Hindawi Publishing Corporation
International Journal of Genomics
Volume 2016, Article ID 7604641, 11 pages
http://dx.doi.org/10.1155/2016/7604641

Research Article

Accurate Identification of Cancerlectins through Hybrid Machine Learning Technology

Jieru Zhang,[1,2] Ying Ju,[3] Huijuan Lu,[4] Ping Xuan,[5] and Quan Zou[2,6]

[1]*School of Software, Tianjin University, Tianjin, China*
[2]*School of Computer Science and Technology, Tianjin University, Tianjin, China*
[3]*School of Information Science and Technology, Xiamen University, Xiamen, China*
[4]*College of Information Engineering, China Jiliang University, Hangzhou, Zhejiang, China*
[5]*School of Computer Science and Technology, Heilongjiang University, Harbin, China*
[6]*State Key Laboratory of Medicinal Chemical Biology, NanKai University, Tianjin, China*

Correspondence should be addressed to Ping Xuan; 2004058@hlju.edu.cn and Quan Zou; zouquan@tju.edu.cn

Received 18 April 2016; Revised 24 May 2016; Accepted 14 June 2016

Academic Editor: Qin Ma

Cancerlectins are cancer-related proteins that function as lectins. They have been identified through computational identification techniques, but these techniques have sometimes failed to identify proteins because of sequence diversity among the cancerlectins. Advanced machine learning identification methods, such as support vector machine and basic sequence features (n-gram), have also been used to identify cancerlectins. In this study, various protein fingerprint features and advanced classifiers, including ensemble learning techniques, were utilized to identify this group of proteins. We improved the prediction accuracy of the original feature extraction methods and classification algorithms by more than 10% on average. Our work provides a basis for the computational identification of cancerlectins and reveals the power of hybrid machine learning techniques in computational proteomics.

1. Introduction

Lectins, which can combine with sugars, are proteins that are produced and secreted by animal and plant cells. These proteins play a key role in cell-to-cell recognition and cellular adhesion, especially cellular interactive adhesion, because they contain many carbohydrate-combining sites. Cancerlectins are well-known lectins because of their source, sequences, binding site architecture, quaternary structure, and carbohydrate specificity. They participate in cancer-related processes, such as tumor cell differentiation, cancer cell monitoring, tumor tissue cell marking, and cancer metastasis.

Cancerlectins are typically identified through biological experiments, but these are costly and inefficient. As such, computational prediction approaches have been employed to verify novel cancerlectin protein sequences and to obtain cancerlectin candidates. Prediction accuracy is an important parameter, which when optimized can reduce the cost of computational prediction approaches. However, the accuracy rates of existing calculation and prediction methods are approximately 70%, which is unsatisfactory and thus should be improved. In the current study, we evaluated different feature extraction algorithms and classifiers to establish novel combinatorial machine learning strategy that can improve prediction accuracy.

Machine learning techniques instead of traditional sequence alignment methods, such as PSI-BLAST [1], HMMER [2], and HAlign [3], are often used to identify special proteins. Among these identification techniques, a support vector machine is the most common classifier used in computational proteomics, which involves various processes, such as classifying protein subfamilies [4–6], predicting protein structural classes [7], and identifying thermophilic proteins [8]. Random forest is also a common classifier that works via an ensemble learning strategy and performs well in protein fold recognition [9]. In addition to random forest, heterogeneous basic classifiers are combined to classify imbalances [10] and improve accuracy [11–13]. Bioinspired

computing models and algorithms can also be used to design promising classifiers, such as spiking neural models [14–18] and evolutionary algorithms [19, 20]. All of these advanced machine learning methods have demonstrated satisfactory performance in cancerlectin identification, which has inspired us to combine different classifiers and feature extractors to optimize the accuracy of prediction. After comparing their efficiency and popularity, we chose the feature extraction methods and classification algorithms mentioned above to demonstrate the impact of machine learning on the field of cancerlectin identification.

Protein features are more important than machine learning techniques for achieving the high accuracy of protein prediction. The protein features most commonly used for feature extraction and classification are *k*-mer and Chou's PseACC representation [21, 22]. They perform well in a range of applications, including predicting protein submitochondrial locations [23], identifying Golgi-resident protein types [24], predicting microkit protein localization [25], and identifying bacteriophage virion proteins [26]. Position-specific scoring matrix is another good option, but obtaining it is time-consuming [16], which limits its application. In some instances, an analysis of protein secondary structures helps improve classification accuracy. However, the extraction of secondary structure features is time-consuming. Some studies have reduced the feature dimensions for biological sequences, such as by using the minimum Redundancy Maximum Relevance (mRMR) [27, 28] and Max-Relevance-Max-Distance (MRMD) [29]. Nevertheless, studies have yet to combine hybrid multisource features, which is the main contribution of the current work.

Related machine learning strategies have yet to be applied to distinguish cancerlectins from other lectins. Song and Pan [30] and Kumar et al. [31] employed SVM but obtained only approximately 70% accuracy. They tested basic sequence features and disregarded multiview feature combination. In addition Damodaran et al. [32] collected more than 500 cancerlectins, which are used here as a positive training set for machine learning. In this study, we aim to examine additional features and classifiers and to determine the optimal combination of hybrid machine learning techniques that can be used to achieve optimal accuracy in cancerlectin prediction.

2. Methods

2.1. Main Flow. Machine learning, which can be used in protein mapping, has evolved from computational learning theory and the field of pattern recognition. Algorithms are initially used to extract the features of amino acids; different classifiers are then employed to predict cancerlectins. Various machine learning algorithms, which are more efficient and accurate than traditional methods, such as SVM-Prot-based feature extraction algorithm [33] and libSimpleVote classifier, are also utilized to predict cancerlectins. Therefore, the efficient combination of feature extraction algorithms and classifiers has been extensively investigated.

Although numerous feature extraction algorithms and classifiers have been widely used and studied in the field of bioinformatics and in the computing industry, the combination of these two strategies has rarely been investigated and the development of efficient cancerlectin prediction methods has seldom been performed. Furthermore, the combination of feature extraction and classifiers has been disregarded by most researchers because of the large data requirement and laboriousness of the work.

