PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

RBRIdent: An algorithm for improved identification of RNA-binding residues in proteins from primary sequences

Dapeng Xiong,[1] Jianyang Zeng,[2] and Haipeng Gong[1]*

[1] MOE Key Laboratory of Bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China

[2] Institute for Interdisciplinary Information Sciences, Tsinghua University, Beijing 100084, China

ABSTRACT

Rapid and correct identification of RNA-binding residues based on the protein primary sequences is of great importance. In most prevalent machine-learning-based identification methods; however, either some features are inefficiently represented, or the redundancy between features is not effectively removed. Both problems may weaken the performance of a classifier system and raise its computational complexity. Here, we addressed the above problems and developed a better classifier (RBRIdent) to identify the RNA-binding residues. In an independent benchmark test, RBRIdent achieved an accuracy of 76.79%, Matthews correlation coefficient of 0.3819 and F-measure of 75.58%, remarkably outperforming all prevalent methods. These results suggest the necessity of proper feature description and the essential role of feature selection in this project. All source data and codes are freely available at http://166.111.152.91/RBRIdent.

Proteins 2015; 83:1068–1077.

Key words: RNA-binding residues; sequence-based prediction; machine learning; feature selection.

INTRODUCTION

Proper RNA-protein interactions are critical in many biological processes, including protein synthesis, gene expression, post-transcriptional regulation, and viral replication.[1,2] Disruption of the physiological RNA-protein interactions; however, may cause numerous diseases, ranging from neurological disorders to cancer.[3,4] In principle, deciphering the molecular mechanism by which a protein uses a combination of amino acid residues to specifically recognize and discriminate its RNA partners can facilitate comprehending the functional implications of physiological RNA-protein interactions.[5] Meanwhile, efficient identification of RNA-binding residues can further promote the rational design of RNA drugs, which may block the pathogenic RNA-protein interaction *in vivo*.[6,7]

RNA-binding residues in proteins can be identified unambiguously from the three-dimensional structures of RNA-protein complexes derived using X-ray crystallography, electron microscopy, and nuclear magnetic resonance. Unfortunately, these experimental methods are both costly and time consuming. Considering the presence of strong relationship between interacting residues and their physicochemical properties in RNA-protein complexes,[8–10] the development of effective and reliable computational approaches to identify the RNA-binding residues has become in urgent demand. Currently available computer programs mainly fall into two broad categories: sequence-based methods and structure-based (also named as homology-based or template-based) ones.[11–14] The former is obviously more valuable because of the lack of requirement on structural information, considering the fact that the sequence data of proteins are accumulating in a much faster rate than structures deposited in the database.[15] Therefore, we only focus on the sequence-based methods in this study.

Several sequence-based approaches, including BindN,[16] RNABindR,[5] Pprint,[17] RNAProB,[18] PiRaNhA,[19] BindN+,[20] PRBR,[21] PRNA,[7] and meta-predictor,[13]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: National Natural Science Foundation of China; Grant numbers: 31470033, 31321062; Grant sponsor: National Basic Research Program of China; Grant numbers: 2011CBA00300, 2011CBA00301; Grant sponsor: National Natural Science Foundation of China; Grant numbers: 61033001, 61361136003, 61472205.

*Correspondence to: Haipeng Gong, MOE key laboratory of bioinformatics, School of Life Sciences, Tsinghua University, Beijing 100084, China.
E-mail: hgong@tsinghua.edu.cn

Received 29 January 2015; Revised 23 March 2015; Accepted 24 March 2015
Published online 3 April 2015 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24806

have been proposed to identify the RNA-binding residues in proteins with various levels of success. These methods generally extract various features from amino acid residues and use them as the input to train the machine-learning models for classification. However, the relative importance of those ever proposed features has not been evaluated uniformly. Furthermore, all existing methods contain at least one of the following two deficiencies. On one hand, some features are poorly described and are therefore unable to properly reflect the physicochemical and environmental properties of amino acid residues. On the other hand, the features are merely combined in the simplest way and their intrinsic overlaps are therefore not effectively removed. Both problems may impair the prediction power and restrict the space for further improvement in performance.

In light of all the previous caveats, in this study, we developed a new strategy, named RBRIdent (identification of RNA-binding residues), to improve the sequence-based identification of RNA-binding residues. In specific, we designed several novel features by statistically analyzing the interaction preferences between amino acid residues and their RNA partners from the structure database[5,7,9,22] and combined them with good features reported in the literature to compose a complete feature set, which was subsequently optimized using the genetic algorithm (GA) to a feature subset. Not only is the identification performance greatly improved after feature selection, but also the relative importance of all known features is roughly evaluated in such an analysis. Moreover, we constructed a completely independent benchmark testing dataset, which contains 60 nonredundant RNA-binding proteins, to objectively evaluate the methods. According to the results, RBRIdent markedly prevails all rival methods in the benchmark test.

MATERIALS AND METHODS

Datasets

The RNA-protein complexes used in this research were extracted from the Protein Data Bank (PDB)[23] released until August 20th, 2013. Only those structures solved by X-ray crystallography with a resolution better than 3 Å were retained. The redundant proteins with sequence identity $\geq$25% were removed using the BLASTCLUST program.[24] The amino acid residues at N- and C-termini as well as those adjacent to the chain gaps were uniformly abandoned both because they lack rigorous secondary structure definition and because many of them have missing atoms. After the above preprocessing, our dataset contains 281 nonredundant protein chains, which consist of 63,406 amino acid residues in total. The program ENTANGLE[25] was used to define the RNA-interacting residues in the proteins. Using the default parameters, 8060 and 55,346 of the amino acid residues were defined as binding (positive class) and nonbinding (negative class), respectively.

