Hindawi Publishing Corporation
BioMed Research International
Volume 2016, Article ID 5813645, 11 pages
http://dx.doi.org/10.1155/2016/5813645

Research Article

Protein Remote Homology Detection Based on an Ensemble Learning Approach

Junjie Chen,[1] Bingquan Liu,[2] and Dong Huang[1,3]

[1]*School of Computer Science and Technology, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*
[2]*School of Computer Science and Technology, Harbin Institute of Technology, Harbin, Heilongjiang 150001, China*
[3]*Key Laboratory of Network Oriented Intelligent Computation, Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, Guangdong 518055, China*

Correspondence should be addressed to Bingquan Liu; liubq@hit.edu.cn

Received 29 January 2016; Accepted 21 February 2016

Academic Editor: Xun Lan

Protein remote homology detection is one of the central problems in bioinformatics. Although some computational methods have been proposed, the problem is still far from being solved. In this paper, an ensemble classifier for protein remote homology detection, called SVM-Ensemble, was proposed with a weighted voting strategy. SVM-Ensemble combined three basic classifiers based on different feature spaces, including Kmer, ACC, and SC-PseAAC. These features consider the characteristics of proteins from various perspectives, incorporating both the sequence composition and the sequence-order information along the protein sequences. Experimental results on a widely used benchmark dataset showed that the proposed SVM-Ensemble can obviously improve the predictive performance for the protein remote homology detection. Moreover, it achieved the best performance and outperformed other state-of-the-art methods.

1. Introduction

In computational biology, protein remote homology detection is the classification of proteins into structural and functional classes given their amino acid sequences, especially, with low sequence identities. Protein remote homology detection is a critical step for basic research and practical application, which can be applied to the protein 3D structure and function prediction [1, 2]. Although remote homology proteins have similar structures and functions, they lack easily detectable sequence similarities, because the protein structures are more conserved than protein sequences. When the protein sequence similarity is below 35% at the amino acid level, the alignment score usually falls into a twilight zone [3, 4]. Therefore, it is often a failure to detect protein remote homology by computational approaches only based on protein sequence features. To improve the specificity and sensitivity of the detection, we proposed an ensemble learning method, which can combine basic classifiers based on different feature spaces.

Up to now, many methods for protein remote homology detection have been proposed, which can be categorized into three groups [5]: pairwise alignment algorithms, generative models, and discriminative classifiers. Early computational approaches for protein remote homology detection are pairwise alignment methods, which detect sequence similarities between any given two protein sequences by using Needleman-Wunsch global alignment algorithm [6, 7] and Smith-Waterman local alignment algorithm [8]. Later, some trade-off methods were proposed so as to trade reduced accuracy for improved efficiency, such as BLAST [9] and FASTA [10]. PSI-BLAST [11] iteratively builds a probabilistic profile of a query sequence and therefore a more sensitive sequence comparison score can be calculated [12]. After pairwise alignment methods, the predictive accuracy was significantly improved by using the generative algorithms. Generative models were iteratively trained by using positive samples of a protein family or superfamily; for example, HHblits [13] generates a profile hidden Markov model (profile-HMM) [14,

15] from the query sequence and iteratively searches through a large database.

Currently the discriminative methods achieve the state-of-the-art performance [16–19]. Different from pairwise algorithm and generative methods, the discriminative methods can easily embed various characteristics of protein sequences and learn the information from both positive and negative samples in a given benchmark dataset. A key feature of discriminative method is that its input requires fixed length feature vectors. Therefore, some researchers proposed various feature vectors for protein representation. Some methods are based on sequence information, physical and chemical properties of proteins [20–22], or secondary structure information [23, 24], such as SVM-DR [25]. Some methods are based on kernel method, such as SVM-Pairwise [5], SVM-LA [26], motif kernel [27], mismatch [28], SW-PSSM [29], and profile kernel [30]. Later, the performance of discriminative approaches is further improved by Top-n-gram, because it can transform protein profiles into pseudo protein sequences, which contain the evolutionary information [31–33].

Although many discriminative methods for protein remote homology detection have been proposed based on various feature extracting techniques, there is no attempt to combine these methods using an ensemble learning method to improve predictive performance. An ensemble classifier [34, 35] is built by combining a set of basic classifiers in weighted voting strategy to give a final determination in classifying a query sample. Ensemble classifiers have achieved great success in many fields, including protein-protein interaction sites [36], protein fold pattern recognition [22, 37], tRNA detection [38, 39], microRNA identification [40–44], DNA binding protein identification [45], and eukaryotic protein subcellular location prediction [46], because they are able to learn a more expressive concept in classification compared to a single classifier and reduce the variance caused by a single classifier.

In this study, inspired by the success of ensemble classifier in the other fields, we proposed an ensemble classifier for protein remote homology detection, called SVM-Ensemble, which combined three state-of-the-art discriminative methods with a weighted voting strategy. The three basic classifiers SVM-Kmer, SVM-ACC, and SVM-SC-PseAAC were constructed with Kmer, auto-cross covariance (ACC), and series correlation pseudo amino acid composition (SC-PseAAC), respectively. Experimental results on a widely used benchmark dataset [5] showed that SVM-Ensemble can obviously improve the predictive performance by combining various features. Moreover, SVM-Ensemble achieved an average ROC score of 0.945, outperforming the other start-of-the-art methods, indicating that it would be a useful computational tool for protein remote homology detection.

2. Materials and Methods

2.1. Benchmark Dataset. A widely used superfamily benchmark [5] was used to evaluate the performance of our method for protein remote homology detection. The classification problem definition and benchmark dataset are available at http://noble.gs.washington.edu/proj/svm-pairwise/. The same dataset has been used in a number of earlier studies [26, 47–50], allowing us to perform direct comparisons to the relative performance.

The benchmark contains 54 families and 4352 proteins, which are derived from the SCOP database with version 1.53 and the similarities between any two sequences are less than E-value of 10^{-25}. Remote homology detection can be treated as a superfamily classification problem. For each family, the proteins within the family were regarded as positive test samples, and the proteins outside the family but within the same superfamily were taken as positive training samples. Negative samples were selected from outside of the fold and split into training and testing sets. This process was repeated until each family had been tested. This yielded 54 families with at least 10 positive training examples and 5 positive test examples.

2.2. Profile-Based Protein Representation. Although some methods have achieved certain degree of success only by using amino acid sequence information, their performance is not satisfying. Recent studies demonstrated that the methods over profile-based protein sequences would show better performance because a profile is richer than an individual sequence as far as the evolutionary information is concerned [50, 53].

