Hindawi Publishing Corporation
BioMed Research International
Volume 2015, Article ID 425810, 10 pages
http://dx.doi.org/10.1155/2015/425810

Research Article

Sequence-Based Prediction of RNA-Binding Proteins Using Random Forest with Minimum Redundancy Maximum Relevance Feature Selection

Xin Ma,[1] Jing Guo,[2] and Xiao Sun[2]

[1]*Golden Audit College, Nanjing Audit University, Nanjing 210029, China*
[2]*State Key Laboratory of Bioelectronics, Southeast University, Nanjing 210096, China*

Correspondence should be addressed to Xin Ma; maxin@nau.edu.cn

Received 24 June 2015; Accepted 21 September 2015

Academic Editor: Liam McGuffin

The prediction of RNA-binding proteins is one of the most challenging problems in computation biology. Although some studies have investigated this problem, the accuracy of prediction is still not sufficient. In this study, a highly accurate method was developed to predict RNA-binding proteins from amino acid sequences using random forests with the minimum redundancy maximum relevance (mRMR) method, followed by incremental feature selection (IFS). We incorporated features of conjoint triad features and three novel features: binding propensity (BP), nonbinding propensity (NBP), and evolutionary information combined with physicochemical properties (EIPP). The results showed that these novel features have important roles in improving the performance of the predictor. Using the mRMR-IFS method, our predictor achieved the best performance (86.62% accuracy and 0.737 Matthews correlation coefficient). High prediction accuracy and successful prediction performance suggested that our method can be a useful approach to identify RNA-binding proteins from sequence information.

1. Introduction

RNA-binding proteins are important functional proteins that are pivotal to a cell's function, such as in gene expression, posttranscriptional regulation, protein synthesis, and replication and assembly of many viruses [1–4]. How to discriminate RNA-binding proteins from other proteins is important to understand the mechanisms of these functions. Therefore, the reliable identification of RNA-binding proteins is an important research topic in the field of proteomics and will play a vital role in proteome functional annotation, in the discovery of potential therapeutics for genetic diseases and in reliable diagnostics. Several experimental techniques, such as X-ray crystallography, nuclear magnetic resonance, and filter binding assays have been used to identify RNA-binding proteins. However, using experimental methods to identify RNA-binding proteins is costly and time consuming. It is desirable to develop computational methods to recognize RNA-binding proteins.

Previous studies have investigated the mechanisms by which proteins bind to DNA; however, research on RNA-binding proteins lags behind. Methods to identify RNA-binding proteins could be divided into two categories: recognition from protein structure and prediction from amino acid sequences. The structure-based prediction approach usually produces a better performance; however, obtaining the protein structure is still costly and time consuming. Considering the theory that a protein's amino acid sequence contains all the necessary information to predict its function [5], we hypothesized that it would be an effective approach to predict RNA-binding proteins directly from amino acid sequences. Therefore, machine learning algorithms have been used to build classification systems to discriminate RNA-binding proteins from nonbinding ones.

Support vector machine (SVM) [6] is an effective machine learning algorithm that is the most widely used for prediction of RNA-binding proteins [7–11]. Cai and Lin first built a prediction model by using SVM and incorporated

a comprehensive set of input features based on the amino acid composition and a limited range of correlations with hydrophobicity and solvent accessible surface area [7]. Shao et al. proposed an SVM-based predictor using a conjoint triad feature, which extracts information directly from the amino acids sequences of proteins [10]. Kumar et al. developed a prediction model named RNApred (http://www.imtech.res.in/raghava/rnapred/), which also uses an SVM and uses a position-specific scoring matrix (PSSM) profile and sequence descriptors as inputs [11].

To obtain a good predictive model, two major problems should be considered. One is feature extraction and selection and the other is the selection of the classification algorithm. To solve the first problem, we proposed a novel feature called evolutionary information combined with physicochemical properties (EIPP). The results show that EIPP has a more powerful ability to distinguish RNA-binding proteins from nonbinding ones than PSSM, which dramatically improved the prediction of RNA-binding proteins compared with a previous work [11]. In our study, we used the minimum redundancy maximum relevance (mRMR) method combined with incremental feature selection (IFS) to select the optimal features, which not only reduced the dimension of the features but also improved the performance of the predictor. To solve the second problem, we choose the random forest (RF) algorithm [12] instead of the SVM algorithm because the SVM algorithm is time consuming when searching for appropriate optimal parameters, and the kernel function for the predictor and RF algorithm is an ensemble classifier with fast performance that has been applied successfully in many fields. Therefore, in this study, the mRMR-IFS feature selection approach and the RF algorithm are combined to construct the prediction model. Our results showed that the prediction model achieved 86.62% accuracy, 78.34% sensitivity, and 94.91% specificity, with a Matthews correlation coefficient of 0.737, indicating that it outperformed previous methods in predicting RNA-binding proteins.

2. Materials and Methods

2.1. Dataset. RNA-binding proteins and nonbinding proteins were obtained from release "2014_06" of the UniProtKB database (http://www.uniprot.org/) [13]. By searching with the keyword "RNA binding," we extracted 47,768 RNA-binding proteins from UniProtKB. We followed the procedure by Yu et al. [9] to obtain 545,536 nonbinding proteins. To ensure the reliability of data, we only selected manually annotated and reviewed proteins.

As indicated by previous studies [8–10], the data used in this study were selected strictly according to the following criteria. (1) Protein sequences with more than 6000 amino acids were removed because they might be protein complexes. Protein sequences less than 50 amino acids were also removed because they might be protein fragments. (2) Proteins including irregular amino acid characters such as "x" and "z" were filtered out. (3) To reduce redundancy and homology bias, the BLAST package was used in this research. The BLAST package was downloaded from NCBI utilized to remove those sequences that have 40% sequence identity to any other sequences in the dataset. To create the nonredundant dataset, the longest amino acid sequences were selected in each cluster. Finally, we obtained 2,848 RNA-binding proteins as positive instances and 83,516 nonbinding proteins as negative instances. To achieve a balance between positive instances and negative instances, we randomly selected the same number of negative instances as the number of positive instances. Therefore, the main dataset (MDset) used in this study comprised 2,848 RNA-binding proteins and 2,848 nonbinding proteins.