In the current study, various feature extraction algorithms are investigated and different feature dimensions are combined to determine an accurate feature vector. Feature extraction results are then applied to different classifiers to predict cancerlectins. After performing these trials, the most accurate and efficient combination of feature extraction algorithm and classifier can be determined and the accuracy rate can be calculated. Thus, this study aims to evaluate existing feature extraction methods and to identify the appropriate dimensions that can be used to predict cancerlectins with the highest accuracy. An appropriate classifier is also necessary to predict cancerlectins. Other tools and methods are also utilized to reduce the dimension of feature vectors and to help improve the accuracy of prediction. The following concepts are considered in our study:

(1) Various feature vector files in .arff are calculated on the basis of a specific database (CancerLectinDB), and different dimensions are combined to create .arff files.

(2) Different classifiers are used to predict the mapping of cancerlectin, and different prediction results are compared in one table or graph to determine the most accurate prediction method.

(3) Feature extraction and random forest based on Conjoint Triad and Pseudo-Amino Acid Composition are the most accurate combination of feature extraction algorithms and classifiers to predict cancerlectins.

The main flow process is shown in Figure 1.

2.2. Data Preprocessing. CancerLectinDB, which is from a web server named CalecPred [30] and was provided by Professors Song and Pan, is used in this study to obtain high-quality data regarding cancerlectins. All of the training and the test sets were selected from this server as the data set in this work. Within the data set, 178 cancerlectins and 226 noncancerlectins are used as a training set and 20 other cancerlectins and noncancerlectins are utilized as a test set. In some feature extraction algorithms found in ProtrWeb [34], some cancerlectin and noncancerlectin sequences cannot be included because the protein sequence is too long to fit the methods; as such, we excluded these protein sequences to ensure an appropriate fit with the corresponding feature extraction methods. Table 1 shows the number of lectins used in some feature extraction algorithms in ProtrWeb after the excluded data have been removed.

2.2.1. Sequence Motifs Discovery. In order to clearly visualize the data, MEME [35] was used to analyze the conserved motifs among the cancerlectins. Because there is a limitation in the number of amino acids, we divided the set of cancerlectins into two groups. The five most significant conserved

Table 1: The number of pieces of data used in the ProtrWeb.

	Train set		Test set	
	Cancerlectin	Noncancerlectin	Cancerlectin	Noncancerlectin
Amino Acid Composition	178	226	20	20
Dipeptide Composition	178	226	20	20
Normalized Moreau-Broto Autocorrelation	178	225	20	20
Moran Autocorrelation	178	226	20	20
Geary autocorrelation	178	226	20	20
Conjoint Triad	178	226	20	20
Sequence-Order-Coupling Number	178	225	20	20
Quasi-Sequence-Order Descriptors	178	225	20	20
Pseudo-Amino Acid Composition	178	225	20	20
Amphiphilic Pseudo-Amino Acid Composition	178	225	20	20

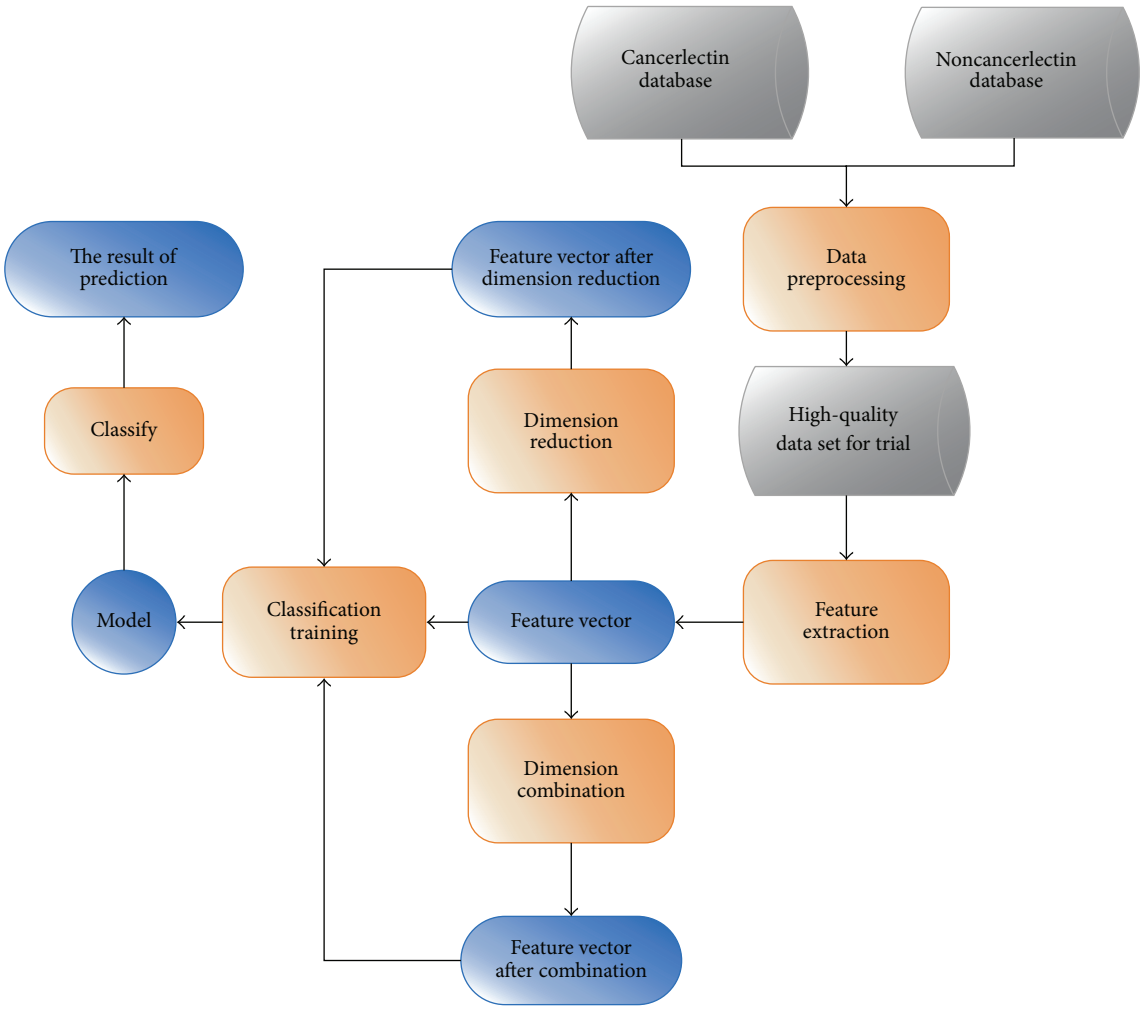


Figure 1: The main flow chart of the identification method of cancerlectin.

motifs of the first group are shown in Figure 5 and Table 5, and the motifs of the second group are shown in Figure 6 and Table 6.