To better evaluate the performance of RBRIdent and all rival methods, the complete dataset was divided into two mutually independent parts: a training set for model optimization as well as cross-validation and a testing set for the objective benchmark test. Considering that an effective method should exhibit good predicting power for the presently unknown and undiscovered RNA-binding proteins, the 221 nonredundant protein chains released before the year 2011 were assigned to the training set (called RBP221), which contains 6406 binding and 44,854 nonbinding residues respectively, while the remaining 60 nonredundant protein chains released in the year 2011 and later were assigned to the testing set (called RBP60), which contains 1654 binding and 10,492 nonbinding residues, respectively.

Feature vectors

Mutual information (MI)

According to previous studies, the statistically derived propensity features have been proved to be highly valuable.[5,7] The mutual information (MI) was first introduced into the identification of RNA-binding residues in PRNA.[7] In this program, MI was obtained by statistically analyzing the interactions between a residue triplet and a nucleotide in the available structure database, and was reported to greatly improve the predicting power. In specific, a residue triplet was considered as RNA-binding if the middle residue was identified to physically interact with one nucleotide in an RNA-protein complex, according to the criterion of ENTANGLE. The MI was defined as:

$$MI(x,y) = \sum_{p,r} f_{p,r}(x,y) log_2\left(\frac{f_{p,r}(x,y)}{f_p(x)\cdot f_r(y)}\right), \quad (1)$$

$$f_{p,r}(x,y) = \frac{N_{p,r}(x,y)}{\sum_{x,y} N_{p,r}(x,y)}, \quad (2)$$

$$f_p(x) = \frac{N_p(x)}{\sum_x N_p(x)}, \quad (3)$$

$$f_r(y) = \frac{N_r(y)}{\sum_y N_r(y)}, \quad (4)$$

where x and y refer to a residue triplet and a nucleotide respectively, $f_{p,r}(x,y)$ is the joint probability of residue triplet x to interact with nucleotide y in the protein-RNA pair (p, r), $f_p(x)$ and $f_r(y)$ refer to the marginal probability of the residue triplet x in protein p and that of the nucleotide y

in RNA r respectively, and each N refers to the counting number in the database for the same type of f.

In the above protocol, each residue in the triplet x can be any of the 20 traditional amino acids and each nucleotide y can be one of the four classical ones, such that the triplet-nucleotide interacting pairs may have 32,000 (20 × 20 × 20 × 4) categories in total, which greatly outnumbers the available triplet-nucleotide interacting pairs (12,541) in the latest database and therefore impairs the statistical reliability in the probability estimation in Eq. (2). We modified this protocol by regrouping the two flanking residues in the triplet to five classes, based on the physicochemical properties: (a) positively charged residues: His, Lys, and Arg; (b) negatively charged residues: Asp and Glu; (c) polar residues: Asn, Gln, Ser, and Thr; (d) nonpolar alkyl residues: Ala, Cys, Gly, Ile, Leu, Met, Pro, and Val; (e) aromatic residues: Phe, Trp, and Tyr. In this way, the number of categories for the triplet-nucleotide interaction drops to 2000 (5 × 20 × 5 × 4), thereby enabling reliable estimation of MI. Notably, all MIs were calculated based purely on the training dataset and utilized no information from the benchmark testing dataset.

Interaction propensity (IP)

We expanded the original IP proposed in RNABindR[5] to 5 features so as to reflect the intrinsic tendencies of a residue to interact with an RNA molecule through hydrogen bonding, electrostatic, van der Waals, hydrophobic, and stacking interactions, respectively. They are accordingly named as hydrogen bonding interaction propensity (HBIP), electrostatic interaction propensity (EIP), van der Waals interaction propensity (VDWIP), hydrophobic interaction propensity (HIP), and stacking interaction propensity (SIP), respectively. These IPs were obtained in a similar way to that presented in RNABindR (here we only take HBIP as an example),

$$HBIP = log_2\left(\frac{f_{HBI}(x)}{f_{ED}(x)}\right), \tag{5}$$

where x refers to the residue type, while $f_{\mathrm{HBI}}(x)$ and $f_{\mathrm{ED}}(x)$ refer to the probabilities of the residue type x making hydrogen bonding interactions (HBI) with RNAs and in the entire dataset (ED), respectively. Decomposing the protein-RNA interactions into the five physical sources was achieved using ENTANGLE. EIP was finally abandoned because of the insufficient electrostatic interaction reported by ENTAGLE. All IPs were calculated purely from the training dataset and utilized no information from the benchmark testing dataset.

Other popular properties

Besides the features described above, we also incorporated several popular properties of amino acids published previously, in the belief that a combination of good features can better reflect the information encoded in the protein-RNA interaction and therefore improve the identification performance. These descriptors include number of atoms (NA), electrostatic charge (EC), hydrophobicity (Hy), relative accessible surface area (RASA), secondary structure (SS), smoothed position specific scoring matrix (Smoothed PSSM) and side-chain pKa value (pKa). The details are listed as follows:

NA and EC: the NA and EC values of amino acids were obtained from Li, *et al.*[26].

Hy: the Hy value of an amino acid was represented by the hydrophobic index designed by Sweet and Eisenberg.[27]

RASA: RASA was used to represent the solvent exposure of a residue. SABLE (version 2.0)[28] was used to predict the RASA of each residue considering its high performance.

SS: the SS of an amino acid was predicted by applying the powerful PSSpred program (http://zhanglab.ccmb.med.umich.edu/PSSpred) against the nonredundant database of protein sequence at NCBI (August 23rd, 2013 release). Here it is classified into three states: helix, extended structure and coil.