To evaluate the performance of our method in comparison with previously well-known studies, we used an independent test dataset (Testset). The Testset comprised 144 RNA-binding proteins and 144 nonbinding proteins obtained from MDset that had not been used in previous studies [11, 14]. The remaining proteins in MDset were designated as the training dataset (TRset). Therefore, TRset contained 2,704 RNA-binding proteins and 2,704 nonbinding proteins.

2.2. Protein Features

2.2.1. Binding Propensity and Nonbinding Propensity (BP and NBP). Prediction of RNA-binding residues was used to identify the RNA-binding proteins from nonbinding ones. We had already developed an RNA-binding residues prediction model, PRBR [15] (http://www.cbi.seu.edu.cn/PRBR/). Each amino acid could be identified by submitting the protein to the PRBR webserver. Consequently, the binding propensity measures and nonbinding propensity measures were adopted in this study, which were made based on the prediction results of RNA-binding residues and nonbinding residues, respectively.

RNA-binding proteins have many more binding residues than nonbinding proteins and RNA-binding residues tend to gather together spatially; therefore, two binding propensity measures were defined as follows:

$$\mathrm{BP}\,(1) = \frac{\sum_{i=1}^{n} \mathrm{RI}\,(i)}{10N}, \tag{1}$$

where N and n are the number of amino acids in this protein and the number of RNA-binding residues, respectively. RI(i) is the reliability index of the prediction result of RNA-binding residue i obtained from PRBR. The reliability index is a positive integer ranging from 0 to 10. Consider

$$\mathrm{BP}\,(2) = \frac{\sum_{i=1}^{N-1} 2^{-i+1} \sum_{k=1}^{n(i)} \overline{\mathrm{RI}}\,(k)}{10\,(N-1)}, \tag{2}$$

where N and $n(i)$ are the number of amino acids in this protein and the number of two RNA-binding residues at a distance i, respectively. $\overline{\mathrm{RI}}(i)$ is the average value of the reliability index for RNA-binding residue k and binding residue $k + i$.

We used predicted RNA-binding residues; therefore, the reliability index is applied in those two formulas. The BP(1) and BP(2) represent the information of the frequency and

correlation of RNA-binding residues in the query protein, respectively. Furthermore, BP(2) formula represents the relevance of the two RNA-binding residues combined with different distances from 1 to $N-1$ and takes into account the fact that the correlation value between two residues is smaller when the distance k is larger which proves the rationality of the definition.

We also defined two nonbinding propensities for nonbinding proteins. The definitions of NBP(1) and NBP(2) are similar to the definitions of BP(1) and BP(2). Consider

$$\mathrm{NBP}\,(1) = \frac{\sum_{i=1}^{n} \mathrm{RI}\,(i)}{10N}, \tag{3}$$

where N and n are the number of amino acids and the number of nonbinding residues in this protein, respectively. RI(i) is the reliability index of the prediction result of nonbinding residue i obtained from PRBR. Consider

$$\mathrm{NBP}\,(2) = \frac{\sum_{i=1}^{N-1} 2^{-i+1} \sum_{k=1}^{n(i)} \overline{\mathrm{RI}}\,(k)}{10\,(N-1)}, \tag{4}$$

where N is the number of amino acids in this protein, $n(i)$ is the number of two nonbinding residues at a distance i, and $\overline{\mathrm{RI}}(k)$ is the average value of the reliability index for nonbinding residue k and nonbinding residue $k+i$.

NBP(1) and NBP(2) describe the information of the appearance and correlation of nonbinding residues in the query protein, respectively, which are similar to BP(1) and BP(2). We also used the reliability index because the prediction result of nonbinding residues is applied in those formulas.

2.2.2. Evolutionary Information Combined with Physicochemical Properties (EIPP). Evolutionary information in the form of a position-specific scoring matrix (PSSM) has been used successfully to represent proteins in many applications, such as prediction of DNA-binding residues [16–21] and RNA-binding residues [15, 22, 23]. Here, PSSM profiles were generated using the PSI-BLAST program [24] to search the nonredundant (NR) database through three iterations, with 0.001 as the e-value cutoff for multiple sequence alignment. The PSSM scoring matrix has $20 * L$ elements, where L is the length of protein. However, different proteins may have different numbers of amino acids. Therefore, the PSSM could not be used directly as feature in the prediction work because all the machine learning methods require the input feature to have a fixed length. Therefore, we generated a PSSM-400, which has a vector of dimension of 400 from the PSSM. PSSM-400 is composition of occurrences of each type of amino acid corresponding to each type of amino acids in sequences. We pooled all rows that belonged to the same amino acid in this PSSM to form a new matrix. We then converted each new matrix to a vector and added all the normalized values in each column for the new matrixes. Therefore, we produced a 20-dimensional vector for each new matrix to generate PSSM-400.

The physicochemical property feature has been used effectively in many fields, such as the identification of DNA\RNA-binding proteins [7, 9, 14, 25, 26] and the identification and prediction of protein-protein interactions [27]. Thus, an EIPP was generated by merging 20 amino acid columns of the PSSM-400 into a single column containing the information for a certain physicochemical property. Six physicochemical properties that we used successfully in previous works [15] were considered for combining with PSSM-400 to generate the EIPP: the pKa values of the amino group, the pKa values of the carboxyl group [28], the molecular mass [6], the lowest free energy [29], the Balaban index [30], and the Wiener index [31]. The entry e_{ak} of kth type of amino acid in a protein sequence for a certain physicochemical property a in EIPP was calculated with

$$e_{ak} = \sum_{i=1}^{20} \sqrt{d_a\,(i)}\, f_k\,(i)\,, \tag{5}$$

where a is the index of a certain physicochemical property, k is the index of the type of amino acids in the query protein sequence, i is the index of the type of naïve amino acids, $f_k(i)$ is the normalized value of the ith type of naïve amino acid for the kth type of amino acid in the protein sequence of the PSSM-400, and $d_a(i)$ is the normalized physicochemical property values of a for the ith type amino acids. Therefore, the vector size of EIPP feature is 6×20.