The frequency profile $\mathbb{M}$ for protein $\mathbf{P}$ with L amino acids can be represented as

$$\mathbb{M} = \begin{bmatrix} m_{1,1} & m_{1,2} & \cdots & m_{1,L} \\ m_{2,1} & m_{2,2} & \cdots & m_{2,L} \\ \vdots & \vdots & \vdots & \vdots \\ m_{20,1} & m_{20,2} & \cdots & m_{20,L} \end{bmatrix}, \quad (1)$$

where $m_{i,j}$ $(0 \leq m_{i,j} \leq 1)$ is the target frequency which reflects the probability of amino acid i $(i = 1, 2, \ldots, 20)$ occurring at the sequence position j $(j = 1, 2, \ldots, L)$ in protein $\mathbf{P}$ during evolutionary processes. For each column in $\mathbb{M}$, the elements add up to 1. Each column can therefore be regarded as an independent multinomial distribution. The target frequency was calculated from the multiple sequence alignments generated by running PSI-BLAST [11] against the NCBI's NR with default parameters except that the number of iterations was set at 10 in the current study. The details of how to build a frequency profile can be found in [50].

Given the frequency profile $\mathbb{M}$ for protein $\mathbf{P}$, we can find the amino acid with maximum frequency in each column of $\mathbb{M}$. These amino acids are combined to produce the profile-based protein representation. In a frequency profile $\mathbb{M}$, the target frequencies reflect the probabilities of the corresponding amino acids appearing in the specific sequence positions. The higher the frequency is, the more likely the corresponding amino acid occurs. Thus, the produced profile-based protein sequence contains evolutionary information in the frequency profile. We convert the frequency profiles into a series of profile-based proteins. The existing sequence-based methods can therefore be directly performed on the protein representations for further processing.

2.3. Feature Vector Representations for Protein Sequences. In this study, three kinds of features have been employed to construct the SVM-Ensemble predictor, including Kmer, auto-cross covariance (ACC), and series correlation pseudo amino acid composition (SC-PseAAC).

Suppose a protein sequence $\mathbf{P}$ with L amino acid residues can be represented as

$$\mathbf{P} = R_1R_2R_3R_4R_5R_6 \cdots R_L, \tag{2}$$

where R_i represents the amino acid residue at the sequence position i, such that R_1 represents the amino acid residue at the sequence position 1 and R_2 represents the amino acid residue at position 2 and so on. The three used representation methods can be described as follows.

2.3.1. Kmer. Kmer [56] is the simplest approach to represent the proteins, in which the protein sequences are represented as the occurrence frequencies of k neighboring amino acids.

2.3.2. Auto-Cross Covariance (ACC). ACC transformation [60–62] is to build two signal sequences and then calculate the correlation between them. ACC results in two kinds of variables: autocovariance (AC) transformation and cross covariance (CC) transformation. AC variable measures the correlation of the same property between two residues separated by a distance of lag along the sequence. CC variable measures the correlation of two different properties between two residues separated by lag along the sequence.

Autocovariance (AC) Transformation. Given a protein sequence $\mathbf{P}$ in (2), the AC variable can be calculated by

$$\mathrm{AC}(u, \mathrm{lag}) = \sum_{i=1}^{L-\mathrm{lag}} \frac{\left(P_u(R_i) - \overline{P}_u\right)\left(P_u\left(R_{i+\mathrm{lag}}\right) - \overline{P}_u\right)}{L - \mathrm{lag}}, \tag{3}$$

where u is a physicochemical index, L is the length of the protein sequence, $P_u(R_i)$ means the numerical value of the physicochemical index u for the amino acid R_i, and $\overline{P}_u$ is the average value for physicochemical index u along the whole sequence:

$$\overline{P}_u = \sum_{j=1}^{L} \frac{P_u\left(R_j\right)}{L}. \tag{4}$$

In such a way, the length of AC feature vector is $N * \mathrm{LAG}$, where N is the number of physicochemical indices. LAG is the maximum of lag (lag $= 1, 2, \ldots,$ LAG).

Cross Covariance (CC) Transformation. Given a protein sequence $\mathbf{P}$ in (2), the CC variable can be calculated by

$$\mathrm{CC}(u_1, u_2, \mathrm{lag}) = \sum_{i=1}^{L-\mathrm{lag}} \frac{\left(P_{u_1}(R_i) - \overline{P}_{u_1}\right)\left(P_{u_2}\left(R_{i+\mathrm{lag}}\right) - \overline{P}_{u_2}\right)}{L - \mathrm{lag}}, \tag{5}$$

where u_1, u_2 are two different physicochemical indices, L is the length of the protein sequence, and $P_{u_1}(R_i)$, $P_{u_2}(R_{i+\mathrm{lag}})$ are the numerical value of the physicochemical indices u_1, u_2 for the amino acids R_i, $R_{i+\mathrm{lag}}$. $\overline{P}_{u_1}$, $\overline{P}_{u_2}$ are the average value for physicochemical index values u_1, u_2 along the whole sequence and they can be calculated by (4).

In such way, the length of the CC feature vector is $N * (N - 1) * \mathrm{LAG}$, where N is the number of physicochemical indices. LAG is the maximum of lag (lag $= 1, 2, \ldots,$ LAG).

Therefore, the length of the ACC feature vector is $N * N * \mathrm{LAG}$. In current implementation, three physicochemical properties were employed, including hydrophobicity, hydrophilicity, and mass (see Table S1 in Supplementary file, available online at http://dx.doi.org/10.1155/2016/5813645) extracted from AAindex [57, 63].

2.3.3. Series Correlation Pseudo Amino Acid Composition (SC-PseAAC). SC-PseAAC [64] is an approach incorporating the contiguous local sequence-order information and the global sequence-order information into the feature vector of the protein sequence. Given a protein sequence $\mathbf{P}$ in (2), the SC-PseAAC [64] feature vector of $\mathbf{P}$ is defined:

$$\mathbf{P} = \left[x_1 \;\; x_2 \;\; \cdots \;\; x_{20} \;\; x_{20+1} \;\; x_{20+2} \;\; \cdots \;\; x_{20+\lambda} \;\; x_{20+\lambda+1} \;\; x_{20+\lambda+2} \;\; \cdots \;\; x_{20+3\lambda}\right]^T, \tag{6}$$

where

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j} & (1 \le u \le 20) \\ \dfrac{w\tau_{u-20}}{\sum_{i=1}^{20} f_i + w\sum_{j=1}^{2\lambda} \tau_j} & (20 + 1 \le u \le 20 + 3\lambda), \end{cases} \tag{7}$$

where f_i ($i = 1, 2, \ldots, 20$) is the normalized occurrence frequency of the 20 native amino acids in the protein $\mathbf{P}$; the parameter λ is an integer, representing the highest counted rank (or tier) of the correlation along a protein sequence; w is the weight factor ranging from 0 to 1; and τ_j is the j-tier sequence-correlation factor that reflects the sequence-order correlation between all of the most contiguous residues along a protein sequence, which is defined as

$$\tau_1 = \frac{1}{L-1}\sum_{i=1}^{L-1} H^1_{i,i+1}$$

$$\tau_2 = \frac{1}{L-1}\sum_{i=1}^{L-1} H^2_{i,i+1}$$

$$\tau_3 = \frac{1}{L-1}\sum_{i=1}^{L-1} M_{i,i+1}$$

$$\vdots$$

$$\tau_{3\lambda-2} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{1}$$

$$\tau_{3\lambda-1} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} H_{i,i+\lambda}^{2}$$