Smoothed PSSM: the PSSM was obtained using PSI-BLAST[24] search against the non-redundant database of protein sequences at NCBI as aforementioned. The substitution matrix, round of iteration and E value were set to BLOSUM62, 3 and 0.001, respectively. The original PSSM describes the evolutionary conservation at the corresponding residue position. Cheng *et al.* modified the original PSSM by substituting the PSSM value of each target residue by the sum of PSSM values of all residues within a preset window centered at the target one. Since Smoothed PSSM considers the effect of environmental residues during the evolution, it was reported to be superior to the standard one.[18] Therefore, we adopted the Smoothed PSSM in this work. The size of the preset window for the smoothing procedure (s) was optimized to 19 (see Results and Discussion).

pKa: the side-chain pKa value of an amino acid was used to represent its chemical characteristics. They were obtained from Nelson and Cox.[29]

A sliding window of odd size w was applied to encode the environmental information around the amino acid residue. The properties of all residues in the window were used to describe the target residue located at the center. The RNA-binding status of the center residue was attached to each window. Using the 12 descriptors mentioned above, the feature vectors were constructed by joining the properties of the w residues in the window. (1, 0, 0), (0, 1, 0), and (0, 0, 1) were used to represent the SS type of helix, extended structure and coil, respectively. For a Smoothed PSSM profile, each amino acid was encoded as a vector of 20 terms, each of which records the log-likelihood for substitution by a specific one of the overall 20 amino acids in evolution. The MI of each residue contains four columns to reflect the association propensities between the corresponding residue triplet and four types of nucleotides. In summary, a single candidate residue was represented as a feature vector of $w \times 12$ descriptors with $w \times 36$ elements altogether. Various sliding window sizes (w), from 1 to 19, were tested for optimization, and the final value was chosen as 9 (see Results and Discussion).

Selection of the random forest (RF) model

In this study, we chose RF as our classification paradigm due to its strong ability to find global classification solutions,[30] and successful application in bioinformatics and other practical domains.[7,31,32] RF is a system consisting of a collection of tree-structured classifiers,[30] which takes advantage of two powerful machine learning techniques: bagging[33] and random subspace. The original bagging requires that each tree is trained on a bootstrap sample of training data and that the predictions are made based on the majority votes of the trees. RF further refines bagging by selecting about two-thirds of the training samples for the construction of each tree and using the remaining samples to test the tree. The random subspace method randomly selects a subset of features to split at each node when growing a tree. This method can decrease the correlation between the trees in the forest and thereby reduce the forest error rate.

Here, the RF algorithm was implemented by the randomForest R package,[34] and the default parameters of this package were adopted. The RF model only outputs voting scores for the residues, and a tradeoff threshold is required for the final determination of positive and negative identifications. In other words, only when the RF voting score exceeds the threshold, is the residue predicted as RNA-binding. The optimal value for this threshold was chosen as the one that produced the best performance (or the highest F-measure, see the following sections for more details) in the training dataset.

Feature selection

Selecting the most discriminative set of features would increase the performance, efficiency and comprehensibility of a classifier system by reducing its redundancy and complexity.[35–37] In addition, the relative importance of all features in the identification of RNA-binding residues can be roughly estimated by finding the optimal feature subsets using feature selection. Notably, the importance of features cannot be evaluated by the RF model, since all features are represented as vectors containing at least w elements (w is the sliding window size). Here, GA[38] was chosen as our feature selection paradigm due to its strong random search ability to find the convincingly optimal feature subset.[39–41] The GA algorithm can randomly change the status of many features rather than one individual feature in every iteration, which enables the global optimization.[39–41]

The initial feature subset was generated randomly and was then updated through many iterations of selection, crossover, and mutation until convergence. The fivefold cross validation was used to test the power of feature subsets. The feature subset that achieves the best classification performance would be considered as the optimal one.

In specific, a binary string was chosen to encode the feature data of the population: 1 indicated that the corresponding feature was active, while 0 meant the reverse. The fitness function for selection was set as $f(x) = 10{,}000 \times$ F-measure, considering that F-measure has been chosen as our primary evaluation criterion because of its comprehensiveness (see the following section for more details). Each generation contained 100 individuals. The crossover probability, mutation probability and iteration number were set to 0.9, 0.1, and 100, respectively. We employed the classical proportion selection operator, and the two optimal individuals in every generation were directly passed to the next generation.

Performance evaluation

The performance of RBRIdent was evaluated against five popular sequence-based methods, including Pprint,[17] PiRaNhA,[19] BindN+,[20] PRNA,[7] and meta-predictor.[13] For a fair comparison, all methods were reimplemented using our own code and their parameters were optimized in the uniform training dataset RBP221. The correct reimplementation can be reflected from the agreement of performances reported in this study and those in the original literature. All tunable parameters remained unchanged when the methods were evaluated using the testing dataset RBP60. The original PRNA directly obtains the information of SS and RASA from the PDB structure and therefore utilizes the structural information. For a fair comparison, we fed the PRNA classifier with predicted SS and RASA values (using PSSpred and SABLE, respectively, see the previous section). The modified classifier is therefore called PRNA*.

Five measures were calculated for performance evaluation, including sensitivity (SN), specificity (SP), accuracy (ACC), Matthews correlation coefficient (MCC), and F-measure, as defined in the following equations:

$$\mathrm{SN} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{6}$$

$$\mathrm{SP} = \frac{\mathrm{TN}}{\mathrm{TN}+\mathrm{FP}}, \tag{7}$$

$$\mathrm{ACC}=\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FN}+\mathrm{TN}+\mathrm{FP}}, \tag{8}$$

$$\mathrm{MCC}=\frac{\mathrm{TP}\times\mathrm{TN}-\mathrm{FP}\times\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FN})(\mathrm{TP}+\mathrm{FP})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}, \tag{9}$$

$$F\mathrm{-measure}=\frac{2\times\mathrm{SN}\times\mathrm{SP}}{\mathrm{SN}+\mathrm{SP}}, \tag{10}$$

where TP, FN, TN, and FP refer to true positives, false negatives, true negatives, and false positives, respectively. SN and SP measure the proportion of correctly identified residues within the RNA-binding and nonbinding ones, respectively, while ACC estimates the overall identification accuracy of both interacting and noninteracting residues. MCC indicates the degree of correlation between the real and the predicted interacting status of the residues, and ranges between 1 (all predictions are correct) and −1 (none are correct). F-measure is the harmonic mean of SN and SP and is generally believed as a more comprehensive and effective evaluator.[7,42] Therefore, it was chosen as our primary evaluation criterion in this work.