2.2.3. Conjoint Triad (CT). Electrostatic and hydrophobic interactions influence protein-nucleic acid interactions and may be reflected by the dipoles and volumes of the side chains of amino acids, respectively. Based on the dipoles and volumes of the side chains, the 20 kinds of amino acids could be clustered into seven classes [32]. Considering that disulfide bonds have no special effect on protein-nucleic acid interactions, the unique amino acid cysteine in the seventh class was put back to the third class in this study. Therefore, the 20 kinds of amino acids were clustered into six classes as follows. Class a: Ala, Gly, and Val; Class b: Ile, Leu, Phe, and Pro; Class c: Tyr, Met, Thr, Ser, and Cys; Class d: His, Asn, Gln, and Tpr; Class e: Arg and Lys; and Class f: Asp and Glu. According to the similar feature construction method used in [32], a protein is described by the conjoint triads feature with $6 \times 6 \times 6 = 216$ dimensions, where each component of the feature vector has the value of the frequency of the corresponding triad.

As mentioned above, for each query protein, the vector size of a feature is $4 + 120 + 216 = 340$.

2.3. Algorithms to Classify and Measure a Classifier's Performance. The random forest (RF) algorithm [12] is a classification algorithm that uses an ensemble of tree-structured classifiers, which has been used successfully in many applications for data classification and achieves high performance. The random forest R package [33] was used to implement the RF algorithm.

To evaluate the performance of the classifier, a 5-fold cross-validation procedure for the training dataset was used in this research. During the procedure, we randomly divided the data instances into five parts. Four of these parts were input into the RF to establish a model for classification,

and every instance of the remaining part was predicted by the model. Ultimately, the prediction performance of the classifier was evaluated by the remaining part.

To evaluate the performance of the RNA-binding proteins predictor, the accuracy, sensitivity, specificity, and Matthews correlation coefficient (MCC) were calculated as follows:

$$\begin{aligned}
\text{Accuracy} &= \frac{(\text{TP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})},\\
\text{Sensitivity} &= \frac{\text{TP}}{(\text{TP} + \text{FN})},\\
\text{Specificity} &= \frac{\text{TN}}{(\text{TN} + \text{FP})},\\
\text{MCC} &= \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP})\,(\text{TN} + \text{FN})\,(\text{TP} + \text{FN})\,(\text{TN} + \text{FP})}},
\end{aligned} \tag{6}$$

where TP, TN, FP, and FN represent the number of true positive, true negative, false positive, and false negative results, respectively.

2.4. Minimum Redundancy Maximum Relevance (mRMR) and Incremental Feature Selection (IFS). Considering the successful application on several classification researches [34–42] by using the minimum redundancy Maximum relevance (mRMR) method combining with incremental feature selection (IFS) method, the mRMR-IFS was used in this research to select the prominent features that distinguish the RNA-binding proteins from nonbinding ones.

The mRMR method was developed by Peng et al. [43]. Here, we used it for feature analysis and selection. It selects candidate features with both the maximum relevance for the target and the minimum redundancy relative to the features already selected. To calculate relevance and redundancy, we used mutual information (MI), which is defined as follows:

$$\text{MI}(x, y) = \iint p(x, y) \log \frac{p(x, y)}{p(x)\, p(y)}\, dx\, dy. \tag{7}$$

In (7), $p(x, y)$ is the joint probabilistic density of random vectors x and y, and $p(x)$ and $p(y)$ are the marginal probabilities.

Let Ω, Ω_s, and Ω_t denote the whole feature set, the already-selected feature set containing m features, and the to-be-selected feature set containing n features, respectively.

To obtain the feature f_t in Ω_t with the maximal relevance for the target c and the minimal redundancy relative to the features in Ω_s, the mRMR function is defined as

$$f_t = \max_{f_j \in \Omega_t} \left[\text{MI}\left(f_j, c\right) - \frac{1}{m} \sum_{f_i \in \Omega_s} \text{MI}\left(f_j, f_i\right) \right], \quad \left(j = 1, 2, \ldots, n\right). \tag{8}$$

In this study, after using the mRMR method, all of the 340 features were ordered as follows:

$$S = \left\{ f'_1, f'_2, \ldots, f'_h, \ldots, f'_{340} \right\}. \tag{9}$$

In (9), the earlier the feature satisfying (8), and the smaller the index h, the better the feature.

To select the optimal features, we used incremental feature selection (IFS) [35, 36], which is based on the results of mRMR. We first built 340 feature sets from the ordered feature set S (9), with the ith feature set being

$$S_i = \left\{ f'_1, f'_2, \ldots, f'_i \right\}, \quad (i = 1, 2, \ldots, 340). \tag{10}$$

We then constructed 340 individual predictors for the 340 feature sets to predict RNA-binding proteins. Each predictor was constructed by the RF algorithm and evaluated by 5-fold cross-validation. The 340 MCC values were calculated from all the predictors and obtained the IFS curve with feature index i of S_i as the x-axis and the MCC value as the y-axis. Finally, the optimal feature set was obtained when the IFS curve reached its peak.