$$\tau_{3\lambda} = \frac{1}{L-\lambda}\sum_{i=1}^{L-\lambda} M_{i,i+\lambda} \quad (\lambda < L-1), \tag{8}$$

where $H_{i,j}^{1}$, $H_{i,j}^{2}$, and $M_{i,j}$ are the hydrophobicity, hydrophilicity, and mass correlation functions given by

$$H_{i,j}^{1} = \widehat{h}^{1}(R_i) \cdot \widehat{h}^{1}(R_j)$$

$$H_{i,j}^{2} = \widehat{h}^{2}(R_i) \cdot \widehat{h}^{2}(R_j) \tag{9}$$

$$M_{i,j} = \widehat{m}(R_i) \cdot \widehat{m}(R_j),$$

where $\widehat{h}^{1}(R_i)$, $\widehat{h}^{2}(R_i)$, and $\widehat{m}(R_i)$ are the substituting values of hydrophobicity, hydrophilicity, and mass values for amino acid R_i. They are all subjected to a standard conversion as described by the following equation:

$$\widehat{h}^{1}(R_i) = \frac{h^{1}(R_i) - \sum_{k=1}^{20}\left(h^{1}(\mathbb{R}_k)/20\right)}{\sqrt{\sum_{u=1}^{20}\left[h^{1}(\mathbb{R}_u) - \sum_{k=1}^{20}\left(h^{1}(\mathbb{R}_k)/20\right)\right]^{2}/20}}$$

$$\widehat{h}^{2}(R_i) = \frac{h^{2}(R_i) - \sum_{k=1}^{20}\left(h^{2}(\mathbb{R}_k)/20\right)}{\sqrt{\sum_{u=1}^{20}\left[h^{2}(\mathbb{R}_u) - \sum_{k=1}^{20}\left(h^{2}(\mathbb{R}_k)/20\right)\right]^{2}/20}} \tag{10}$$

$$\widehat{m}(R_i) = \frac{m(R_i) - \sum_{k=1}^{20}\left(m(\mathbb{R}_k)/20\right)}{\sqrt{\sum_{u=1}^{20}\left[m(\mathbb{R}_u) - \sum_{k=1}^{20}\left(m(\mathbb{R}_k)/20\right)\right]^{2}/20}},$$

where we use $\mathbb{R}_i$ $(i = 1, 2, \ldots, 20)$ to represent the 20 native amino acids. The symbols $h^{1}(R_i)$, $h^{2}(R_i)$, and $m(R_i)$ represent the original hydrophobicity, hydrophilicity, and mass values (see Table S1 in Supplementary file) of the amino acid R_i.

These aforementioned features can be generated by a web-server called Pse-in-one [56], which can be used to generate the desired feature vectors for protein/peptide and DNA/RNA sequences according to the need of user's studies. It covers a total of 28 different modes, of which 14 are for DNA sequences, 6 are for RNA sequences, and 8 are for protein sequences.

2.4. Support Vector Machine. Support vector machine (SVM) is a supervised machine learning technique for classification task based on statistical theory [65, 66]. Given a set of fixed length training vectors with labels (positive and negative input samples), SVM can learn a linear decision boundary to discriminate the two classes. The result is a linear classification rule that can be used to classify new test samples. When the samples are linearly nonseparable, the kernel function can be used to map the samples to a high-order feature space in which the optimal hyper plane as decision boundary can be found. SVM has exhibited excellent performance in practice [54, 58, 67–73] and has a strong theoretical foundation of statistical learning.

In this study, the publicly available Gist SVM package (http://www.chibi.ubc.ca/gist/) is employed. The SVM parameters are used by default of the Gist Package except that the kernel function is set as radial basis function.

2.5. Ensemble Classifier. The ensemble classifier is able to learn a more expressive concept in classification compared to a single classifier and reduces the variance caused by a single classifier. Therefore, it was employed in many fields and achieved great success [36, 37].

In this paper, we proposed a weighted voting strategy for protein remote homology detection, as shown in Figure 1. The ensemble framework of SVM-Ensemble was constructed by combining SVM-Kmer, SVM-ACC, and SVM-SC-PseAAC with weighted factors. The processing can be formulated as below.

Suppose the ensemble classifier is expressed by

$$\mathbb{C} = \max\left\{\mathbb{C}_{S_1}, \mathbb{C}_{S_2}, \ldots, \mathbb{C}_{S_{54}}\right\} \tag{11}$$

$$\mathbb{C}_{S_j} = \mathbb{C}_{1S_j} \oplus \mathbb{C}_{2S_j} \oplus \mathbb{C}_{3S_j}, \tag{12}$$

where $\mathbb{C}_{iS_j}$ represents the ith basic SVM classifier on superfamily S_j $(1 \le j \le 54)$. That is, $\mathbb{C}_{1S_1}$ represents the classifier SVM-Kmer that operates on the superfamily S_1, $\mathbb{C}_{2S_1}$ represents the classifier SVM-ACC that operates on superfamily S_1, and $\mathbb{C}_{3S_1}$ represents the classifier SVM-SC-PseAAC that operates on superfamily S_1. $\mathbb{C}_{S_j}$ is the average performance of three basic classifiers on superfamily S_j with weighted voting strategy. In (12), the symbol $\oplus$ denotes the weighted voting operator.

The three basic classifiers can be combined by using the following equation:

$$\mathbb{C}_{S_j} = \sum_{i=1}^{3} w_{iS_j}\mathbb{C}_{iS_j}\left(\mathbf{P}, S_j\right) \quad \left(1 \le j \le 54\right), \tag{13}$$

where $\mathbb{C}_{iS_j}(\mathbf{P}, S_j)$ is the belief function or supporting degree for $\mathbf{P}$ belonging to S_j predicted by the ith basic classifier and w_{iS_j} is the weighted factor assigned with the average ROC score of the ith basic classifier on superfamily S_j.