RESULTS AND DISCUSSION

Parameter optimization using the complete feature set

To obtain the best performance, the sizes of the sliding window (*w*) and the smoothing window (*s*; for the Smoothed PSSM) were tested from 1 to 19, using the fivefold cross validation with the complete feature set on the training dataset. The best window size was optimized in terms of the primary evaluation criteria F-measure. Finally, the optimal *w* and *s* values were set to 9 and 19, respectively.

Figure S1, Supporting Information presents the performance derived from varied *s* values with a fixed *w* equal to 9 [Fig. S1(A), Supporting Information], and varied *w* values with a fixed *s* equal to 19 [Fig. S1(B), Supporting Information] on the training dataset. In Figure S1(A), the curve reaches the optimal performance at the largest *s* value, although the growth in performance becomes more gradual when *s* exceeds 15. The same trend has been observed in all other *w* values ever tested (data not shown) and therefore the optimal value of *s* was finally set to 19. In Figure S1(B), when *s* is fixed to its optimal value, F-measure grows slowly and monotonically with the rise of *w* values. Notably, more terminal residues (the $(w-1)/2$ residues on both ends) become invalid for identification when a larger sliding window size is chosen, since all feature vectors in the whole window centered at the target residue are required for prediction. Considering the tradeoff between wider application and better performance, we followed a compromised strategy to settle down the *w* value, which was chosen as the one with its F-measure closest to the mean of the minimal and maximal F-measures in the plot. Finally, the optimal value of *w* was set to 9. Interestingly, the performance of our program is almost independent of the choice of *w* values (in terms of F-measure) in the benchmark test (Fig. S2, Supporting Information), which supports our optimization strategy. Finally, using the optimal values of *w* and *s*, our program achieved an ACC of 78.67%, MCC of 0.4026 and F-measure of 77.26% (SN = 75.48%, SP = 79.12%) in the fivefold cross validation on the training dataset.

Table I
The Result of Feature Selection

NA	EC	Hy	RASA	SS	Smoothed PSSM	pKa	HBIP	SIP	HIP	VDWIP	MI
Y	Y			Y	Y			Y			Y

“Y” indicates the corresponding feature is retained in the optimal feature subset.

Further improvement by feature selection

In this study, we combined our own features (MI and IPs) with the good descriptors reported previously to represent the specific properties of protein residues in RNA-protein complexes. Although the integration of all features guarantee the sufficient coverage on information encoded by the protein-RNA interaction, the possible redundancy between features may reduce the classification performance and raise the computation complexity. Conventionally, the importance of features can be examined in a simple way, by measuring the performance after iteratively removing one feature from the set. As shown in Table SI, Supporting Information, this naïve protocol cannot effectively evaluate the feature importance here since all features except the Smoothed PSSM cause a similar level of performance loss after removal. Therefore, such complicate correlations between features can only be effectively processed by more advanced techniques.

In this work, to remove redundancy and to obtain the optimal feature set that gave the best results, feature selection was carried out using GA. As shown in Table I, the optimal feature subset is comprised of NA, EC, SS, Smoothed PSSM, SIP, and MI, among which only the Smoothed PSSM can be identified using the naïve protocol. Interestingly, two features (SIP and MI) proposed by us are retained in this optimal feature subset, which reinforces their strong discriminating power in the identification of the RNA-binding residues.

All 6 optimal features retained after feature selection reasonably reflect the physicochemical properties of protein-RNA interactions. The number of atoms (NA) effectively differentiates the various amino acids and is therefore roughly equivalent to the residue identity. The

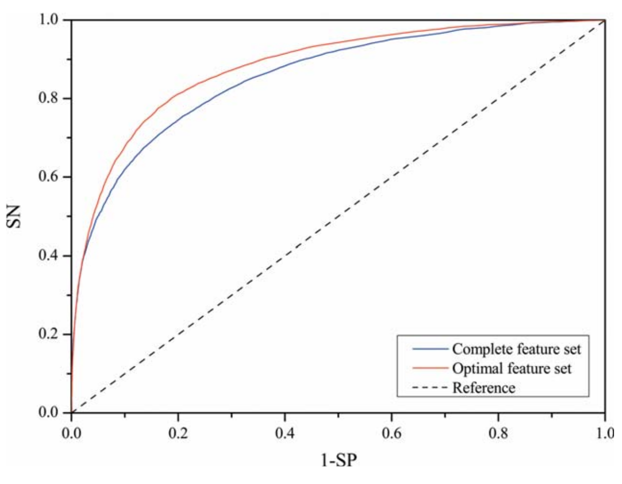

Figure 1

ROC curves to evaluate the performance using the complete (blue) and optimal (red) feature sets by fivefold cross validation in the training dataset. A random guess is plotted as reference (black and dotted). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

electrostatic charge (EC) is generally believed as an important feature around the protein-RNA binding interface, since positively charged residues are required to cluster on the protein binding surface so as to neutralize the RNA molecules that carry negative charges in the phosphor-ribose chain. The secondary structure (SS) reflects the local structural property and various SS elements (helix, strand, and coils) present different interacting patterns in recognizing the RNA molecules. The Smoothed PSSM represents the evolutionary information and describes the general residue substitution pattern of a segment of residues around the center one (in the smoothing window of size *s*). RNA binding residues are usually functionally important and therefore should be more conserved in evolution than the other ones. The stacking interaction propensity (SIP) reflects the tendency of planar side chains in the proteins to insert into the parallel neighboring bases in the RNAs by forming strong π–π stacking. As shown in Table SII, Supporting Information, the occurrence of such interactions is limited to residues with planar and conjugated side groups in both the training and testing dataset. Although this feature has never been reported in any previous researches, its retaining here indicates its critical role in the protein-RNA binding. The mutual information (MI) summarizes the association between a residue triplet in protein and a nucleotide in RNA from the structure database and has been proved to be essential in identifying the RNA-binding residues.[7] In contrast to SIP, the other three IPs (HBIP, HIP, and VDWIP) were unexpectedly excluded from the optimal feature subset, possibly because their corresponding interaction patterns have already been well represented in MI.