3. Results and Discussion

3.1. Prediction of RNA-Binding Proteins Using Various Features. We explored the performance of RF-based predictors for predicting RNA-binding proteins by various features. The prediction results of the individual RF-based predictors using 10 cycles of 5-fold cross-validation over the MDset are shown in Table 1. First, three features, including PSSM-400, EIPP, and CT, were used to construct RF predictors to predict RNA-binding proteins. As shown in Table 1, the classifier using EIPP achieved a higher performance than the other predictors using a single feature, with 83.11% accuracy and an MCC of 0.662. Therefore, we proposed that EIPP, which provides the evolutionary information and physicochemical properties information of the protein, could effectively distinguish RNA-binding proteins from nonbinding ones. It was obvious that the EIPP features are more powerful than the commonly used PSSM-400. Therefore, we used EIPP as a significant feature instead of PSSM-400 in this study. Although the vector dimensions of the BP and NBP features are the lowest of all the features, they play an important role in improving the performance of the classifier. When the BP and NBP features were combined with EIPP, the accuracy and MCC increased dramatically to 84.28% and 0.704, respectively. When they were combined with CT, the accuracy and MCC also increased, to 76.61% and 0.568, respectively, which are not as good as the performance obtained by the combination of the EIPP, BP, and NBP. Finally, we found that the combination of EIPP, BP, NBP, and CT achieved the best performance, with the results for accuracy, sensitivity, specificity, and MCC of 85.73%, 77.64%, 94.24%, and 0.729, respectively. Thus, the mRMR-IFS method was used to select an optimal feature set from all features, including EIPP, BP, NBP, and CT in this study.

3.2. mRMR Results. We ranked a list of 340 features for MDset dataset using the mRMR method, which was downloaded from http://penglab.janelia.org/proj/mRMR/index.htm. Within this mRMR list, a smaller index value for a feature represents higher importance in the prediction of RNA-binding proteins. The ranked 340-feature list was then

Table 1: The prediction performance of the RF model based on various features, evaluated by 10 cycles of 5-fold cross-validation on the MDset dataset.

Feature	Accuracy ± SD	Sensitivity ± SD	Specificity ± SD	MCC ± SD
PSSM-400	0.7967 ± 0.0062	0.7003 ± 0.0093	0.8894 ± 0.0075	0.620 ± 0.016
EIPP	0.8311 ± 0.0105	0.7487 ± 0.0071	0.9107 ± 0.0129	0.662 ± 0.021
CT	0.7482 ± 0.0092	0.6591 ± 0.0067	0.8406 ± 0.0153	0.5096 ± 0.015
EIPP + BP + NBP	0.8428 ± 0.0038	0.7573 ± 0.0082	0.9367 ± 0.0043	0.704 ± 0.008
CT + BP + NBP	0.7661 ± 0.0197	0.7034 ± 0.0132	0.8587 ± 0.0114	0.568 ± 0.026
EIPP + CT	0.8317 ± 0.0139	0.7482 ± 0.0068	0.9202 ± 0.0127	0.671 ± 0.018
EIPP + BP + NBP + CT	0.8573 ± 0.0117	0.7764 ± 0.0143	0.9424 ± 0.0062	0.729 ± 0.020

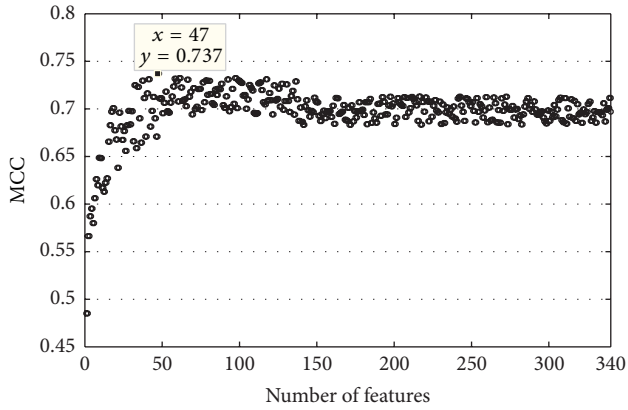

Figure 1: The IFS curve showing MCC values against feature numbers. The maximum MCC value was 0.684 when the top 47 features were selected.

used in the IFS procedure for optimal feature selection and analysis.

3.3. IFS Results. Based on the list of 340 features obtained from the mRMR method, we obtained 340-feature subsets. We then built 340 individual predictors for the 340-subfeature sets to predict RNA-binding proteins, evaluated by 5-fold cross-validation on the MDset dataset. As shown in Figure 1, the IFS curve was plotted by feature indices and MCC values obtained from the corresponding predictor. Using the top 47 features, the maximum MCC value was 0.737. Using these 47 features, the performance of the predictor was better than that of the predictor using all 340 features, with the results for accuracy, sensitivity, specificity, and MCC increasing to 86.62%, 78.34%, 94.91%, and 0.737, respectively. Therefore, these 47 optimal features were considered as the optimal feature set to be used in our final prediction model for predicting RNA-binding protein. The 47 optimal features are shown in Table 2.

3.4. Analysis of 47 Features in the Optimal Feature Set

3.4.1. Analysis of the Optimal Feature Set. As described in Section 2, there are three types of features in this study, namely, BP/NBP, EIPP, and CT. For the MDset dataset, all of the three types of features with 340 dimensions were reduced to 47 dimensions after mRMR-IFS feature selection process. The number of each type of feature in the optimal feature set is shown in Figure 2(a). The selection proportion of each type of feature for the corresponding type of feature is shown in Figure 2(b).

As shown in Figure 2(a), there were four BP and NBP features, 19 EIPP features and 24 CT features. Although the number of BP and NBP features in the optimal feature set was four, which was the least among the three types of features, the number of BP and NBP features in the original feature set was also four; thus, the selection proportion of BP and NBP features was 100%. This result showed that BP and NBP play an important role in distinguishing RNA-binding proteins from nonbinding ones. The EIPP features and CT features have similar numbers in the optimal feature set. However, the selection proportion of EIPP features (15.83%) is almost one and half times as many as the selection proportion of CT features (11.11%). This result indicated that EIPP also plays a vital role in RNA-binding proteins prediction and that CT contributes the least to the prediction of RNA-binding proteins, which is consistent with the result obtained from Table 1.