2.6. Performance Metrics for Evaluation. We evaluated the performance of different methods by employing the receiver operating characteristic (ROC) scores [55, 74–78]. Because

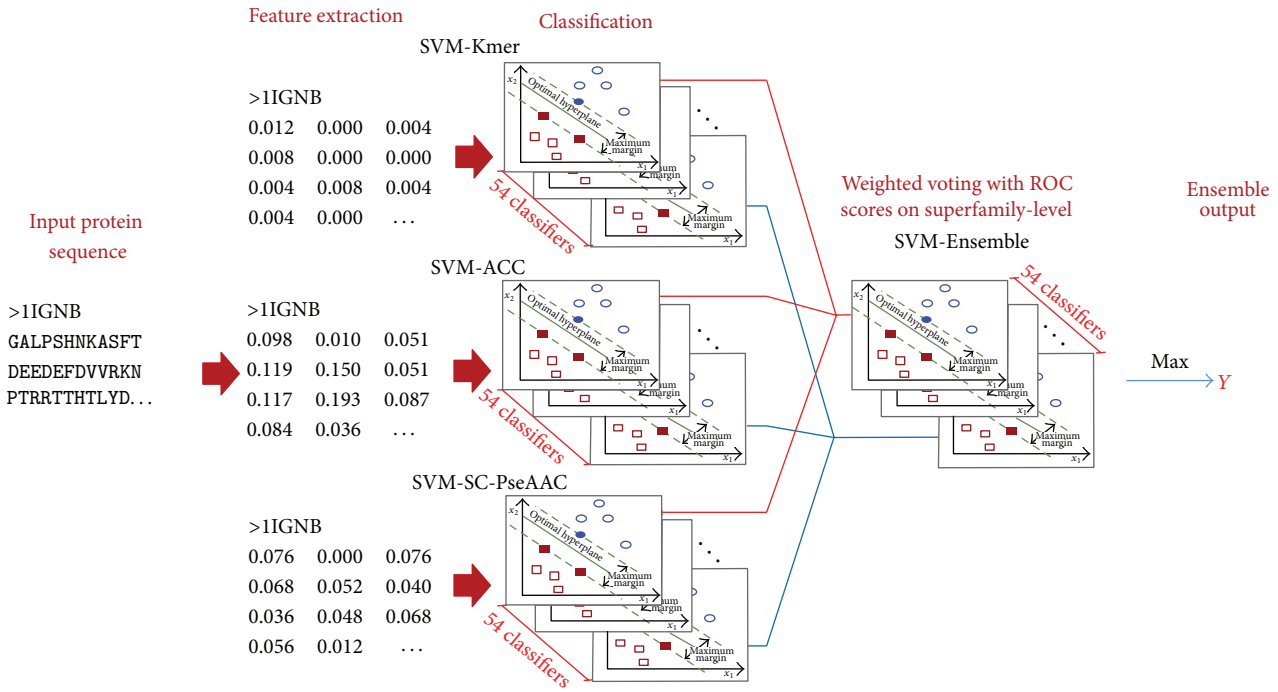

Figure 1: Flowchart to show how the ensemble classifier is formed by combining three basic classifiers on superfamily-level. The ensemble strategy is first employed on superfamily-level, and then the query protein **P** is predicted belonging to the superfamily type with which its score is the highest.

the test sets have more negative than positive samples, simply measuring error-rates will not give a good evaluation of the performance. For the case, the best way to evaluate the trade-off between the specificity and sensitivity is to use ROC score. ROC score is the normalized area under a curve that is plotted with true positives as a function of false positives for varying classification thresholds. ROC score of 1 indicates a perfect separation of positive samples from negative samples, whereas ROC score of 0.5 denotes that random separation. ROC50 score is the area under the ROC curve up to the first 50 false positives.

3. Results and Discussion

3.1. The Influence of Parameters on the Predictive Performance of Basic Predictors. There are several parameters for each basic predictor, which should be optimized. For more information of these parameters, please refer to Materials and Method. In this study, we optimized them by using grid search. The influence of these parameters on the performance was shown in Figure 2, and the optimized values of the parameters and their results were shown in Table 1, from which we can see that SVM-Kmer achieved the best performance, followed by SVM-SC-PseAAC.

3.2. Performance of Ensemble Classifier Based on Various Feature Combinations with Weighted Voting Strategy. As discussed above, predictors based on different feature sets showed different performance. In this study, in order to further improve the performance of protein remote homology detection, we employed an ensemble learning approach to combine various predictors. The performance of ensemble classifier combined various feature combinations was shown in Table 2. The best performance (ROC = 0.943, ROC50 = 0.744) can be achieved with the combination of all the three basic predictors and obviously outperformed all the three basic predictors in terms of both ROC score and ROC50 score. These results were not surprising. The three basic predictors were based on different features, and their predictive results are complementary. The performance can be improved by combining them with an ensemble learning method.

Table 1: The performance of three basic predictors with optimal parameters on benchmark dataset.

Methods	Optimal parameters	ROC[a]	ROC50[a]
SVM-Kmer	$k = 2$	0.912	0.785
SVM-ACC	LAG = 14	0.787	0.483
SVM-SC-PseAAC	$\lambda = 5, w = 0.2$	0.911	0.657

[a] Average ROC and ROC50 scores.

3.3. Feature Analysis for Discriminative Power. To further study the discriminative power of features in the three basic predictors, we employed a feature extraction method, called principal component analysis (PCA) [79], to calculate the discriminative weight vectors in the feature space. The process of PCA for extracting significant features can be found in [32, 80].

