BIOINFORMATICS ORIGINAL PAPER Vol. 26 no. 13 2010, pages 1616–1622
doi:10.1093/bioinformatics/btq253

Structural bioinformatics

Advance Access publication May 18, 2010

Prediction of protein–RNA binding sites by a random forest method with combined features

Zhi-Ping Liu[1], Ling-Yun Wu[2], Yong Wang[2], Xiang-Sun Zhang[2] and Luonan Chen[1,*]

[1]Key Laboratory of Systems Biology, SIBS-Novo Nordisk Translational Research Centre for PreDiabetes, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, Shanghai 200031 and [2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China

Associate Editor: Dmitrij Frishman

ABSTRACT

Motivation: Protein–RNA interactions play a key role in a number of biological processes, such as protein synthesis, mRNA processing, mRNA assembly, ribosome function and eukaryotic spliceosomes. As a result, a reliable identification of RNA binding site of a protein is important for functional annotation and site-directed mutagenesis. Accumulated data of experimental protein–RNA interactions reveal that a RNA binding residue with different neighbor amino acids often exhibits different preferences for its RNA partners, which in turn can be assessed by the interacting interdependence of the amino acid fragment and RNA nucleotide.

Results: In this work, we propose a novel classification method to identify the RNA binding sites in proteins by combining a new interacting feature (interaction propensity) with other sequence- and structure-based features. Specifically, the interaction propensity represents a binding specificity of a protein residue to the interacting RNA nucleotide by considering its two-side neighborhood in a protein residue triplet. The sequence as well as the structure-based features of the residues are combined together to discriminate the interaction propensity of amino acids with RNA. We predict RNA interacting residues in proteins by implementing a well-built random forest classifier. The experiments show that our method is able to detect the annotated protein–RNA interaction sites in a high accuracy. Our method achieves an accuracy of 84.5%, *F*-measure of 0.85 and AUC of 0.92 prediction of the RNA binding residues for a dataset containing 205 non-homologous RNA binding proteins, and also outperforms several existing RNA binding residue predictors, such as RNABindR, BindN, RNAProB and PPRint, and some alternative machine learning methods, such as support vector machine, naive Bayes and neural network in the comparison study. Furthermore, we provide some biological insights into the roles of sequences and structures in protein–RNA interactions by both evaluating the importance of features for their contributions in predictive accuracy and analyzing the binding patterns of interacting residues.

Availability: All the source data and code are available at http://www.aporc.org/doc/wiki/PRNA or http://www.sysbio.ac.cn/datatools.asp

Contact: lnchen@sibs.ac.cn

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on February 5, 2010; revised and accepted on May 13, 2010

*To whom correspondence should be addressed.

1 INTRODUCTION

RNA undergoes diverse post-transcriptional regulation of gene expression, including regulation of its transportation, localization and decay (Glisovic *et al*., 2008). In many cases, such a process occurs through elements on the RNA molecule that interact with hundreds of RNA binding proteins existing in the cell (Lunde *et al*., 2007). Interactions among proteins and RNA molecules play an essential role in a variety of biological activities within a cell, such as the post-transcriptional gene regulation, alternative splicing, translation and infections by RNA viruses (Terribilini *et al*., 2006). Therefore, it is important to understand the principle of protein–RNA interactions and identify their interaction sites when selecting activators and inhibitors in rational drug design. It is commonly believed that RNA recognition by proteins is primarily mediated by certain classes of RNA binding domains and motifs (Morozova *et al*., 2006; Shulman-Peleg *et al*., 2008). Specifically, the correlated pattern of sequence and structure in a RNA binding protein can be recognized and bound by a specific RNA sequence, and then the resulting protein–RNA complex performs important functions somewhere.

In recognition of RNA functional importance in living molecules and close association with protein in its activities, experimental and computational studies of protein–RNA complexes have been substantially increased (Ellis *et al*., 2007; Hall, 2002; Jones *et al*., 2001). Recently, a variety of approaches have been proposed to study protein–RNA interactions (Lunde *et al*., 2007). Though some improvement has been obtained, the precise mechanisms of the protein–RNA interaction are far from being fully understood. In particular, there is no efficient way to experimentally identify protein–RNA binding sites that play a central role in forming protein–RNA interactions (Chen and Lim, 2008; Terribilini *et al*., 2006). Hence, it is strongly demanded to develop a reliable computational method to accurately predict protein–RNA interacting sites by exploring the accumulated data of protein–RNA complexes.