3.4.2. Analysis of BP and NBP Features in the Optimal Feature Set. All four BP and NBP features in the original feature dataset were selected to the optimal feature set, which revealed that BP and NBP features contribute mostly to distinguish RNA-binding proteins from nonbinding ones. We also calculated the p values of BP and NBP features between the binding proteins and the nonbinding ones to measure the discrimination ability. Each of them was less than 0.00005. These results also proved that BP and NBP could successfully discriminate between DNA-binding proteins and nonbinding proteins.

The superior performance of BP and NBP features represents the reliability of the definition of BP and NBP features. The detailed explanation for the reliability of the definitions of BP and NBP could be as follows. Compared with nonbinding proteins, RNA-binding residues should show a higher tendency to exist in binding proteins and RNA-binding residues should tend to gather together spatially on the surface of an RNA-binding protein. The two BP features revealed the character of RNA-binding proteins at the sequence level and the spatial level, respectively. By contrast, the proportion of nonbinding residues should be much higher for nonbinding proteins in comparison to RNA-binding proteins. This phenomenon represents the reliability

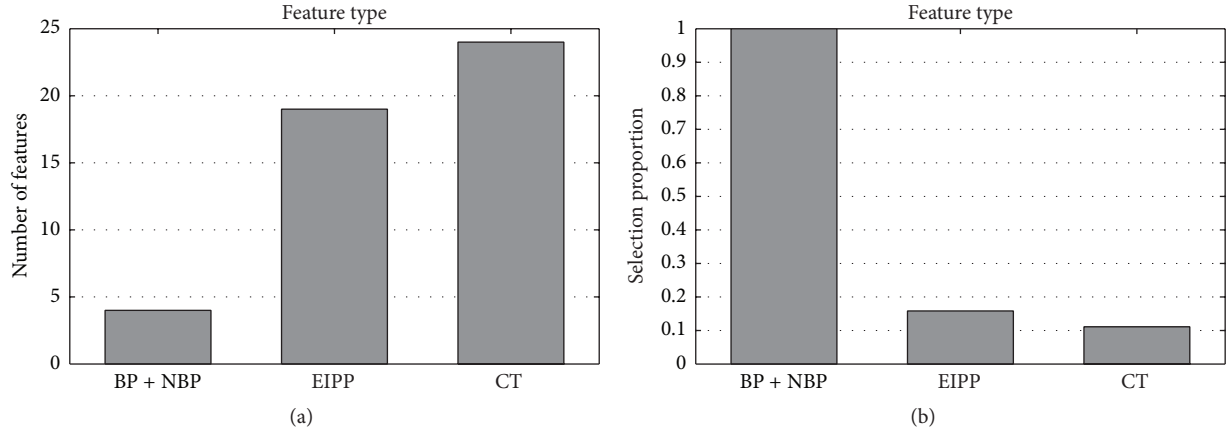

Figure 2: (a) Feature distribution for the 47 optimal features. (b) The selection proportion of each type of feature.

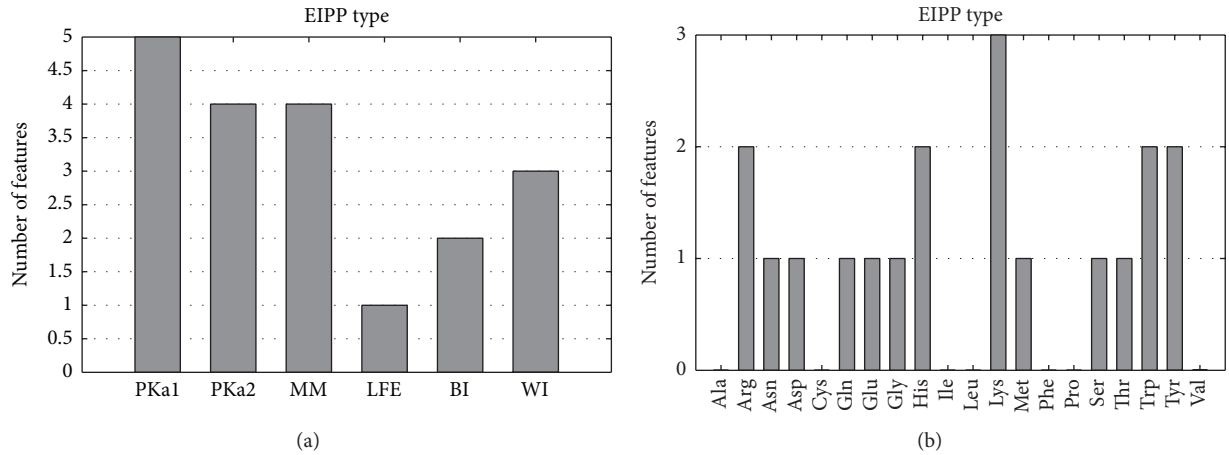

Figure 3: (a) Physicochemical property distribution to construct the 19 EIPP features that were selected in the optimal feature set. (b) The type of amino acids distribution to construct the 19 EIPP features that were selected in the optimal feature set.

of the proposed NBP feature. Therefore, BP and NBP features worked well, as we expected.

3.4.3. Analysis of EIPP Features in the Optimal Feature Set. We selected 19 EIPP features in the optimal feature set after using the mRMR-IFS method. Considering that EIPP was constructed by the evolutionary information of each type of amino acid in sequences and physicochemical property, we collected the statistics of the number of each type of amino acid and the number of each type of physicochemical property that constituted the 19 EIPP features. Figures 3(a) and 3(b) show the contributions of the number of each type of physicochemical property and the number of each type of amino acid, respectively.