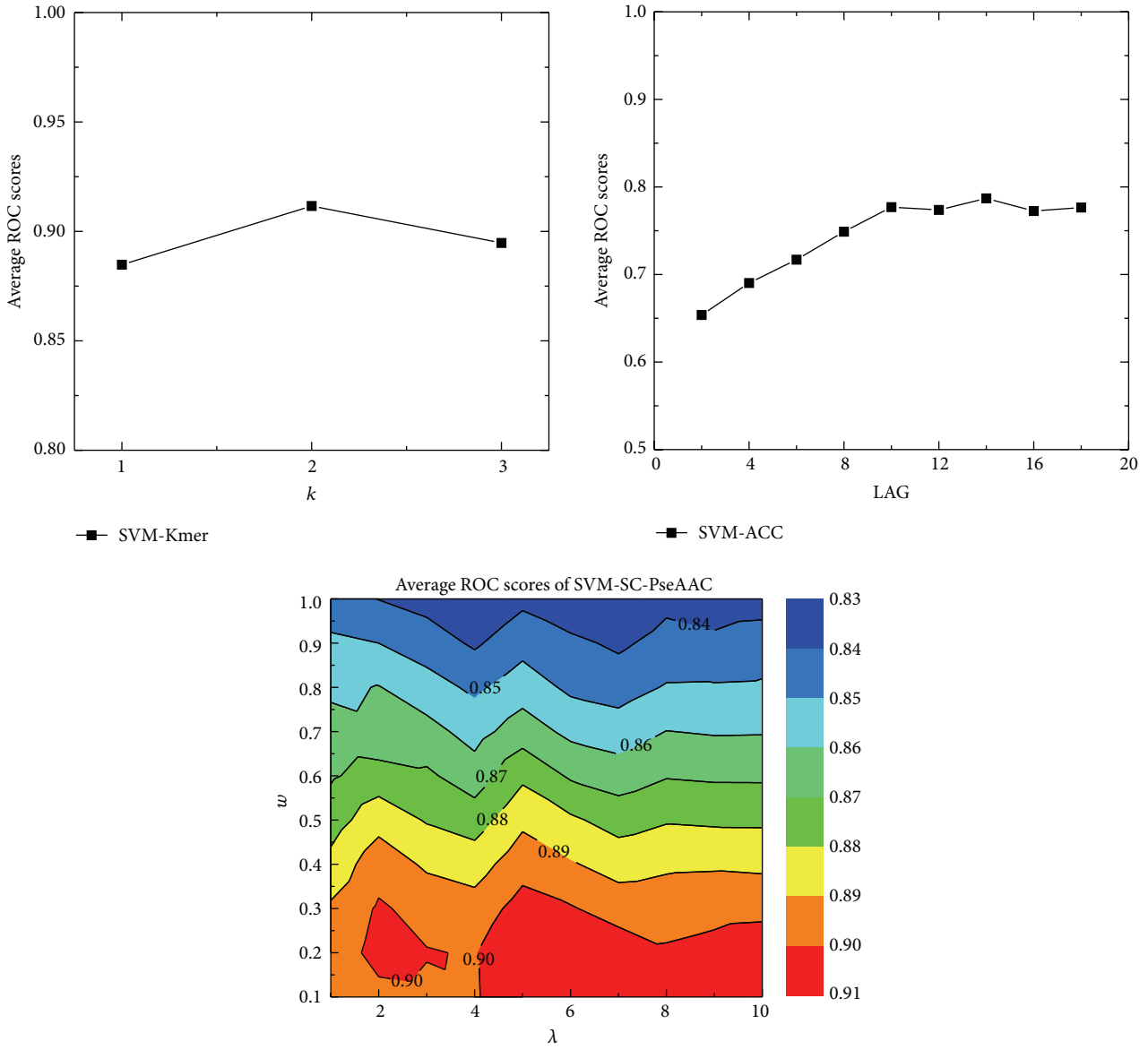

Figure 2: The performance of three basic predictors with all parameter combinations. k value of 2 and the LAG value of 14 were used in SVM-Kmer and SVM-ACC. SVM-SC-PseAAC achieves the best performance with $\lambda = 5$ and $w = 0.2$. Parameter w is mainly impact factor. However, parameter λ has minor impact on the performance.

For each basic predictor, the top 10 most discriminative features in the feature space were shown in Table 3, from which we can see that, for the Kmer features, six of the most discriminative features contain the amino acid M, indicating the importance of this amino acid. For ACC features, the hydrophobicity (h^1) has important impact on the feature discrimination. For SC-PseAAC features, the amino acid M has the most discriminative power and features with small λ value are more important. Both ACC and SC-PseAAC features with strong discriminative power incorporate the sequence-order effects. These three kinds of features consider both sequence composition and sequence order effects. Therefore, SVM-Ensemble can further improve the performance by combining them in an ensemble learning approach.

3.4. Comparison with Other Related Predictors. Some state-of-the-art methods for protein remote homology detection were selected to compare with the proposed SVM-Ensemble. SVM-Pairwise [5] represents each protein as a vector of pairwise similarities to all proteins in the training set. The kernel of SVM-LA [26] measures the similarity between a pair of proteins by taking into account all the optimal local alignment scores with gaps between all possible subsequences. Mismatch kernel [28] is calculated based on occurrences of (k, m)-patterns in the data. Monomer-dist [47] constructs the feature vectors by the occurrences of short oligomers. SVM-DR is based on the distance-pairs; PseAACIndex is based on the pseudo amino acid composition (PseAAC). disPseAAC constructs the feature vectors by combining the

Table 2: Performance of ensemble classifier combining various predictors with weighted voting. The best performance was achieved by combining SVM-Kmer, SVM-ACC, and SVM-SC-PseAAC. The symbol ⊕ denotes the weighted voting operator.

Ensemble methods with superfamily-level strategy	ROC[a]	ROC50[a]
SVM-Kmer ⊕ SVM-ACC	0.929	0.767
SVM-Kmer ⊕ SVM-SC-PseAAC	0.937	0.715
SVM-ACC ⊕ SVM-SC-PseAAC	0.922	0.691
SVM-Kmer ⊕ SVM-ACC ⊕ SVM-SC-PseAAC	**0.943**	**0.744**

[a] Average ROC and ROC50 scores.

Table 3: Top 10 most discriminative features in three feature spaces. These features describe the characteristics of proteins from various perspectives.

Rank	Kmer	ACC	SC-PseAAC
1	MH	$CC_{h^1h^2,\text{lag}=9}$	M
2	WC	$AC_{h^1,\text{lag}=5}$	Y
3	IM	$CC_{h^1h^2,\text{lag}=8}$	$\tau_{h^2,\lambda=1}$
4	MC	$AC_{h^1,\text{lag}=4}$	$\tau_{h^2,\lambda=4}$
5	MY	$CC_{h^1h^2,\text{lag}=7}$	H
6	VM	$AC_{h^1,\text{lag}=14}$	$\tau_{h^1,\lambda=4}$
7	YW	$AC_{m,\text{lag}=13}$	G
8	YR	$CC_{h^1m,\text{lag}=13}$	$\tau_{h^1,\lambda=1}$
9	HW	$CC_{h^1h^2,\text{lag}=10}$	$\tau_{m,\lambda=1}$
10	MQ	$AC_{h^1,\text{lag}=8}$	$\tau_{m,\lambda=3}$

Note: the subscript indexes in ACC features and SC-PseAAC features mean hydrophobicity (h^1), hydrophilicity (h^2), and mass (m).

Table 4: Performance comparison of different methods on the benchmark dataset.

Methods	ROC[a]	ROC50[a]	Source
SVM-Ensemble	0.943	0.744	This study
SVM-Pairwise	0.896	0.464	Liao and Noble, 2003 [5]
SVM-LA ($\beta = 0.5$)	0.925	0.649	Saigo et al., 2004 [26]
Mismatch	0.925	0.649	Leslie et al., 2004 [28]
Monomer-dist	0.919	0.508	Lingner and Meinicke, 2006 [47]
SVM-WCM	0.904	0.445	Lingner and Meinicke, 2008 [51]
SVM-Ngram-LSA	0.859	0.628	Dong et al., 2006 [48]
SVM-Pattern-LSA	0.879	0.626	Dong et al., 2006 [48]
SVM-Motif-LSA	0.859	0.628	Dong et al., 2006 [48]
SVM-Top-n-gram-combine-LSA	0.939	0.767	Liu et al., 2008 [4]
PseAACIndex ($\lambda = 5$)	0.880	0.620	Liu et al., 2013 [31, 52]
PseAACIndex-Profile ($\lambda = 5$)	0.922	0.712	Liu et al., 2013 [31, 52]
SVM-DR	0.919	0.715	Liu et al., 2014 [50, 53–55]
disPseAAC	0.922	0.721	Liu et al., 2015 [2, 32, 44, 45, 56–59]

[a] Average ROC and ROC50 scores.

occurrences of amino acid pairs within Chou's pseudo amino acid composition.

Experimental results of various methods on SCOP 1.53 benchmark dataset were shown in Table 4. The SVM-Ensemble achieved the best performance, indicating that it is correct to combine different predictors via an ensemble learning approach.