2.2.2. Training Set Balancing. There are 178 positive samples (cancerlectins) and 226 negative samples (noncancerlectins) in the training set. This inconsistency between the two groups could result in inaccurate results. In order to optimize the classification, we use the synthetic minority oversampling technique (SMOTE) [36] algorithm in Weka to supervise the instance. We also apply SMOTE to the training set of two main feature extraction methods: Conjoint Triad and Pseudo-Amino Acid Composition. The numbers of positive and negative samples before and after balancing are shown in Table 7. In addition, the comparisons before and after balancing the training set are shown in Table 8. We can see from Table 8 that, after balancing the positive and negative samples, the accuracy of cross-validation increases, but the accuracy of the method with the supplied test set decreases.

2.3. Feature Extraction Algorithm

2.3.1. Conjoint Triad Feature. Conjoint Triad Feature (CTF) is a feature extraction algorithm used to obtain protein dimensions. It is based on neighbor relationships in protein sequences. This algorithm encodes each protein sequence by using a triad frequency distribution, which is extracted from a seven-letter reduced alphabet. It is also applied to formulate protein samples and perform predictions. CTF clusters 20 amino acids into seven classes [37] and regards any three consecutive amino acids among them as a single unit. A total of 343 dimensions of cancerlectin sequences are extracted by using the CTF algorithm. It transfers the file from .csv format into .arff format. These .arff format files are then placed in some classifiers, such as random forests, for analysis and prediction.

A cancerlectin sequence is represented by C and is composed of L amino acids:

$$C = A_1A_2A_3, \ldots, A_L. \quad (1)$$

We can include three amino acids in one group, as follows: $A_1A_2A_3, A_2A_3A_4, A_3A_4A_5, A_4A_5A_6, \ldots, A_{L-3}A_{L-2}A_{L-1}$, $A_{L-2}A_{L-1}A_L$. The CTF of a cancerlectin is considered as the normalized frequency of these corresponding trimers in a sequence of a cancerlectin and is expressed as follows:

$$\text{CTF} = \left[F_1, F_2, F_3, \ldots, F_k\right]^T, \quad (2)$$

where F_i is the frequency of the three consecutive residues and $k = 7^3 = 343$. Because the 20 kinds of amino acids can be divided into seven classes and we have three amino acids in one unit, for each unit, there can be $7 \times 7 \times 7$ different combinations, so we finally obtain 343 dimensions [38].

2.3.2. Pseudo-Amino Acid Composition. Pseudo-Amino Acid Composition (Pse-AAC) [39] is an approach incorporating contiguous local sequence-order information and global sequence-order information into the feature vector of a protein sequence. This approach can be used to obtain a feature vector with 50 dimensions. After some calculations are performed in ProtrWeb, the feature vector file in .arff can be created. The feature extraction vectors can then be placed in classifiers to obtain prediction results.

C can be further expressed as follows:

$$C = A_1A_2A_3, \ldots, A_L. \quad (3)$$

The Pse-AAC feature of a protein is defined as follows:

$$\text{Pse-AAC} = \left[F_1, F_2, F_3, \ldots, F_k\right]^T, \quad (4)$$

where F_i is the frequency of the amino acid calculated by the Pse-AAC algorithm and $k = 50$.

2.4. Classifier Selection and Tools

2.4.1. Weka and Random Forest. Waikato Environment for Knowledge Analysis (Weka) is a well-known suite of machine learning software, which is used for data analysis and predictive modeling. In this study, Weka is used as a classifier. Among the options of Weka, "Classify" provides different modes of classifiers, such as random forest, ZeroR, KStar, and libSVM.

Random forests are used to obtain the average of multiple deep decision trees and are trained on different parts of the same training set to reduce variances. They are also considered a learning method for certain tasks such as classification and regression. Furthermore, random forests are used as a model for the rapid and efficient method of classification. This model applies bagging but uses a modified tree learning algorithm to select and split candidates during learning. In this method, different decision trees are determined for classification.

Weka also includes other test options, such as supplied test set, cross-validation, and percentage split. In this study, supplied test set and cross-validation are used to perform prediction. In the supplied test, training data and test set data should be provided for prediction. In the cross-validation, a single data set is split into a test data set and a training data set by using a specific algorithm.

2.4.2. libSVM and Grid. libSVM [40] is an open-source machine learning library that implements the SMO algorithm for kernelized support vector machines and supports classification and regression; this library has been widely used to solve many tasks in bioinformatics [41, 42]. To apply this tool in our research, we download and install certain configuration files, especially Python. We execute all commands in a command line based on the runtime system of Python.

In this study, Grid was added to libSVM to tune parameters c and g and to enhance the accuracy of the prediction results. c and g are two training parameters provided by SVM with a Gaussian kernel function. Parameter c controls the overfitting of the model and parameter g controls the degree of nonlinearity of the model. g is inversely related to c, which represents the distribution around the statistical mean. Larger values of c will result in a model with low bias and high variance, and smaller g also corresponds to a model with

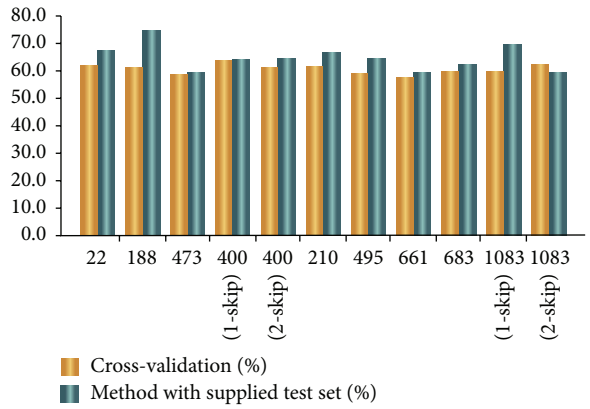


Figure 2: The accuracy rate of prediction in Part I.

low bias and high variance. Thus, the behavior of the kernel is less distributed or more nonlinear. These two parameters are determined by Grid search and cross-validation. The model with the highest estimated performance determines the selected training parameters. Then, these two parameters are used to predict libSVM to establish an SVM model and to obtain a more accurate prediction result. In the following section, the combinations of feature extraction and classifier for which the accuracy rate is >70% are reevaluated in libSVM.