As seen from Figure 3(a), there are five features related to the pKa values of amino group (PKa1), four features related to the pKa values of carboxyl group (PKa2), four features related to the molecular mass (MM), one feature related to the lowest free energy (LFE), two features related to the Balaban index (BI), and three features related to the Wiener index (WI). Compared with all physicochemical properties used in EIPP, PKa1 and PKa2, which determine the ionization state of a residue, are most essential for protein-RNA interaction. The reason is that the ionization state of amino acid side chains affects the interaction with RNA molecules, which have negatively charged phosphate groups. Molecular mass is irreplaceable in protein-RNA interactions because it is related to the volume of space that a residue occupies in the structure. The topological indices of a molecule, such as the Wiener index and the Balaban index, also play an important role in binding activity. Figure 3(b) shows that lysine, arginine, histidine, tryptophan, and tyrosine most frequently constituted the 19 EIPP features. This is most likely because those types of amino acids are abundant in RNA-binding sites and show the highest binding propensities for RNA-protein interactions. This is consistent with several results obtained from previous studies [44–46]. Lysine, arginine, and histidine show the highest binding tendency because they are positively charged amino acids and can easily interact with the negatively charged phosphate backbone of RNA.

3.4.4. Analysis of CT Features in Optimal Feature Set. Twenty-four CT features were selected in the optimal feature set and

Table 2: Optimal 47 features for prediction of RNA-binding proteins.

Rank	Feature
1	EIPP of ASP in protein sequence for the pKa values of amino group
2	EIPP of GLU in protein sequence for the Balaban index
3	BP(2)
4	EIPP of TYR in protein sequence for the pKa values of amino group
5	CT of class a, class b, and class e
6	CT of class d, class b, and class e
7	EIPP of HIS in protein sequence for the pKa values of amino group
8	EIPP of LYS in protein sequence for the pKa values of carboxyl group
9	CT of class b, class d, and class e
10	CT of class d, class c, and class e
11	EIPP of MET in protein sequence for the molecular mass
12	CT of class b, class e, and class a
13	EIPP of ARG in protein sequence for the pKa values of amino group
14	NBP(2)
15	CT of class c, class e, and class d
16	BP(1)
17	EIPP of TRP in protein sequence for the pKa values of amino group
18	CT of class d, class d, and class e
19	EIPP of LYS in protein sequence for the Balaban index
20	NBP(1)
21	CT of class c, class a, and class d
22	CT of class b, class e, and class d
23	CT of class e, class d, and class e
24	EIPP of HIS in protein sequence for the pKa values of carboxyl group
25	CT of class d, class c, and class f
26	CT of class e, class f, and class d
27	CT of class e, class b, and class d
28	CT of class d, class e, and class c
29	EIPP of GLY in protein sequence for the pKa values of carboxyl group
30	EIPP of THR in protein sequence for the molecular mass
31	CT of class c, class b, and class e
32	CT of class c, class e, and class a
33	EIPP of GLN in protein sequence for Wiener index
34	EIPP of SER in protein sequence for Wiener index
35	EIPP of ASN in protein sequence for the molecular mass
36	CT of class b, class a, and class c
37	CT of class e, class d, and class f
38	CT of class e, class b, and class a
39	EIPP of TRP in protein sequence for the pKa values of carboxyl group
40	CT of class a, class e, and class c
41	EIPP of ARG in protein sequence for the lowest free energy
42	CT of class e, class c, and class d
43	EIPP of LYS in protein sequence for the molecular mass
44	CT of class e, class e, and class d
45	EIPP of TYR in protein sequence for Wiener index
46	CT of class e, class c, and class b
47	CT of class f, class c, and class d

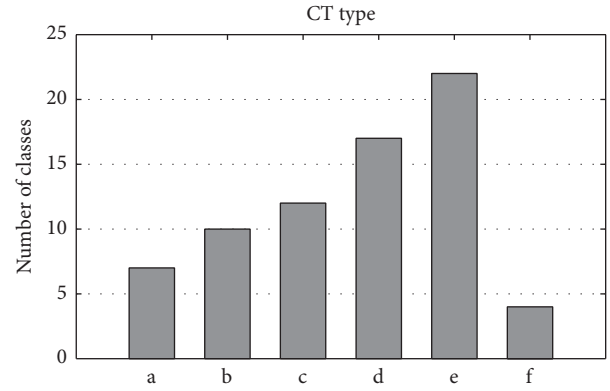

Figure 4: The type of class distribution to construct the 24 CT features that were selected in the optimal feature set.

the number of each type of class, which constituted the 24 CT features, was analyzed and shown in Figure 4. There are 72 classes comprising the 24 CT features. Classes e and d show the highest occurrence numbers, 22 and 17, respectively. The result is rational, because lysine and arginine belong to class e, which showed the easiest interaction ability with RNA. Class c appeared 12 times in the 24 CT features, which ranked third among the six classes, perhaps because class c has five types of amino acids, the most number of types of amino acids among six classes. Class f occurred the least frequently, at four times in 24 CT features. This is because glutamate and aspartate, which constitute class f, are negatively charged amino acids, which would find it harder to interact with the negatively charged phosphate backbone of RNA.

3.5. Comparison with Existing Methods on an Independent Dataset. To evaluate the effectiveness of our protocol, we compared the performance of our method with existing methods. Currently, there are two webservers for identifying RNA-binding proteins based on sequence information. One

Table 3: Comparison of the predicted results by our method and some webservers on the Testset.