4. Conclusions

In this study, we have proposed an ensemble classifier for protein remote homology detection, called SVM-Ensemble. It was constructed by combining three basic classifiers with a weighted voting strategy. Experimental results on a widely used benchmark dataset showed that our method achieved

ROC score of 0.943, which is obviously better than the three basic predictors, including SVM-Kmer, SVM-ACC, and SVM-SC-PseAAC. Compared with some other state-of-the-art methods, the SVM-Ensemble achieved the best performance. Furthermore, by analyzing the discriminative power of these features, some interesting patterns were discovered.

For the future work, more effective features and machine learning techniques will be explored. And evolutionary computation [81], the ensemble learning techniques, and neural-like computing models [82–87] would be applied to other bioinformatics problems, such as gene-disease relationship prediction [52, 88–92] and DNA motif identification [59, 93].

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Bingquan Liu conceived of the study and designed the experiments and participated in designing the study, drafting the paper, and performing the statistical analysis. Junjie Chen participated in coding the experiments and drafting the paper. Dong Huang participated in performing the statistical analysis. All authors read and approved the final paper.

Acknowledgments

This work was supported by Development Program of China (863 Program) [2015AA015405], the National Natural Science Foundation of China (nos. 61300112, 61573118, and 61272383), the Scientific Research Foundation for the Returned Overseas Chinese Scholars, State Education Ministry, the Natural Science Foundation of Guangdong Province (2014A030313695), and Shenzhen Foundational Research Funding (Grant no. JCYJ20150626110425228).

References

[1] P. Bork and E. V. Koonin, "Predicting functions from protein sequences—where are the bottlenecks?" *Nature Genetics*, vol. 18, no. 4, pp. 313–318, 1998.

[2] B. Liu, J. Chen, and X. Wang, "Application of learning to rank to protein remote homology detection," *Bioinformatics*, vol. 31, no. 21, pp. 3492–3498, 2015.

[3] B. Rost, "Twilight zone of protein sequence alignments," *Protein Engineering*, vol. 12, no. 2, pp. 85–94, 1999.

[4] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "A discriminative method for protein remote homology detection and fold recognition combining Top-n-grams and latent semantic analysis," *BMC Bioinformatics*, vol. 9, article 510, 2008.

[5] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for detecting remote protein evolutionary and structural relationships," *Journal of Computational Biology*, vol. 10, no. 6, pp. 857–868, 2003.

[6] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[7] Q. Zou, Q. Hu, M. Guo, and G. Wang, "HAlign: fast multiple similar DNA/RNA sequence alignment based on the centre star strategy," *Bioinformatics*, vol. 31, no. 15, pp. 2475–2481, 2015.

[8] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[9] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[10] W. R. Pearson, "Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms," *Genomics*, vol. 11, no. 3, pp. 635–650, 1991.

[11] S. F. Altschul, T. L. Madden, A. A. Schäffer et al., "Gapped BLAST and PSI-BLAST: a new generation of protein database search programs," *Nucleic Acids Research*, vol. 25, no. 17, pp. 3389–3402, 1997.

[12] B. Liu, X. Wang, L. Lin, Q. Dong, and X. Wang, "Exploiting three kinds of interface propensities to identify protein binding sites," *Computational Biology and Chemistry*, vol. 33, no. 4, pp. 303–311, 2009.

[13] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment," *Nature Methods*, vol. 9, no. 2, pp. 173–175, 2012.

[14] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.

[15] K. Karplus, C. Barrett, and R. Hughey, "Hidden Markov models for detecting remote protein homologies," *Bioinformatics*, vol. 14, no. 10, pp. 846–856, 1998.

[16] H. Ding, H. Lin, W. Chen et al., "Prediction of protein structural classes based on feature selection technique," *Interdisciplinary Sciences: Computational Life Sciences*, vol. 6, no. 3, pp. 235–240, 2014.

[17] H. Ding, L. Liu, F.-B. Guo, J. Huang, and H. Lin, "Identify golgi protein types with modified mahalanobis discriminant algorithm and pseudo amino acid composition," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 58–63, 2011.

[18] H. Lin, W. X. Liu, J. He, X. H. Liu, H. Ding, and W. Chen, "Predicting cancerlectins by the optimal g-gap dipeptides," *Scientific Reports*, vol. 5, Article ID 16964, 2015.

[19] B. Liu, X. Wang, Q. Chen, Q. Dong, and X. Lan, "Using amino acid physicochemical distance transformation for fast protein remote homology detection," *PLoS ONE*, vol. 7, no. 9, Article ID e46633, 2012.

[20] X. Zhao, Q. Zou, B. Liu, and X. Liu, "Exploratory predicting protein folding model with random forest and hybrid features," *Current Proteomics*, vol. 11, no. 4, pp. 289–299, 2014.

[21] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "nDNA-prot: Identification of DNA-binding proteins based on unbalanced classification," *BMC Bioinformatics*, vol. 15, article 298, 2014.

[22] C. Lin, Y. Zou, J. Qin et al., "Hierarchical classification of protein folds using a novel ensemble classifier," *PLoS ONE*, vol. 8, no. 2, Article ID e56499, 2013.

[23] L. Wei, M. Liao, X. Gao, and Q. Zou, "An improved protein structural classes prediction method by incorporating both sequence and structure information," *IEEE Transactions on Nanobioscience*, vol. 14, no. 4, pp. 339–349, 2015.

[24] L. Wei, M. Liao, X. Gao, and Q. Zou, "Enhanced protein fold prediction method through a novel feature extraction

technique," *IEEE Transactions on Nanobioscience*, vol. 14, no. 6, pp. 649–659, 2015.

[25] J. Xu, Q. Zou, R. Xu, X. Wang, and Q. Chen, "Using distances between Top-n-gram and residue pairs for protein remote homology detection," *BMC Bioinformatics*, vol. 15, supplement 2, p. S3, 2014.

[26] H. Saigo, J.-P. Vert, N. Ueda, and T. Akutsu, "Protein homology detection using string alignment kernels," *Bioinformatics*, vol. 20, no. 11, pp. 1682–1689, 2004.

[27] A. Ben-Hur and D. Brutlag, "Remote homology detection: a motif based approach," *Bioinformatics*, vol. 19, supplement 1, pp. i26–i33, 2003.

[28] C. S. Leslie, E. Eskin, A. Cohen, J. Weston, and W. S. Noble, "Mismatch string kernels for discriminative protein classification," *Bioinformatics*, vol. 20, no. 4, pp. 467–476, 2004.

[29] H. Rangwala and G. Karypis, "Profile-based direct kernels for remote homology detection and fold recognition," *Bioinformatics*, vol. 21, no. 23, pp. 4239–4247, 2005.

[30] R. Kuang, E. Ie, K. Wang et al., "Profile-based string kernels for remote homology detection and motif extraction," *Journal of Bioinformatics and Computational Biology*, vol. 3, no. 3, pp. 527–550, 2005.