3. Results and Discussion

3.1. Multidimension Combination Prediction. In this section, the feature extraction algorithms excluded from ProtrWeb are mainly investigated. These algorithms are referred to as multiple dimension combination prediction (MDCP) tools because their use involves different feature extraction methods and their combinations to obtain feature vectors and perform prediction. In the feature extraction part, different methods are employed to determine the vectors: 1-skip, 2-skip, 188-dimension feature extraction, 473-dimension feature extraction, and some algorithm combinations. In general, the 188-dimension feature extraction is based on physicochemical characteristics, and the n-skip algorithm is the same as a k-mer algorithm. In the classification part, the supplied test set and the cross-validation set are used for prediction.

After the combination of various dimensions and the conversion of file format, various .arff files with different dimensions are obtained with a specific file head. We place the .arff files into random forest classifiers in Weka for prediction. Table 2 lists the exact dimensions of the algorithms. Figure 2 shows the prediction results based on cross-validation and supplied test set validation. In Figure 2, 188-dimension feature extraction yields the highest accuracy rate of 75% when the supplied test set validation is applied.

3.2. ProtrWeb-Weka Prediction. In this section, the following algorithms provided by ProtrWeb are examined: Amino Acid Composition, Dipeptide Composition, Normalized Moreau-Broto Autocorrelation, Moran Autocorrelation, Conjoint Triad, Sequence-Order-Coupling Number, Quasi-Sequence-Order Descriptors, Pseudo-Amino Acid Composition, Amphiphilic Pseudo-Amino Acid Composition, Tripeptide Composition, and C/T/D. Conjoint Triad and Pseudo-Amino Acid Composition are among the most commonly used algorithms. Tripeptide Composition is characterized by 8000 dimensions, which are too numerous to calculate. C/T/D is an algorithm composed of three different methods and is too complicated for prediction. As such, these two algorithms are excluded, leaving the first 10 items in the list to be evaluated. The classifier provided by Weka is used for classification.

Figure 3 illustrates the prediction results of cross-validation and supplied test set validation. We use the random forest as the classifier of extraction in Weka. The prediction accuracy rate of Conjoint Triad and Pseudo-Amino Acid Composition is 70%, which is higher than that of other algorithms. We also reduce the number of dimensions of the feature extraction algorithms by using MRMD [29]. Table 3 also lists the number of dimensions after they have been reduced. Figure 4 reveals the accuracy rates of the prediction before and after the dimensions have been reduced.

Table 2: Dimensions of feature extraction algorithms in Part I.

Mode	Dimension
Pse-in-one	22
188 dimensions	188
473 dimensions	473
1-skip	400
2-skip	400
188 dimensions + Pse-in-one	210
473 dimensions + Pse-in-one	495
473 dimensions + 188 dimensions	661
473 dimensions + 188 dimensions + Pse-in-one	683
473 dimensions + 188 dimensions + Pse-in-one + 1-skip	1083
473 dimensions + 188 dimensions + Pse-in-one + 2-skip	1083

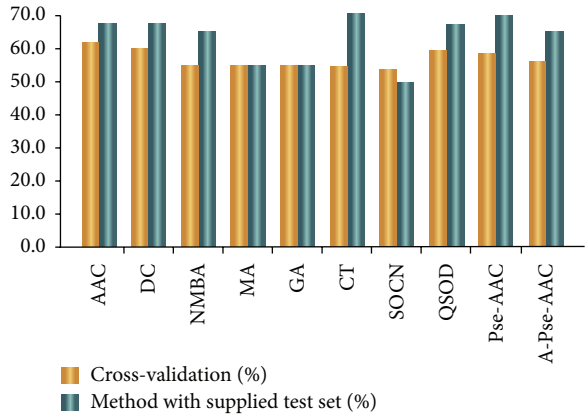


Figure 3: The accuracy rate of prediction in ProtrWeb.

Table 3: Dimensions of feature extraction algorithms in ProtrWeb.

Mode	Dimension	Dimension reduction
Amino Acid Composition	20	19
Dipeptide Composition	400	49
Normalized Moreau-Broto Autocorrelation	240	47
Moran Autocorrelation	240	43
Geary autocorrelation	240	220
Conjoint Triad	343	81
Sequence-Order-Coupling Number	60	17
Quasi-Sequence-Order Descriptors	100	42
Pseudo-Amino Acid Composition	50	23
Amphiphilic Pseudo-Amino Acid Composition	80	15

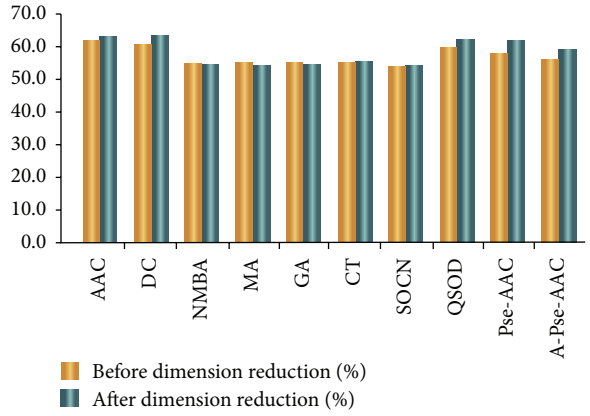


Figure 4: The accuracy rate of prediction before and after dimension reduction.

3.3. ProtrWeb-libSVM Prediction. In this section, Conjoint Triad and Pseudo-Amino Acid Composition from ProtrWeb are included in libSVM for another cycle of evaluation. Considering the high accuracy rate of these two algorithms, which are classified by the classifiers in Weka, we aim to determine whether a more accurate prediction result can be obtained when a different classifier is used. Although Weka is a software suite into which various classification tools are integrated, some methods of prediction cannot be used with it. Hence, we employ libSVM for prediction. Each step in libSVM should be executed in the command line. For libSVM, the parameters c and g are set as the default values. Notably, $g = 1/k$, where k is the number of the cancerlectins.