More importantly, apart from evaluating the feature contribution, the feature selection process can effectively improve the classification power of RBRIdent. The ACC, MCC, and F-measure increase from 78.67%, 0.4026 and 77.26% (SN = 75.48% and SP = 79.12%) in the complete feature set to 81.17%, 0.4597 and 80.67% (SN = 80.02% and SP = 81.34%) in the optimal feature subset, respectively. In contrast, the naïve protocol does not exhibit such capability (Table SI, Supporting Information), which reinforces the necessity of advanced feature selection in the systems containing heavy and complicate correlations between features. The receiver operating characteristic (ROC) curves for the performance evaluated using the complete and the optimal feature sets are shown in Figure 1. The area under the curve (AUC) values obtained through the complete and the optimal feature sets are 0.8570 and 0.8826, respectively. Clearly, the optimal feature subset exhibits stronger identification power and reduced computational complexity.

Comparison with other popular methods in the training set

In this section, the performance of our program was compared with several popular methods, including Pprint,[17] PiRaNhA,[19] BindN+,[20] PRNA,[7] and meta-predictor,[13] using the fivefold cross validation on the training dataset. The results are summarized in Table II. Notably, the SS and RASA information in the original PRNA program was extracted from the structure database. For a fair comparison, all structural information should be excluded. Therefore, the PRNA method was modified here to acquire the SS and RASA information from the predicting programs PSSpred and SABLE, respectively. The modified method is referred as PRNA*. As compared with PRNA, the power loss in PRNA* indicates that the application of structural information can effectively improve the classifier performance.

As shown in Table II, our method remarkably outperforms Pprint, PiRaNhA, BindN+, and meta-predictor in terms of every evaluation criterion, and exhibits a comparable power to PRNA*. In particular, both our method

Table II

Performance Comparison for All Methods Using Fivefold Cross-Validation in the Training Dataset

Method	SN (%)	SP (%)	ACC (%)	MCC	F-measure (%)
Pprint	71.32	73.83	73.52	0.3169	72.55
PiRaNhA	69.43	76.74	75.86	0.3305	72.90
BindN+	69.82	75.51	74.82	0.3234	72.55
PRNA	*83.60*	*80.51*	*80.90*	*0.4741*	*82.03*
PRNA*	**83.38**	77.62	78.34	0.4392	80.40
Meta-predictor	72.39	74.18	73.97	0.3253	73.27
RBRIdent	80.02	**81.34**	**81.17**	**0.4597**	**80.67**

The winner in each evaluating measure (column) is shown in bold. The original PRNA is excluded from comparison due to its inclusion of structural information and its results are shown in italic as a reference.

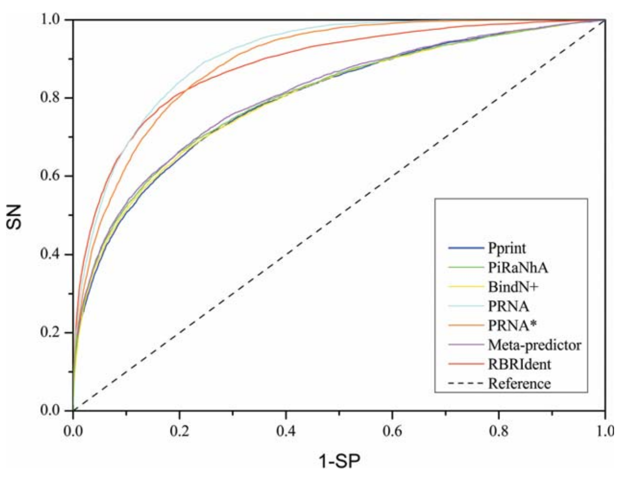

Figure 2

ROC curves to evaluate the performance for all methods using the five-fold cross validation in the training dataset. The curves for Pprint, PiRaNhA, BindN+, PRNA, PRNA*, meta-predictor, and RBRIdent are colored as blue, green, yellow, cyan, orange, magenta, and red, respectively. A random guess is plotted as reference (black and dotted).

and PRNA* exhibit advantages of >7% in the F-measure. The same trend is further confirmed in the ROC analysis (Fig. 2), where the AUC values for Pprint, PiRaNhA, BindN+, PRNA*, meta-predictor and RBRIdent are 0.7969, 0.8016, 0.7977, 0.8902, 0.8057, and 0.8826, respectively. Since the other methods do not have MI in their feature sets, we speculate that the superiority of RBRIdent and PRNA* may arise from the inclusion of MI.

Evaluation in the benchmark test

Considering that some statistical features (e.g., MI and IPs) were obtained from the overall training dataset, cross validation within the training dataset may be incapable of fairly evaluating the performance of various methods. Therefore, all methods were further evaluated in a completely independent benchmark testing dataset (RBP60). For each model, all tunable parameters remained unchanged from their best-performance values obtained in the training dataset.