[31] B. Liu, X. Wang, Q. Zou, Q. Dong, and Q. Chen, "Protein remote homology detection by combining chou's pseudo amino acid composition and profile-based protein representation," *Molecular Informatics*, vol. 32, no. 9-10, pp. 775–782, 2013.

[32] B. Liu, J. Chen, and X. Wang, "Protein remote homology detection by combining Chou's distance-pair pseudo amino acid composition and principal component analysis," *Molecular Genetics and Genomics*, vol. 290, no. 5, pp. 1919–1931, 2015.

[33] Y. Zhang, B. Liu, Q. Dong, and V. X. Jin, "An improved profile-level domain linker propensity index for protein domain boundary prediction," *Protein and Peptide Letters*, vol. 18, no. 1, pp. 7–16, 2011.

[34] T. G. Dietterich, "Ensemble methods in machine learning," in *Multiple Classifier Systems*, pp. 1–15, Springer, Berlin, Germany, 2000.

[35] C. Lin, W. Chen, C. Qiu, Y. Wu, S. Krishnan, and Q. Zou, "LibD3C: ensemble classifiers with a clustering and dynamic selection strategy," *Neurocomputing*, vol. 123, pp. 424–435, 2014.

[36] L. Deng, J. Guan, Q. Dong, and S. Zhou, "Prediction of protein-protein interaction sites using an ensemble method," *BMC Bioinformatics*, vol. 10, no. 1, article 426, 2009.

[37] H.-B. Shen and K.-C. Chou, "Ensemble classifier for protein fold pattern recognition," *Bioinformatics*, vol. 22, no. 14, pp. 1717–1722, 2006.

[38] Q. Zou, J. Guo, Y. Ju, M. Wu, X. Zeng, and Z. Hong, "Improving tRNAscan-SE annotation results via ensemble classifiers," *Molecular Informatics*, vol. 34, no. 11-12, pp. 761–770, 2015.

[39] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "repRNA: a web server for generating various feature vectors of RNA sequences," *Molecular Genetics and Genomics*, vol. 291, no. 1, pp. 473–481, 2016.

[40] C. Y. Wang, L. Hu, M. Z. Guo, X. Y. Liu, and Q. Zou, "imDC: an ensemble learning method for imbalanced classification with miRNA data," *Genetics and Molecular Research*, vol. 14, no. 1, pp. 123–133, 2015.

[41] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human microRNAs by incorporatinga high-quality negative set," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 1, pp. 192–201, 2014.

[42] J. Chen, X. Wang, and B. Liu, "iMiRNA-SSF: improving the identification of MicroRNA precursors by combining negative sets with different distributions," *Scientific Reports*, vol. 6, Article ID 19062, 2016.

[43] B. Liu, L. Fang, F. Liu, X. Wang, and K.-C. Chou, "iMiRNA-PseDPC: microRNA precursor identification with a pseudo distance-pair composition approach," *Journal of Biomolecular Structure and Dynamics*, vol. 34, no. 1, pp. 220–232, 2016.

[44] B. Liu, L. Fang, S. Wang, X. Wang, H. Li, and K.-C. Chou, "Identification of microRNA precursor with the degenerate K-tuple or Kmer strategy," *Journal of Theoretical Biology*, vol. 385, pp. 153–159, 2015.

[45] B. Liu, S. Wang, and X. Wang, "DNA binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific Reports*, vol. 5, Article ID 15479, 2015.

[46] L. Li, Y. Zhang, L. Zou et al., "An ensemble classifier for eukaryotic protein subcellular location prediction using gene ontology categories and amino acid hydrophobicity," *PLoS ONE*, vol. 7, no. 1, Article ID e31057, 2012.

[47] T. Lingner and P. Meinicke, "Remote homology detection based on oligomer distances," *Bioinformatics*, vol. 22, no. 18, pp. 2224–2231, 2006.

[48] Q.-W. Dong, X.-L. Wang, and L. Lin, "Application of latent semantic analysis to protein remote homology detection," *Bioinformatics*, vol. 22, no. 3, pp. 285–290, 2006.

[49] L. Liao and W. S. Noble, "Combining pairwise sequence similarity and support vector machines for remote protein homology detection," in *Proceedings of the 6th Annual International Conference on Computational Biology (RECOMB '02)*, pp. 225–232, Washington, DC, USA, April 2002.

[50] B. Liu, D. Zhang, R. Xu et al., "Combining evolutionary information extracted from frequency profiles with sequence-based kernels for protein remote homology detection," *Bioinformatics*, vol. 30, no. 4, pp. 472–479, 2014.

[51] T. Lingner and P. Meinicke, "Word correlation matrices for protein sequence analysis and remote homology detection," *BMC Bioinformatics*, vol. 9, no. 1, article 259, 13 pages, 2008.

[52] B. Liu, J. Yi, A. Sv et al., "QChIPat: a quantitative method to identify distinct binding patterns for two biological ChIP-seq samples in different experimental conditions," *BMC Genomics*, vol. 14, supplement 8, article S3, 2013.

[53] B. Liu, B. Liu, F. Liu, and X. Wang, "Protein binding site prediction by combining hidden markov support vector machine and profile-based propensities," *The Scientific World Journal*, vol. 2014, Article ID 464093, 6 pages, 2014.

[54] W.-X. Liu, E.-Z. Deng, W. Chen, and H. Lin, "Identifying the subfamilies of voltage-gated potassium channels using feature selection technique," *International Journal of Molecular Sciences*, vol. 15, no. 7, pp. 12940–12951, 2014.

[55] B. Liu, J. Xu, X. Lan et al., "IDNA-Prot—dis: identifying DNA-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PLoS ONE*, vol. 9, no. 9, Article ID e106691, 2014.

[56] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K.-C. Chou, "Pse-in-One: a web server for generating various modes of pseudo components of DNA, RNA, and protein sequences," *Nucleic Acids Research*, vol. 43, no. 1, pp. W65–W71, 2015.

[57] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "PseDNA-Pro: DNA-binding protein identification by combining chou's PseAAC and Physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015.

[58] B. Liu, L. Fang, J. Chen, F. Liu, and X. Wang, "MiRNA-dis: MicroRNA precursor identification based on distance structure status pairs," *Molecular BioSystems*, vol. 11, no. 4, pp. 1194–1204, 2015.

[59] B. Liu, F. Liu, L. Fang, X. Wang, and K.-C. Chou, "repDNA: a Python package to generate various modes of feature vectors for DNA sequences by incorporating user-defined physicochemical properties and sequence-order effects," *Bioinformatics*, vol. 31, no. 8, pp. 1307–1309, 2015.

[60] D.-S. Cao, Q.-S. Xu, and Y.-Z. Liang, "Propy: a tool to generate various modes of Chou's PseAAC," *Bioinformatics*, vol. 29, no. 7, pp. 960–962, 2013.

[61] Q. Dong, S. Zhou, and J. Guan, "A new taxonomy-based protein fold recognition approach based on autocross-covariance transformation," *Bioinformatics*, vol. 25, no. 20, pp. 2655–2662, 2009.