Despite the advantages of libSVM, this method is still unable to achieve sufficient accuracy of prediction. Further studies should include additional parameters in the command line to obtain a prediction result that is close to the actual findings. To improve the predictive accuracy of this method, Grid is used to optimize the parameters c and g. Table 4 summarizes the prediction results obtained in libSVM. The two methods fail to obtain high accuracy rates when classification is performed after these parameters have optimized.

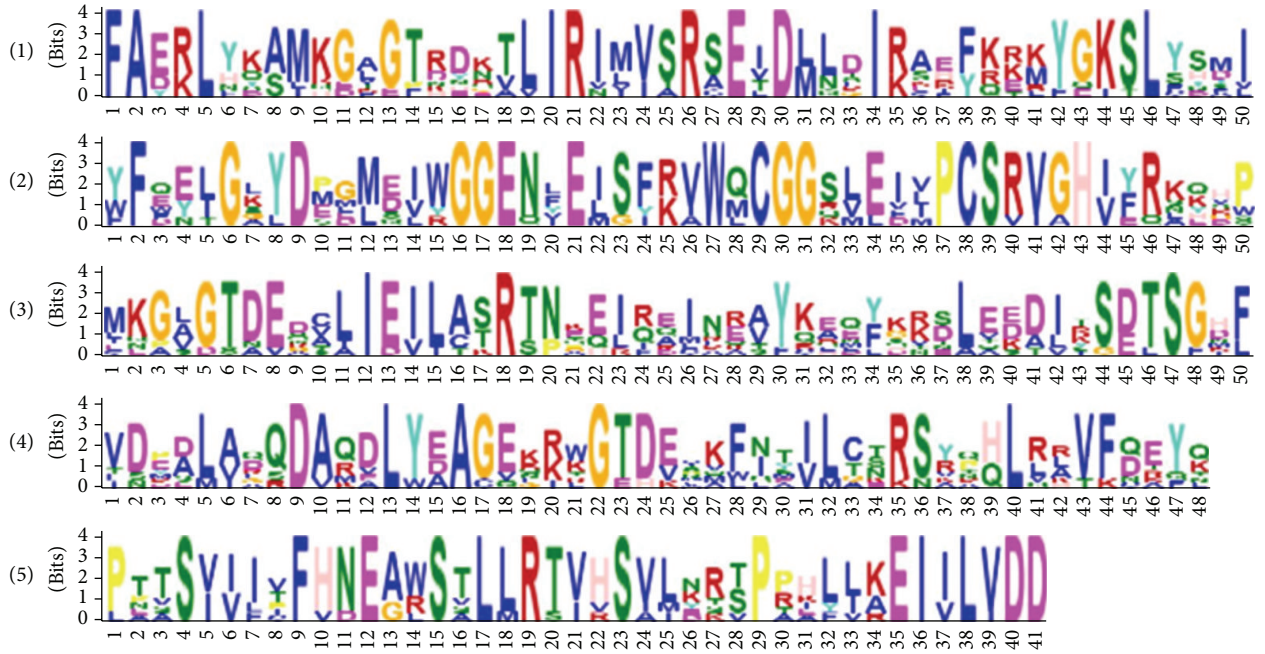
Table 4: The prediction results of libSVM.

Mode	libSVM (%)	libSVM + Grid (%)
Conjoint Triad	55.9406	81.1881
Pseudo-Amino Acid Composition	86.1042	70.9677

4. Conclusions

Amino acid feature extraction and classification are major components of the prediction and classification of protein function. With advances in biology, medicine, and the biopharmaceutical industry, it should be possible to determine the positions of different proteins in cells. Although various amino acid feature extraction, fusion, and classification algorithms have been developed [43], they are independent of one another and are used for analyses in only one specific field of study. The combination of the two algorithms of feature extraction and classification has rarely been investigated and efficient methods for protein function prediction have seldom been developed. In this study, we comprehensively considered the two algorithms of feature extraction and classification in terms of their data set and basic logic. In this way, we determined the optimal strategy for combining feature extraction algorithms and classifiers. Thus, we performed numerous experiments and trials involving different algorithms. After conducting a substantial number of tests, we proposed a prediction method of predicting protein function comprising feature extraction and random forest classification based on Conjoint Triad and Pseudo-Amino Acid Composition. By using this combination, the accuracy rate reached 70%, which is higher than those of other prediction methods.

Our newly proposed method can thus be used to identify cancerlectins with reasonably high accuracy. Several network-based computational methods have already been applied to identify oncogenes [44] or oncomiRNA [45]. In addition, advanced social network algorithms have helped to predict the relationship between diseases and miRNA [46, 47]. However, network-based methods involve similar computation methods between miRNAs [48] or genes [49].