As shown in Table III, our method yields an ACC of 76.79%, MCC of 0.3819 and F-measure of 75.58% (SN = 74.02% and SP = 77.22%), respectively. In terms of both MCC and F-measure values (our primary criterion), RBRIdent remarkably prevails all rival methods (by >0.02 in MCC and >2% in F-measure, respectively). When evaluated using ACC, our method ranks the second and is only slightly worse than PiRaNhA. In general, the comparison results demonstrate that RBRIdent achieves better and more reliable performance for identifying the RNA-binding residues in proteins, considering the comprehensiveness of F-measure as our primary evaluation criterion. Note that 2.12% increase in the F-measure is nontrivial and already indicates significant improvement, since this criterion reflects the general performance of an algorithm. To validate this statement, we performed a statistical test using a bootstrap protocol. In specific, we randomly removed 10 proteins from the testing dataset that contains 60 proteins in total and reevaluated the F-measures. By repeating the process for five times, we finally derived a distribution of F-measures, which allow the statistical test for significance using the student *t* test. As shown in Table SIII, Supporting Information, our method is significantly better than all other methods. In addition, Pprint, PiRaNhA, BindN+, and meta-predictor show comparable behavior here, with PiRaNhA performing slightly better than the others.

Notably, the performance of PRNA* (and PRNA) drops tremendously in the benchmark test as compared with the cross validation in the training set (cf. Tables II and III). In particular, the F-measure decreases by >15%, indicating the presence of over fitting in this method. The over fitting actually arises from the improper treatment on MI, since all methods without using the MIs (Pprint, PiPaNhA, BindN+, and meta-predictor) exhibit steady performance in the training and testing datasets. In PRNA, the residue triplet was constructed using all possible residue combination, which leads to overall 32000 (20 × 20 × 20 × 4) categories of triplet-nucleotide interacting pairs. Unfortunately, there were only 10,337 triplet-nucleotide interacting pairs in the complete structure dataset used in this method, a number significantly smaller than the number of categories. Consequently, the joint probability $f_{p,r}(x,y)$ [Eq. (2)] becomes too sparse to be reliably estimated with statistical significance, and the errors will transfer to MI automatically [Eq. (1)]. Indeed, even the latest dataset only includes 12,541 triplet-nucleotide interacting pairs, which can hardly guarantee the correct estimation of MI using the processing protocol of PRNA. In the PRNA article, the problem was unintentionally covered by data processing of MI, which was calculated from the complete (training + testing) dataset and inevitably caused information leak from the testing dataset into model training.

Table III

Performance Comparison for All Methods in the Benchmark Testing Dataset

Method	SN (%)	SP (%)	ACC (%)	MCC	F-measure (%)
Pprint	72.81	70.05	70.41	0.3030	71.40
PiRaNhA	69.20	**78.28**	**77.10**	0.3553	73.46
BindN+	68.57	76.08	75.09	0.3307	72.13
PRNA	*55.03*	*80.51*	*77.10*	*0.2822*	*65.37*
PRNA*	53.80	78.08	74.83	0.2467	63.70
Meta-predictor	72.40	74.16	73.93	0.3352	73.27
RBRIdent	**74.02**	77.22	76.79	**0.3819**	**75.58**

The winner in each evaluating measure (column) is shown in bold. The original PRNA is excluded from comparison due to its inclusion of structural information and its results are shown in italic as a reference.

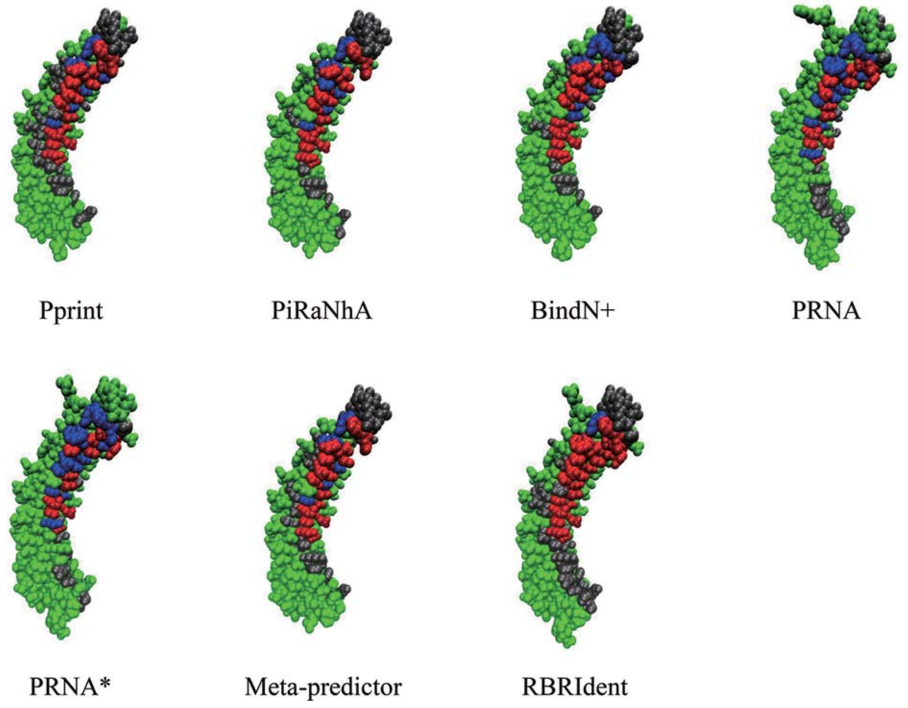

Figure 3

An example of predicted RNA-binding residues for the Puf-domain containing protein 7. All residues are shown in the ball representation. The true positives, false negatives, true negatives, and false positives are colored in red, blue, green, and gray, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

In our method, the over fitting problem is greatly alleviated by effectively reducing the triplet types, although such treatment may impair the discriminative power of MIs. After regrouping, the number of categories for the triplet-nucleotide interaction shrinks to 2000 (5 × 20 × 5 × 4), which can roughly handle the 9964 data points in the training dataset and therefore produces significantly less errors in the estimation of joint probability $f_{p,r}(x,y)$. Accordingly, the performance of RBRIdent drops only mildly (by ~4% in F-measure) in the benchmark test. The loss in performance may be completely rescued in the near future when more structures of RNA-protein complexes are deposited into the PDB database.