[62] X. Liu, L. Zhao, and Q. Dong, "Protein remote homology detection based on auto-cross covariance transformation," *Computers in Biology and Medicine*, vol. 41, no. 8, pp. 640–647, 2011.

[63] S. Kawashima and M. Kanehisa, "AAindex: amino acid index database," *Nucleic Acids Research*, vol. 28, no. 1, p. 374, 2000.

[64] K.-C. Chou, "Using amphiphilic pseudo amino acid composition to predict enzyme subfamily classes," *Bioinformatics*, vol. 21, no. 1, pp. 10–19, 2005.

[65] C.-C. Chang and C.-J. Lin, "LIBSVM: a Library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, no. 3, article 27, 2011.

[66] L. Fang, F. Liu, X. Wang, J. Chen, K.-C. Chou, and B. Liu, "Identification of real microRNA precursors with a pseudo structure status composition approach," *PLoS ONE*, vol. 10, no. 3, Article ID e0121501, 2015.

[67] H. Ding, E.-Z. Deng, L.-F. Yuan et al., "ICTX-type: a sequence-based predictor for identifying the types of conotoxins in targeting ion channels," *BioMed Research International*, vol. 2014, Article ID 286419, 10 pages, 2014.

[68] H. Ding, S.-H. Guo, E.-Z. Deng et al., "Prediction of Golgi-resident protein types by using feature selection technique," *Chemometrics and Intelligent Laboratory Systems*, vol. 124, pp. 9–13, 2013.

[69] S.-H. Guo, E.-Z. Deng, L.-Q. Xu et al., "INuc-PseKNC: a sequence-based predictor for predicting nucleosome positioning in genomes with pseudo k-tuple nucleotide composition," *Bioinformatics*, vol. 30, no. 11, pp. 1522–1529, 2014.

[70] H. Lin, E.-Z. Deng, H. Ding, W. Chen, and K.-C. Chou, "IPro54-PseKNC: a sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition," *Nucleic Acids Research*, vol. 42, no. 21, pp. 12961–12972, 2014.

[71] H. Lin, H. Ding, F.-B. Guo, A.-Y. Zhang, and J. Huang, "Predicting subcellular localization of mycobacterial proteins by using Chou's pseudo amino acid composition," *Protein and Peptide Letters*, vol. 15, no. 7, pp. 739–744, 2008.

[72] L.-F. Yuan, C. Ding, S.-H. Guo, H. Ding, W. Chen, and H. Lin, "Prediction of the types of ion channel-targeted conotoxins based on radial basis function network," *Toxicology in Vitro*, vol. 27, no. 2, pp. 852–856, 2013.

[73] B. Liu, L. Fang, R. Long, X. Lan, and K.-C. Chou, "iEnhancer-2L: a two-layer predictor for identifying enhancers and their strength by pseudo *k*-tuple nucleotide composition," *Bioinformaitcs*, vol. 32, no. 3, pp. 362–369, 2016.

[74] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[75] H. Ding, P.-M. Feng, W. Chen, and H. Lin, "Identification of bacteriophage virion proteins by the ANOVA feature selection and analysis," *Molecular BioSystems*, vol. 10, no. 8, pp. 2229–2235, 2014.

[76] H. Ding, L. Luo, and H. Lin, "Prediction of cell wall lytic enzymes using chou's amphiphilic pseudo amino acid composition," *Protein and Peptide Letters*, vol. 16, no. 4, pp. 351–355, 2009.

[77] B. Liu, X. Wang, L. Lin, B. Tang, Q. Dong, and X. Wang, "Prediction of protein binding sites in protein structures using hidden Markov support vector machine," *BMC Bioinformatics*, vol. 10, article 381, 2009.

[78] B. Liu and L. Fang, "Identification of microRNA precursor based on gapped *n*-tuple structure status composition kernel," *Computational Biology and Chemistry*, 2016.

[79] S. Wold, K. Esbensen, and P. Geladi, "Principal component analysis," *Chemometrics and Intelligent Laboratory Systems*, vol. 2, no. 1–3, pp. 37–52, 1987.

[80] Q.-S. Du, Z.-Q. Jiang, W.-Z. He, D.-P. Li, and K.-C. Chou, "Amino acid principal component analysis (AAPCA) and its applications in protein structural class prediction," *Journal of Biomolecular Structure and Dynamics*, vol. 23, no. 6, pp. 635–640, 2006.

[81] X. Zhang, Y. Tian, and Y. Jin, "A knee point driven evolutionary algorithm for many-objective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 6, pp. 761–776, 2015.

[82] T. Song and L. Pan, "On the universality and non-universality of spiking neural P systems with rules on synapses," *IEEE Transactions on NanoBioscience*, vol. 14, no. 8, pp. 960–966, 2015.

[83] X. Zeng, X. Zhang, T. Song, and L. Pan, "Spiking neural P systems with thresholds," *Neural Computation*, vol. 26, no. 7, pp. 1340–1361, 2014.

[84] X. Chen, M. J. Pérez-Jiménez, L. Valencia-Cabrera, B. Wang, and X. Zeng, "Computing with viruses," *Theoretical Computer Science*, vol. 623, pp. 146–159, 2016.

[85] L. P. Xiangxiang Zeng and M. J. Pérez-Jiménez, "Small universal simple spiking neural P systems with weights," *Science China Information Sciences*, vol. 57, no. 9, pp. 1–11, 2014.

[86] X. Zhang, L. Pan, and A. Păun, "On the universality of axon P systems," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 11, pp. 2816–2829, 2015.

[87] X. Zhang, Y. Liu, B. Luo, and L. Pan, "Computational power of tissue P systems for generating control languages," *Information Sciences*, vol. 278, pp. 285–297, 2014.

[88] X. Zeng, Y. Liao, Y. liu, and Q. Zou, "Prediction and validation of disease genes using HeteSim Scores," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[89] Q. Zou, J. Li, Q. Hong et al., "Prediction of microRNA-disease associations based on social network analysis methods," *BioMed Research International*, vol. 2015, Article ID 810514, 9 pages, 2015.

[90] X. Zeng, X. Zhang, and Q. Zou, "Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks," *Briefings in Bioinformatics*, vol. 17, no. 2, pp. 193–203, 2016.

[91] Q. Zou, J. Li, L. Song, X. Zeng, and G. Wang, "Similarity computation strategies in the microRNA-disease network: a survey," *Briefings in Functional Genomics*, vol. 15, no. 1, pp. 55–64, 2016.

[92] H.-Z. Chen, M. M. Ouseph, J. Li et al., "Canonical and atypical E2Fs regulate the mammalian endocycle," *Nature Cell Biology*, vol. 14, no. 11, pp. 1192–1202, 2012.

[93] X. Wang, Y. Miao, and M. Cheng, "Finding motifs in DNA sequences using low-dispersion sequences," *Journal of Computational Biology*, vol. 21, no. 4, pp. 320–329, 2014.