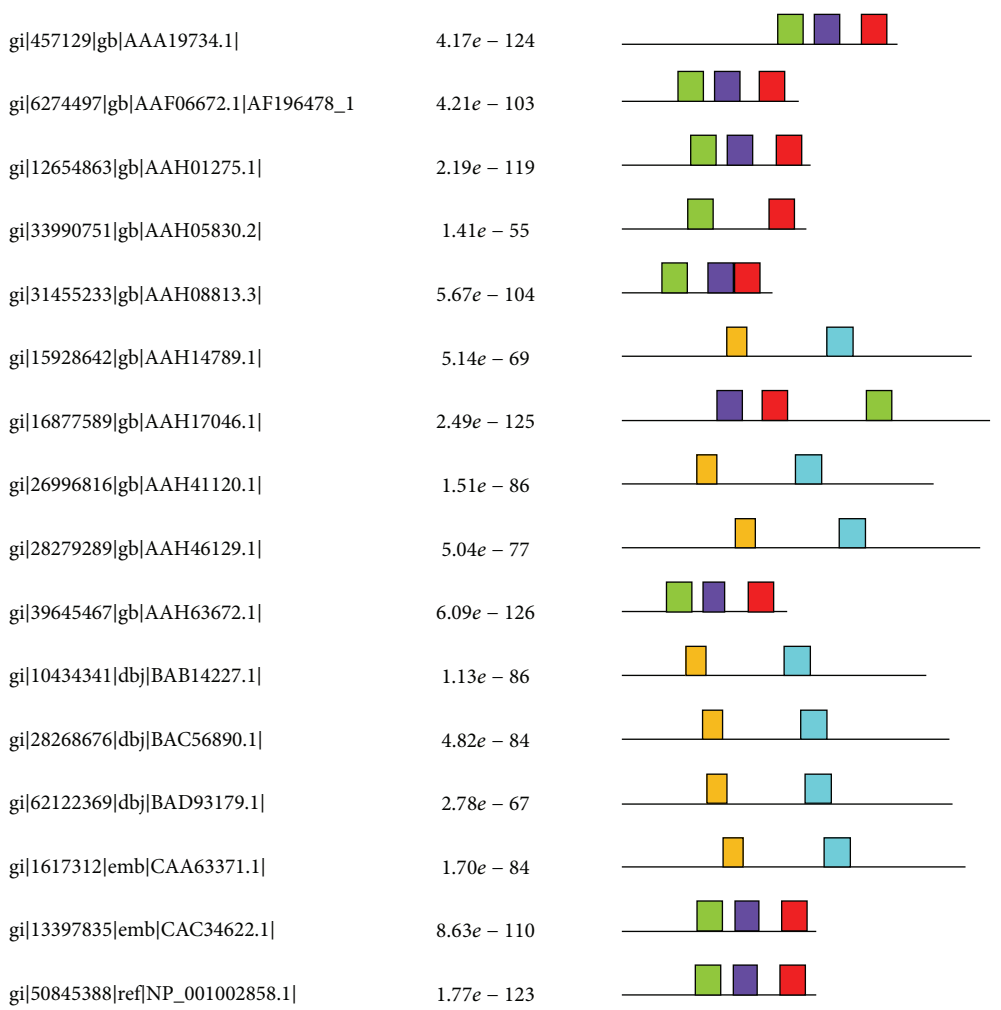
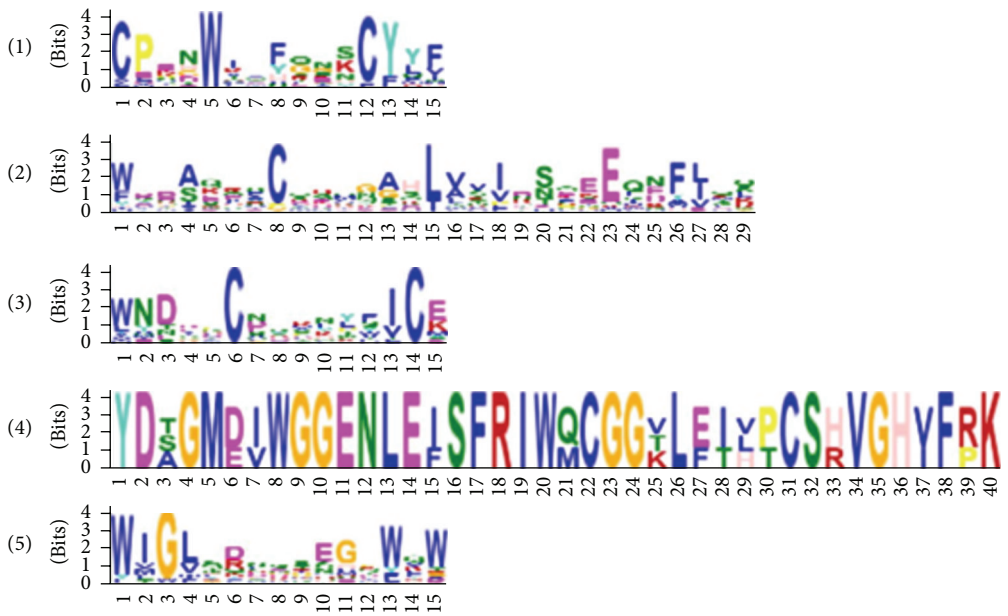


Figure 5: The most significant 5 conserved motifs of the first group.



gi|13346506|ref|NP_008857.1|	1.11$e-16$
gi|75709177|ref|NP_031359.2|	5.73$e-20$
gi|126090892|ref|NP_037359.2|	7.99$e-34$
gi|7019447|ref|NP_037401.1|	3.35$e-24$
gi|42542390|ref|NP_055072.3|	6.81$e-45$
gi|13124891|ref|NP_065207.2|	1.88$e-50$
gi|112807221|ref|NP_078918.3|	2.83$e-40$
gi|18641360|ref|NP_569057.1|	1.28$e-43$
gi|33667103|ref|NP_878910.1|	9.69$e-43$
gi|74720063|sp|Q9UBG0|MRC2_HUMAN	1.79$e-35$
gi|193108|gb|AAA37577.1|	3.68$e-24$
gi|2138183|gb|AAC53207.1|	6.87$e-5$
gi|7677469|gb|AAF67177.1|AF240357_1	5.09$e-44$
gi|13784939|gb|AAH02001.3|	7.10$e-16$
gi|13277590|gb|AAH03705.1|	1.53$e-18$
gi|13879298|gb|AAH06623.1|	1.35$e-34$
gi|15030306|gb|AAH11428.1|	6.35$e-41$
gi|17512381|gb|AAH19154.1|	1.18$e-9$
gi|20381202|gb|AAH27742.1|	9.27$e-40$
gi|34784347|gb|AAH57147.1|	8.50$e-39$
gi|34980877|gb|AAH57165.1|	7.60$e-5$
gi|87299635|ref|NP_033860.2|	2.71$e-11$
gi|7305051|ref|NP_038545.1|	4.68$e-36$
gi|16716407|ref|NP_444339.1|	2.95$e-35$
gi|46559408|ref|NP_573466.1|	4.59$e-30$
gi|31560764|ref|NP_659192.2|	1.87$e-37$
gi|30520371|ref|NP_705818.2|	3.28$e-28$
gi|121674803|ref|NP_766320.2|	2.73$e-38$
gi|21541998|sp|O89103|C1QR1_MOUSE	5.51$e-15$
gi|114638554|ref|XP_522073.2|	7.88$e-3$
gi|9633932|ref|NP_052011.1|	1.89$e-13$