A case study

Figure 3 shows a case study on the prediction of RNA-binding residues in the Puf-domain containing protein 7 (PDB ID 3V71, chain A), which belongs to the testing dataset RBP60. For this protein, the achieved ACC, MCC, and F-measure are 79.61%, 0.3667, and 78.36% (SN = 76.92% and SP = 79.86%), respectively in Pprint, 86.30%, 0.4474 and 80.69% (SN = 75.00% and SP = 87.31%), respectively in PiRaNhA, 88.06%, 0.4750 and 80.43% (SN = 73.08% and SP = 89.44%), respectively in BindN+, 88.24%, 0.3792, and 69.01% (SN = 55.56% and SP = 91.05%), respectively in PRNA, 88.62%, 0.3503 and 63.26% (SN = 48.15% and SP = 92.18%), respectively in PRNA*, 84.59%, 0.4394 and 82.01% (SN = 79.17% and SP = 85.07%), respectively in meta-predictor, as well as 84.73%, 0.5202 and 89.57% (SN = 96.30% and SP = 83.71%), respectively in RBRIdent. Although RBRIdent identifies more nonbinding residues as binding ones (large false positives), it rarely mistakenly assigns the RNA-binding residues (small false negatives), which on average guarantees its best performance on this protein target. The low false negative rate achieved by RBRIdent may bring further benefits to experimental community since fewer RNA-binding residues are omitted by this method. The detailed sites predicted by different methods are listed in the Supporting Information. In addition, proteins generally use a continuous surface for RNA recognition. This behavior, however, cannot be properly reflected in the crystal structure that only contains a small fragment of RNA chain. Notably, RBRIdent is the only method that exhibits such behavior in the prediction: all predicted residues (true positives + false positives) compose a concave and continuous surface to favor the binding of a large RNA molecule rather than a small fragment.

As stated above, PiRaNhA ranks second in the independent benchmark test (Table III). The head-to-head comparison between our method and PiRaNhA is shown in Figure S3, Supporting Information where our method

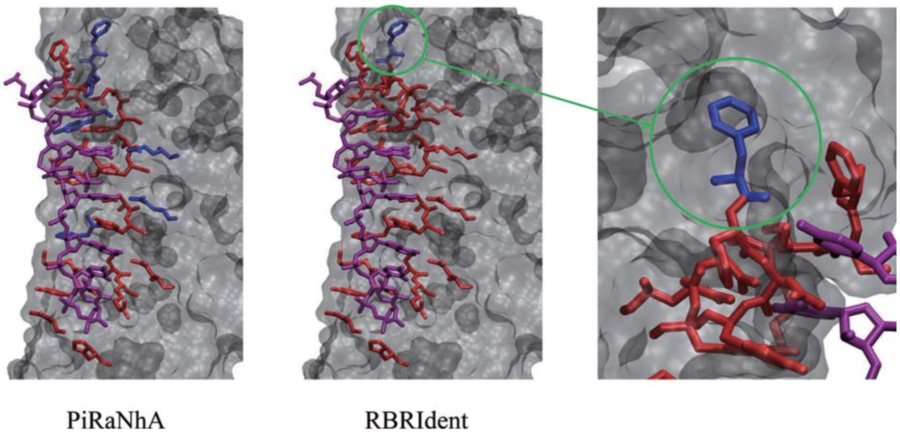

Figure 4

Comparison of the RNA-binding residues in an example (the Puf-domain containing protein 7) identified by RBRIdent and the best rival method PiRaNhA. The RNA molecule and the RNA-binding residues are both shown in the Licorice representation while the overall protein is represented as a transparent and gray surface. The RNA, true positives (correctly identified RNA-binding residues) and false negatives (unidentified RNA binding residues) are colored in purple, red and blue, respectively. [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

prevails in most protein targets within the testing dataset. Here, we also made a close-up comparison between PiRaNhA and RBRIdent on this specific protein target. As shown in Figure 4, among all RNA-binding residues of this protein, the six failures in the PiRaNhA identification (blue in the left panel) interact with the RNA molecule from nearly all directions. As a contrast, only one of these residues (blue in the middle and right panels), which loosely interacts with one terminus of the RNA, cannot be correctly predicted by RBRIdent. The better identification of RNA-binding residues in this case may arise from the introduction of MI and SIP in our method, both of which are derived by statistically analyzing the RNA-protein interacting pattern in the structure database. Current database is insufficient for the accurate estimation of MI, which probably leads to the high rate of false positives. With the accumulation of structure deposits in the database in the future, the performance of our method is expected to be further improved.

CONCLUSION

In this work, we presented a better and more reliable method, RBRIdent, for identifying the RNA-binding residues in proteins from the primary sequences. For a target protein that is known for RNA-binding, the program takes the protein sequence as the input and predicts the RNA-interacting residues. We proposed novel discriminating features to describe the interaction between residues in protein and nucleotides in RNA, and combined them with other good features reported in previous researches. In addition, to our knowledge, the feature selection was, for the first time, applied to this kind of study to reduce the feature space and to further improve performance. We also conducted an objective benchmark test to evaluate the performance of our method as well as five prevalent ones. In the test, RBRIdent prevails all rival methods. Future work is stilled needed to further reduce the rate of false positives and negatives.

ACKNOWLEDGMENTS

The authors gratefully thank Sai Zhang for the unstated trial of model building using the deep learning technique. The authors declare no conflict of interest.

