate prediction of protein structural classes using functional domains and predicted secondary structure sequences," J. Biomol. Struct. Dyn., Vol.29, pp. 1127-1137, 2012.
[49] C. Cortes and V. Vapnik, "Support-vector networks," Machine Learning, Vol.20, pp. 273-297, 1995.
[50] K. C. Chou and Y. D. Cai, "Using functional domain composition and support vector machines for prediction of protein subcellular location," J. Biol. Chem., Vol.277, pp. 45765-45769, 2002.

[51] Y. D. Cai, G. P. Zhou, and K. C. Chou, "Support vector machines for predicting membrane protein types by using functional domain composition," Biophys. J., Vol.84, pp. 3257-3263, 2003.
[52] P. M. Feng, W. Chen, H. Lin, and K. C. Chou, "iHSP-PseRAAAC: identifying the heat shock protein families using pseudo reduce damino acid alphabet composition," Anal. Biochem., Vol.442, pp. 118-125, 2013.
[53] W. Chen, P. M. Feng, H. Lin, and K. C. Chou, "iRSpot-PseDNC: identify recombination spots with pseudo dinucleotide composition," Nucl. Acids Res., Vol.41, pp. e69, 2013.
[54] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," ACM Trans. Intell. Syst. Technol., Vol.2, pp. 1-27, 2011, software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm [Accessed April 3, 2015].
[55] K. C. Chou and C. T. Zhang, "Prediction of protein structural classes," Crit. Rev. Biochem. Mol. Biol., Vol.30, pp. 275-349, 1995.
[56] Z. X. Wang and Z. Yuan, "How good is prediction of protein structural class by the component-coupled method?" Proteins, Vol.38, pp. 165-175, 2000.
[57] T. Liu, X. Zheng, and J. Wang, "Prediction of protein structural class using a complexity-based distance measure," Amino Acids, Vol.38, pp. 721-728, 2010.
[58] Y. Cai, K. Feng, W. Lu, and K. C. Chou, "Using LogitBoost classifier to predict protein structural classes," J. Theor. Biol., Vol.238, pp. 172-176, 2006.
[59] L. Dong, Y. Yuan, and T. Cai, "Using Bagging classifier to predict protein domain structural class," J. Biomol. Struct. Dyn., Vol.24, pp. 239-242, 2006.
[60] K. C. Chou and H. B. Shen, "Review: recent advances in developing web-servers for predicting protein attributes," Natural Science, Vol.2, pp. 63-92, 2009.

Name:
Liang Kong

Affiliation:
Associate Professor, School of Mathematics and Information Science & Technology, Hebei Normal University of Science & Technology

Address:
No.360, West Hebei Street, Qinhuangdao City, Hebei Province, China

Brief Biographical History:
2004 Received the B.S. degree, Ocean University of China
2007 Received the M.S. degree, Ocean University of China
2010- Ph.D. Student at Yanshan University

Main Works:
• "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., Vol.344, pp. 12-18, 2014.
• "Novel structure-driven features for accurate prediction of protein structural class," Genomics, Vol.103, No.4, pp. 292-297, 2014.

Name:
Lingfu Kong

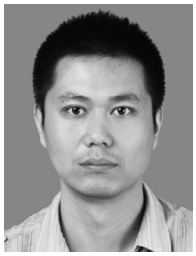
Affiliation:
Professor, School of Information Science and Engineering, Yanshan University

Address:
No.438, West Hebei Street, Qinhuangdao City, Hebei Province, China

Brief Biographical History:
1986 Received the M.S. degree from Northeast Heavy Machinery Institute
1995 Received the Ph.D. degree from Harbin Institute of Technology

Main Works:
• "Salient region detection with hierarchical image abstraction," J. of Information Science and Engineering, Vol.31, pp. 861-878, 2015.
• "Salient region detection an integration approach based on image pyramid and region property," IET Computer Vision, Vol.9, pp. 85-97, 2015.

Membership in Academic Societies:
• Senior Member, China Computer Federation
• Senior Member, Chinese Institute of Electronics



Name:
Rong Jing

Affiliation:
College of Information Science and Engineering, Yanshan University

Address:
No.438, West Hebei Street, Qinhuangdao City, Hebei Province, China

Brief Biographical History:
2004 Received the B.S. degree, Liaoning University of Technology
2007 Received the M.S. degree, Yanshan University
2010- Ph.D. Student at Yanshan University

Main Works:
• "Boundary detection method for large-scale coverage holes in wireless sensor network based on minimum critical threshold constraint," J. of Sensors, 2014.