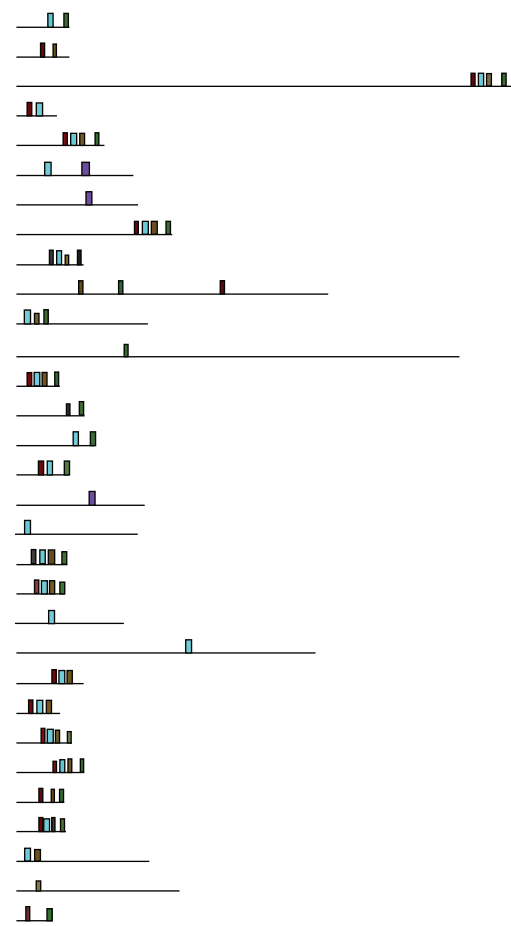


Figure 6: The most significant 5 conserved motifs of the second group.

Table 5: The 5 most significant conserved motifs of the first group.

Motif	Width	*E* value	Best possible match
1	50	1.3*e* − 157	FA[ED][RK]L[YH][KQ][AS]MKG[AL]GT[RD]D[KN][TV]LIRI[ML][VI]SR[SA]E[ITV]D[LM][LN]DI[RK][AS][EH][FY][KQR][KRE][KM]YGKSL[YS][SH][MD]I
2	50	3.3*e* − 131	[YW]F[EQ][EY][LI]G[KL]YD[EMP]G[ML][ED][IV]WGGEN[FL]E[IL]SF[RK]VW[QM]CGGS[LV]EI[ILV]PCSRVGH[IV][FY]RK[KQ]HP
3	50	6.2*e* − 105	MKG[ALV]GTDED[CAV]LIE[IV]L[AC][ST]R[TS][NP][EK][EQ][IL][RQ][EAQ]IN[ER][AV]Y[KQ][EA][QE][FY][KG][KR][DS]LE[DEK][DA][IL][KRT]S[DE]TSG[HD][FL]
4	50	1.6*e* − 089	VD[EP][AD]L[AV][DQ]QDA[QR]DLY[EAD]AGEK[RK][WK]GTD[EV]XKF[IN]T[IV]L[CT][NST]RS[YR][PQ][HQ]L[RL][ALR]VF[DQ]EY[QK]
5	41	2.1*e* − 086	PTTS[VI][IV]I[TV]FHNE[AG][WR]STLLRT[VI]HSVL[KN]R[ST]P[PR]HL[LI][KA]EI[IV]LVDD

Table 6: The 5 most significant conserved motifs of the second group.

Motif	Width	*E* value	Best possible match
1	15	8.2*e* − 066	CPENWIX[FY][GQ]N[KS]CY[YL]F
2	29	2.1*e* − 071	[WF]XD[AS][QEK]XXCXXXG[AG]HL[VA][VS][IV]D[SN]XEEQ[NDE]F[LI]QQ
3	15	2.6*e* − 034	WNDXXC[ND]XK[LN][YL][FS][IV]C[EK]
4	40	7.7*e* − 033	YD[AST]GM[DE][IV]WGGENLE[IF]SFRIW[QM]CGG[KTV]L[EF][IT][HLV][PT]CS[HR]VGH[VI]F[RP]K
5	15	1.6*e* − 030	WIG[LV]S[DR]XXSEGXWQW

Table 7: The numbers of positive and negative samples of training set.

	Before balancing			After balancing		
	Cancerlectin	Noncancerlectin	Total	Cancerlectin	Noncancerlectin	Total
Conjoint Triad	178	226	404	356	226	582
Pseudo-Amino Acid Composition	178	225	403	356	225	581

Table 8: The comparisons before and after balancing the training set.

	Before balancing		After balancing	
	Cross-validation	Method with supplied test set	Cross-validation	Method with supplied test set
Conjoint Triad	54.9505%	70%	71.134%	67.5%
Pseudo-Amino Acid Composition	57.8164%	70%	72.4613%	67.5%

Information on interactions involving lncRNA [50, 51] and cell death [52] systems can improve prediction of the relationship between RNA and diseases. As another example of network constructing, a random walk [53] technique has been commonly employed to construct networks and predict unknown relationships. Nevertheless, network-based methods have been disregarded in cancerlectin identification. It is suggested that network features should be considered to improve classification accuracy. Further studies should also focus on the combination of network-based methods and classification techniques. Moreover, big data technologies, including Mahout and Hadoop, could be utilized to cope with large-scale data [54].

Competing Interests

The authors declare that they have no competing interests.

Authors' Contributions

Jieru Zhang analyzed data and designed the project and coordinated it. Huijuan Lu helped in the classification analysis. Ping Xuan and Ying Ju were involved in drafting the paper. Quan Zou helped revise the paper and gave helpful suggestion. All authors read and approved the final paper.

Acknowledgments

The work was supported by the Natural Science Foundation of China (no. 61370010, no. 61302139, and no. 61272315), the Natural Science Foundation of Fujian Province of China (no. 2014J01253), and the State Key Laboratory of Medicinal Chemical Biology in China.

References

[1] A. A. Schäffer, L. Aravind, T. L. Madden et al., "Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements," *Nucleic Acids Research*, vol. 29, no. 14, pp. 2994–3005, 2001.

[2] R. D. Finn, J. Clements, and S. R. Eddy, "HMMER web server: interactive sequence similarity searching," *Nucleic Acids Research*, vol. 39, no. 2, Article ID gkr367, pp. W29–W37, 2011.

[3] 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.

[4] W. Chen, P. Feng, and H. Lin, "Prediction of ketoacyl synthase family using reduced amino acid alphabets," *Journal of Industrial Microbiology and Biotechnology*, vol. 39, no. 4, pp. 579–584, 2012.

[5] Q. Zou, Y. Mao, L. Hu, Y. Wu, and Z. Ji, "miRClassify: an advanced web server for miRNA family classification and annotation," *Computers in Biology and Medicine*, vol. 45, no. 1, pp. 157–160, 2014.