REFERENCES

1. Chen Y, Varani G. Protein families and RNA recognition. FEBS J 2005;272:2088–2097.
2. Glisovic T, Bachorik JL, Yong J, Dreyfuss G. RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 2008;582:1977–1986.
3. Cooper TA, Wan L, Dreyfuss G. RNA and disease. Cell 2009;136:777–793.
4. Lukong KE, Chang K-W, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. Trends Genet 2008;24:416–425.
5. Terribilini M, Lee J-H, Yan C, Jernigan RL, Honavar V, Dobbs D. Prediction of RNA binding sites in proteins from amino acid sequence. RNA 2006;12:1450–1462.
6. Wang C, Fang Y, Xiao J, Li M. Identification of RNA-binding sites in proteins by integrating various sequence information. Amino Acids 2011;40:239–248.
7. Liu Z, Wu L, Wang Y, Zhang X, Chen L. Prediction of protein–RNA binding sites by a random forest method with combined features. Bioinformatics 2010;26:1616–1622.
8. Ellis JJ, Broom M, Jones S. Protein–RNA interactions: structural analysis and functional classes. Proteins: Struct Funct Bioinformatics 2007;66:903–911.

9. Kim OTP, Yura K, Go N. Amino acid residue doublet propensity in the protein–RNA interface and its application to RNA interface prediction. Nucleic Acids Res 2006;34:6450–6460.
10. Doherty EA, Batey RT, Masquida B, Doudna JA. A universal mode of helix packing in RNA. Nat Struct Mol Biol 2001;8:339–343.
11. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, Honavar V. RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of predicted RNA-binding residues in proteins. PLoS One 2014;9:e97725
12. Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 2012;13:89.
13. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein–RNA interactions. J Struct Biol 2012;179:261–268.
14. Li S, Yamashita K, Amada KM, Standley DM. Quantifying sequence and structural features of protein–RNA interactions. Nucleic Acids Res 2014;42:10086–10098.
15. Wang L, Brown SJ. Prediction of RNA-binding residues in protein sequences using support vector machines. Conf Proc IEEE Eng Med Biol Soc 2006;1:5830–5833.
16. Wang L, Brown SJ. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 2006;34(Suppl 2):W243–W248.
17. Kumar M, Gromiha MM, Raghava GPS. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins: Struct Funct Bioinformatics 2008;71:189–194.
18. Cheng C-W, Su E, Hwang J-K, Sung T-Y, Hsu W-L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 2008;9(Suppl 12):S6.
19. Spriggs RV, Murakami Y, Nakamura H, Jones S. Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics 2009;25:1492–1497.
20. Wang L, Huang C, Yang M, Yang J. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 2010;4:S3.
21. Ma X, Guo J, Wu J, Liu H, Yu J, Xie J, Sun X. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins: Struct Funct Bioinformatics 2010;79:1230–1239.
22. Kim H, Jeong E, Lee S-W, Han K. Computational analysis of hydrogen bonds in protein–RNA complexes for interaction patterns. FEBS Lett 2003;552:231–239.
23. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. Nucleic Acids Res 2000;28:235–242.
24. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
25. Allers J, Shamoo Y. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J Mol Biol 2001;311:75–86.
26. Li N, Sun Z, Jiang F. Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC Bioinformatics 2008;9:553.
27. Sweet RM, Eisenberg D. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. J Mol Biol 1983;171:479–488.
28. Wagner M, Adamczak R, Porollo A, Meller J. Linear regression models for solvent accessibility prediction in proteins. J Comput Biol 2005;12:355–369.
29. Nelson DL, Cox MM. Amino acids, peptides, and proteins. Lehninger principles of biochemistry, 4th ed. New York: W.H. Freeman; 2004. pp 75–115.
30. Breiman L. Random forests. Mach Learn 2001;45:5–32.
31. Kandaswamy KK, Chou K-C, Martinetz T, Möller S, Suganthan PN, Sridharan S, Pugalenthi G. AFP-pred: a random forest approach for predicting antifreeze proteins from sequence-derived properties. J Theor Biol 2011;270:56–62.
32. Jiang P, Wu H, Wang W, Ma W, Sun X, Lu Z. MiPred: classification of real and pseudo microRNA precursors using random forest prediction model with combined features. Nucleic Acids Res 2007; 35(Suppl 2):W339–W344.
33. Breiman L. Bagging predictors. Mach Learn 1996;24:123–140.
34. Liaw A, Wiener M. Classification and regression by random Forest. R News 2002;2:18–22.
35. Saeys Y, Inza I, Larrañaga P. A review of feature selection techniques in bioinformatics. Bioinformatics 2007;23:2507–2517.
36. Liu H, Yu L. Toward integrating feature selection algorithms for classification and clustering. IEEE Trans Knowl Data En 2005;17: 491–502.
37. Guyon I, Elisseeff A. An introduction to variable and feature selection. J Mach Learn Res 2003;3:1157–1182.
38. Holland J. Genetic algorithms. Sci Am 1992;267:66–72.
39. Huang C-L, Wang C-J. A GA-based feature selection and parameters optimizationfor support vector machines. Expert Syst Appl 2006;31: 231–240.
40. Raymer ML, Punch WF, Goodman ED, Kuhn LA, Jain AK. Dimensionality reduction using genetic algorithms. IEEE Trans Evol Comput 2000;4:164–171.
41. Leardi R, Boggia R, Terrile M. Genetic algorithms as a strategy for feature selection. J Chemometr 1992;6:267–281.
42. Pizzuti C, Rombo S. Restricted neighborhood search clustering revisited: an evolutionary computation perspective. In: Ngom A, Formenti E, Hao J-K, Zhao X-M, van Laarhoven T, editors. Pattern recognition in bioinformatics. Vol.7986, Lecture Notes in Computer Science: Springer Berlin Heidelberg; 2013. pp 59–68.