Method	ACC (%)	SE (%)	SP (%)	MCC
Our method	0.7674	0.7222	0.8125	0.537
SVMprot	0.5764	0.7639	0.3889	0.165
RNApred	0.6111	0.6389	0.5833	0.223

is SVMprot by Han et al. [8] (http://jing.cz3.nus.edu.sg/cgi-bin/svmprot.cgi), which predicts RNA-binding proteins using SVM with encoded representations of tabulated residue properties as features. The other is RNApred by Kumar et al. [11] (http://www.imtech.res.in/raghava/rnapred/), which predicts RNA-binding proteins by SVM with PSSMs. To ensure that the comparison result is fair, a dataset Testset, with 288 proteins, was used as an independent test dataset, which did not include the proteins mentioned in [8, 11]. The mRMR-IFS selection model was reconstructed based on the training dataset TRset and used to predict the putative RNA-binding proteins in the Testset. Our method correctly predicted 117 out of 144 RNA-binding proteins and 104 out of 144 nonbinding proteins. Then we submitted the proteins in Testset to SVMprot. Out of 144 RNA-binding proteins, SVMprot correctly predicted 56 as RNA-binding proteins. Out of 144 nonbinding proteins, it correctly predicted 110 as nonbinding proteins. When we tested the RNApred, we found that after a protein sequence was submitted to the server, no prediction results were received from the servers. Therefore, we repeated the same method as Kumar et al.'s work and reconstructed RNApred model based on the Main dataset mentioned in [11]. The reconstructed RNApred was then used to predict RNA-binding proteins in Testset. It correctly predicted 84 out of 144 RNA-binding proteins and 92 out of 144 negatives. The detailed comparison results of the three methods are shown in Table 3. The results demonstrated that our method outperformed those previous methods in the prediction of RNA-binding proteins. The excellent results were due to the effective features and the mRMR-IFS feature selection.

4. Conclusions

Accurate identification of new RNA-binding proteins is important to understand RNA-protein interactions. In this study, an accurate method was developed to predict RNA-binding proteins using only sequence information. We proposed three novel features, binding propensity (BP), nonbinding propensity (NBP), and evolutionary information combining with physicochemical properties (EIPP). BP and NBP were constructed based on the prediction results of RNA-binding residues and nonbinding residues, respectively. The EIPP features were improved on those of PSSM by combining evolutionary information with physicochemical properties. The results showed that using those novel features dramatically improved the prediction performance and were effective in distinguishing RNA-binding proteins from nonbinding ones. The mRMR-IFS feature selection method and RF algorithm are then utilized to construct the prediction model. This is the first study in which the mRMR-IFS feature selection method has been successfully used to predict RNA-binding proteins. The prediction model achieved excellent performance, with 86.62% accuracy, 78.34% sensitivity, and 94.91% specificity and an MCC of 0.737. These results indicated that our predictor is a useful tool to predict RNA-binding proteins.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grant no. 61305072 and the Natural Science Foundation of the Jiangsu Higher Education Institutions of China under Grant no. 14KJB520020 and sponsored by Qing Lan project.

References

[1] M. Ibba and D. Soll, "Protein-RNA molecular recognition," *Nature*, vol. 381, no. 6584, p. 656, 1996.

[2] R. N. De Guzman, R. B. Turner, and M. F. Summers, "Protein-RNA recognition," *Biopolymers*, vol. 48, no. 2-3, pp. 181–195, 1998.

[3] S. Cusack, "Aminoacyl-tRNA synthetases," *Current Opinion in Structural Biology*, vol. 7, no. 6, pp. 881–889, 1997.

[4] S. M. Fernández-Moya and A. M. Estévez, "Posttranscriptional control and the role of RNA-binding proteins in gene regulation in trypanosomatid protozoan parasites," *Wiley Interdisciplinary Reviews: RNA*, vol. 1, no. 1, pp. 34–46, 2010.

[5] Y.-D. Cai and A. J. Doig, "Prediction of *Saccharomyces cerevisiae* protein functional class from functional domain composition," *Bioinformatics*, vol. 20, no. 8, pp. 1292–1300, 2004.

[6] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, New York, NY, USA, 1998.

[7] Y.-D. Cai and S. L. Lin, "Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence," *Biochimica et Biophysica Acta*, vol. 1648, no. 1-2, pp. 127–133, 2003.

[8] L. Y. Han, C. Z. Cai, S. L. Lo, M. C. M. Chung, and Y. Z. Chen, "Prediction of RNA-binding proteins from primary sequence by a support vector machine approach," *RNA*, vol. 10, no. 3, pp. 355–368, 2004.

[9] X. Yu, J. Cao, Y. Cai, T. Shi, and Y. Li, "Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines," *Journal of Theoretical Biology*, vol. 240, no. 2, pp. 175–184, 2006.

[10] X. Shao, Y. Tian, L. Wu, Y. Wang, L. Jing, and N. Deng, "Predicting DNA- and RNA-binding proteins from sequences with kernel methods," *Journal of Theoretical Biology*, vol. 258, no. 2, pp. 289–293, 2009.

[11] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "SVM based prediction of RNA-binding proteins using binding residues and evolutionary information," *Journal of Molecular Recognition*, vol. 24, no. 2, pp. 303–313, 2011.

[12] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[13] T. U. Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. 1, pp. D71–D75, 2012.

[14] C. R. Peng, L. Liu, B. Niu et al., "Prediction of RNA-binding proteins by voting systems," *Journal of Biomedicine and Biotechnology*, vol. 2011, Article ID 506205, 8 pages, 2011.

[15] X. Ma, J. Guo, J. Wu et al., "Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature," *Proteins*, vol. 79, no. 4, pp. 1230–1239, 2011.

[16] S. Ahmad and A. Sarai, "PSSM-based prediction of DNA binding sites in proteins," *BMC Bioinformatics*, vol. 6, article 33, 2005.

[17] S.-Y. Ho, F.-C. Yu, C.-Y. Chang, and H.-L. Huang, "Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM-PSSM method," *BioSystems*, vol. 90, no. 1, pp. 234–241, 2007.

[18] L. Wang, M. Q. Yang, and J. Y. Yang, "Prediction of DNA-binding residues from protein sequence information using random forests," *BMC Genomics*, vol. 10, supplement 1, article S1, 2009.