[6] W. Chen and H. Lin, "Identification of voltage-gated potassium channel subfamilies from sequence information using support vector machine," *Computers in Biology and Medicine*, vol. 42, no. 4, pp. 504–507, 2012.

[7] T. Song, L. Pan, J. Wang, I. Venkat, K. G. Subramanian, and R. Abdullah, "Normal forms of spiking neural P systems with anti-spikes," *IEEE Transactions on Nanobioscience*, vol. 11, no. 4, pp. 352–359, 2012.

[8] H. Lin and W. Chen, "Prediction of thermophilic proteins using feature selection technique," *Journal of Microbiological Methods*, vol. 84, no. 1, pp. 67–70, 2011.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

[16] X. Zeng, L. Xu, X. Liu, and L. Pan, "On languages generated by spiking neural P systems with weights," *Information Sciences*, vol. 278, pp. 423–433, 2014.

[17] J. X. Tao Song and L. Pan, "Spiking neural P systems with request rules," *Neurocomputing*, vol. 193, pp. 193–200, 2016.

[18] T. Song, J. Xu, 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.

[19] X. Zhang, Y. Tian, R. Cheng, and Y. Jin, "An efficient approach to nondominated sorting for evolutionary multiobjective optimization," *IEEE Transactions on Evolutionary Computation*, vol. 19, no. 2, pp. 201–213, 2015.

[20] Y. T. Xingyi Zhang 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.

[21] B. Liu, F. Liu, X. Wang, J. Chen, L. Fang, and K. 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. W1, pp. W65–W71, 2015.

[22] 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 15479, 2015.

[23] H. Lin, W. Chen, L.-F. Yuan, Z.-Q. Li, and H. Ding, "Using over-represented tetrapeptides to predict protein submitochondria locations," *Acta Biotheoretica*, vol. 61, no. 2, pp. 259–268, 2013.

[24] 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.

[25] W. Chen and H. Lin, "Prediction of midbody, centrosome and kinetochore proteins based on gene ontology information," *Biochemical and Biophysical Research Communications*, vol. 401, no. 3, pp. 382–384, 2010.

[26] 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.

[27] Y. Cai, T. Huang, L. Hu, X. Shi, L. Xie, and Y. Li, "Prediction of lysine ubiquitination with mRMR feature selection and analysis," *Amino Acids*, vol. 42, no. 4, pp. 1387–1395, 2012.

[28] G. Huang, L. Lu, K. Feng et al., "Prediction of S-nitrosylation modification sites based on kernel sparse representation classification and mRMR algorithm," *BioMed Research International*, vol. 2014, Article ID 438341, 10 pages, 2014.

[29] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016.

[30] T. Song and L. Pan, "Spiking neural P systems with rules on synapses working in maximum spikes consumption strategy," *IEEE Transactions on Nanobioscience*, vol. 14, no. 1, pp. 37–43, 2015.

[31] R. Kumar, B. Panwar, J. S. Chauhan, and G. P. Raghava, "Analysis and prediction of cancerlectins using evolutionary and domain information," *BMC Research Notes*, vol. 4, article 237, 2011.

[32] D. Damodaran, J. Jeyakani, A. Chauhan, N. Kumar, N. R. Chandra, and A. Surolia, "CancerLectinDB: adatabase of lectins relevant to cancer," *Glycoconjugate Journal*, vol. 25, no. 3, pp. 191–198, 2008.

[33] C. Z. Cai, L. Y. Han, Z. L. Ji, X. Chen, and Y. Z. Chen, "SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence," *Nucleic Acids Research*, vol. 31, no. 13, pp. 3692–3697, 2003.

[34] N. Xiao, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu, "Protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequences," *Bioinformatics*, vol. 31, no. 11, pp. 1857–1859, 2015.

[35] T. L. Bailey, M. Boden, F. A. Buske et al., "MEME Suite: tools for motif discovery and searching," *Nucleic Acids Research*, vol. 37, no. 2, pp. W202–W208, 2009.

[36] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, 2002.

[37] J. Shen, J. Zhang, X. Luo et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[38] H. Wang and X. Hu, "Accurate prediction of nuclear receptors with conjoint triad feature," *BMC Bioinformatics*, vol. 16, no. 1, article 402, pp. 1–13, 2015.

[39] 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.

[40] 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.

[41] 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.

[42] R. Wang, Y. Xu, and B. Liu, "Recombination spot identification based on gapped k-mers," *Scientific Reports*, vol. 6, article 23934, 2016.

[43] J. Zeng, D. Li, Y. Wu, Q. Zou, and X. Liu, "An empirical study of features fusion techniques for protein-protein interaction prediction," *Current Bioinformatics*, vol. 11, no. 1, pp. 4–12, 2016.

[44] Q. Zou, J. Li, C. Wang, and X. Zeng, "Approaches for recognizing disease genes based on network," *BioMed Research International*, vol. 2014, Article ID 416323, 10 pages, 2014.

[45] 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.

[46] 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.

[47] X. Zeng, X. Zhang, Y. Liao, and L. Pan, "Prediction and validation of association between microRNAs and diseases by multipath methods," *Biochimica et Biophysica Acta (BBA)—General Subjects*, 2016.

[48] 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.

[49] 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*, no. 99, p. 1, 2016.

[50] C. Yang, D. Wu, L. Gao et al., "Competing endogenous RNA networks in human cancer: hypothesis, validation, and perspectives," *Oncotarget*, vol. 7, no. 12, pp. 13479–13490, 2016.

[51] 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 19062, 2016.

[52] D. Wu, Y. Huang, J. Kang et al., "ncRDeathDB: a comprehensive bioinformatics resource for deciphering network organization of the ncRNA-mediated cell death system," *Autophagy*, vol. 11, no. 10, pp. 1917–1926, 2015.

[53] Y. Liu, X. Zeng, Z. He, and Q. Zou, "Inferring microRNA-disease associations by random walk on a heterogeneous network with multiple data sources," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2016.

[54] Q. Zou, X.-B. Li, W.-R. Jiang, Z.-Y. Lin, G.-L. Li, and K. Chen, "Survey of MapReduce frame operation in bioinformatics," *Briefings in Bioinformatics*, vol. 15, no. 4, pp. 637–647, 2014.