[19] J. Wu, H. Liu, X. Duan et al., "Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature," *Bioinformatics*, vol. 25, no. 1, pp. 30–35, 2009.

[20] X. Ma, J.-S. Wu, H.-D. Liu, X.-N. Yang, J.-M. Xie, and X. Sun, "SVM-based approach for predicting DNA-binding residues in proteins from amino acid sequences," in *Proceedings of the International Joint Conference on Bioinformatics, Systems Biology and Intelligent Computing*, pp. 225–229, Shanghai, China, August 2009.

[21] L. Wang, C. Huang, M. Q. Yang, and J. Y. Yang, "BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features," *BMC Systems Biology*, vol. 4, no. 1, article S3, 2010.

[22] Y.-F. Huang, L.-Y. Chiu, C.-C. Huang, and C.-K. Huang, "Predicting RNA-binding residues from evolutionary information and sequence conservation," *BMC Genomics*, vol. 11, supplement 4, article S2, 2010.

[23] Y. Murakami, R. V. Spriggs, H. Nakamura, and S. Jones, "PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences," *Nucleic Acids Research*, vol. 38, supplement 2, pp. W412–W416, 2010.

[24] 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.

[25] S. Ahmad, M. M. Gromiha, and A. Sarai, "Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information," *Bioinformatics*, vol. 20, no. 4, pp. 477–486, 2004.

[26] S. Ahmad and A. Sarai, "Moment-based prediction of DNA-binding proteins," *Journal of Molecular Biology*, vol. 341, no. 1, pp. 65–71, 2004.

[27] J. R. Bock and D. A. Gough, "Predicting protein—protein interactions from primary structure," *Bioinformatics*, vol. 17, no. 5, pp. 455–460, 2001.

[28] J. Wang, *Biochemistry*, Higher Education, 2002 (Chinese).

[29] R. E. Buntrock, "ChemOffice ultra 7.0," *Journal of Chemical Information and Computer Sciences*, vol. 42, no. 6, pp. 1505–1506, 2002.

[30] M. Rose, "Re: Balaban et al.—low volume bowel preparation for colonoscopy: randomized endoscopist-blinded trial of liquid sodium phosphate versus tablet sodium phosphate," *The American Journal of Gastroenterology*, vol. 98, no. 10, pp. 2328–2329, 2003.

[31] D. Bonchev, "The overall Wiener index—a new tool for characterization of molecular topology," *Journal of Chemical Information and Computer Sciences*, vol. 41, no. 3, pp. 582–592, 2001.

[32] 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.

[33] A. Liaw and M. Wiener, "Classification and regression by random forest," *R News*, vol. 2, no. 3, pp. 18–22, 2002.

[34] C. Zou, J. Gong, and H. Li, "An improved sequence based prediction protocol for DNA-binding proteins using SVM and comprehensive feature analysis," *BMC Bioinformatics*, vol. 14, article 90, 2013.

[35] Y.-F. Gao, B.-Q. Li, Y.-D. Cai, K.-Y. Feng, Z.-D. Li, and Y. Jiang, "Prediction of active sites of enzymes by maximum relevance minimum redundancy (mRMR) feature selection," *Molecular Biosystems*, vol. 9, no. 1, pp. 61–69, 2013.

[36] T. Gui, X. Dong, R. Li, Y. Li, and Z. Wang, "Identification of hepatocellular carcinoma-related genes with a machine learning and network analysis," *Journal of Computational Biology*, vol. 22, no. 1, pp. 63–71, 2015.

[37] B.-Q. Li, Y.-D. Cai, K.-Y. Feng, and G.-J. Zhao, "Prediction of protein cleavage site with feature selection by random forest," *PLoS ONE*, vol. 7, no. 9, Article ID e45854, 2012.

[38] B.-Q. Li, K.-Y. Feng, L. Chen, T. Huang, and Y.-D. Cai, "Prediction of protein-protein interaction sites by random forest algorithm with mRMR and IFS," *PLoS ONE*, vol. 7, no. 8, Article ID e43927, 2012.

[39] B.-Q. Li, L.-L. Hu, L. Chen, K.-Y. Feng, Y.-D. Cai, and K.-C. Chou, "Prediction of protein domain with mRMR feature selection and analysis," *PLoS ONE*, vol. 7, no. 6, Article ID e39308, 2012.

[40] X. Ma and X. Sun, "Sequence-based predictor of ATP-binding residues using random forest and mRMR-IFS feature selection," *Journal of Theoretical Biology*, vol. 360, pp. 59–66, 2014.

[41] J. Wang, D. Zhang, and J. Li, "PREAL: prediction of allergenic protein by maximum Relevance Minimum Redundancy (mRMR) feature selection," *BMC Systems Biology*, vol. 7, supplement 5, article S9, 2013.

[42] N. Zhang, Y. Zhou, T. Huang et al., "Discriminating between lysine sumoylation and lysine acetylation using mRMR feature selection and analysis," *PLoS ONE*, vol. 9, no. 9, Article ID e107464, 2014.

[43] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005.

[44] M. Treger and E. Westhof, "Statistical analysis of atomic contacts at RNA-protein interfaces," *Journal of Molecular Recognition*, vol. 14, no. 4, pp. 199–214, 2001.

[45] M. Terribilini, J.-H. Lee, C. Yan, R. L. Jernigan, V. Honavar, and D. Dobbs, "Prediction of RNA binding sites in proteins from amino acid sequence," *RNA*, vol. 12, no. 8, pp. 1450–1462, 2006.

[46] M. Kumar, M. M. Gromiha, and G. P. S. Raghava, "Prediction of RNA binding sites in a protein using SVM and PSSM profile," *Proteins: Structure, Function and Genetics*, vol. 71, no. 1, pp. 189–194, 2008.