Many studies indicate that there is a strong relationship between interaction residues and their compositions in protein–RNA complexes (Doherty *et al*., 2001; Ellis *et al*., 2007; Kim *et al*., 2006). A number of machine learning techniques have been applied to detect RNA binding residues in protein sequences. For instance, BindN (Wang and Brown, 2006) uses an SVM-based classifier to predict potential RNA or DNA binding residues in proteins by sequence features. RNAProB (Cheng *et al*., 2008)

and PPRint (Kumar *et al*., 2008) implement promising SVM-based methods based on evolutionary profiles in the identification. RNABindR (Terribilini *et al*., 2007) generates a naive Bayes classifier to predict RNA binding amino acid residues in proteins. Jeong *et al*. (2004) proposed a neural network method for predicting RNA binding sites by using amino acid and secondary structure elements. Spriggs *et al*. (2009) improved the prediction performance using four sequence properties by a SVM-based classifier (see Supplementary Table S3 for the summary). However, most of the methods have not considered the interaction features underlying the partnership between protein residue and RNA nucleotide in the identification process. Actually, various interaction features extracted from the interacting sites give us valuable information to understand how a protein interacts with a RNA. If integrated together, these features can describe the interacting patterns in a more comprehensive way. On the other hand, for many existing methods such as RNABindR (Terribilini *et al*., 2007), although the interface propensity of residue with RNA is designed to describe interactions, it is simply defined as the proportion of a given amino acid in interaction sites divided by the proportion of the residue in the dataset. In this article, by noting that amino acids with different neighbor amino acids or in different local structures often exhibit preferences for their RNA partners (Kim *et al*., 2003, 2006; Terribilini *et al*., 2006), we design a new feature, i.e. interaction propensity of an amino acid, which considers the neighbors of the amino acid simultaneously instead of examining itself alone, based on the message transforming between a protein residue with its neighbors and the interacting nucleotides of RNA molecule. Moreover, we consider sequence and structure features of a residue which are also the contributors of the RNA binding events (Morozova *et al*., 2006). Clearly, all of those information are indispensable descriptors of the interaction patterns between protein residues and RNA nucleotide from different aspects, and thereby are expected to have strongly representative power as a combined feature when integrated together.

In the present work, we propose a novel random forest (RF) method for predicting RNA binding sites in proteins from both sequence and structure features. The RF (Breiman, 2001) is an ensemble classifier with many decision trees, which has many advantages, e.g. producing a highly accurate classifier, handling a very large number of input variables and estimating the importance of variables in determining classification (Breiman, 2001). We compute the mutual interaction propensity between amino acids and nucleotides by using representative protein–RNA complexes in PDB (Berman *et al*., 2000). In addition to the interaction propensity feature, several important residue properties have been combined as hybrid features. The cross-validation test has confirmed the effectiveness of our method. The comparison study shows that the accuracy (ACC) of our method is better than previous methods such as RNABindR, BindN and PPRint, and achieves an ACC of 84.5%, F-measure of 0.85 and area under curve (AUC) of 0.92 prediction of the RNA binding residues for a test dataset. Computational experimental results also show that the proposed interaction propensity is a very powerful feature for predicting RNA binding sites in proteins. Moreover, we identified the importance of individual features and their combinations in contributing both the specificity (SP) of protein–RNA binding and the ACC of the prediction, which provides biological insights into the roles of sequences and structures in protein–RNA interactions.

2 MATERIALS AND METHODS

2.1 Datasets

The protein–RNA complexes used in our experiments were downloaded from RsiteDB (Shulman-Peleg *et al*., 2008) which were retrieved from the Nucleic Acid Database (Berman *et al*., 1992). In total, there are 339 complexes with resolution better than 3Å and 1182 protein–RNA chains of these complexes. After removing protein and RNA chains with sequence identity above 25% and 60%, respectively (Altschul *et al*., 1997), we got 205 non-redundant protein–RNA chains in 164 complexes. We used ENTANGLE (Allers and Shamoo, 2001) to define the interaction sites between protein chain and RNA. Protein–RNA interactions that include hydrogen bonding, electrostatic, hydrophobic and van der Waals interactions (Allers and Shamoo, 2001) have been considered. Residues in protein–RNA interface were extracted as the interacting residues. We identified the interacting pairs between amino acids in protein chains and their partner nucleotides. Totally, 5261 (9.87%) amino acid residues were defined as the RNA binding sites from the total 53 315 residues, the rest 48 054 (90.13%) amino acids were defined as non-binding protein residues. The RNA binding chains are listed in the Supplementary Materials.

2.2 Interaction propensity

We identify and quantify the mutual dependence between the protein residues and RNA nucleotide by calculating a new measure, i.e. interaction propensity. We highlight the important role of the nearest neighbor residues in determining the SP of biochemical features and the preference of interacting with nucleotides for an amino acid residue (Terribilini *et al*., 2006; Wang *et al*., 2007). Hence, we define the mutual interaction propensity of a residue triplet and a nucleotide. A triplet is regarded as interacting with a nucleotide when its central residue interacts with the nucleotide. The mutual interaction propensity is defined as follows:

$$S(x,y)=\sum_{p,r} f_{p,r}(x,y)\log_2\frac{f_{p,r}(x,y)}{f_p(x)f_r(y)},$$

where x represents a residue triplet, y represents a nucleotide (i.e. $y\in\{A,G,C,U\}$). $f_{p,r}(x,y)=N_{p,r}(x,y)/\sum_{x,y}N_{p,r}(x,y)$ represents the frequency of x interacting y in the protein–RNA pair (p,r), where $N_{p,r}(x,y)$ is the number of residue triplet x binding to nucleotide y and $\sum_{x,y}N_{p,r}(x,y)$ is the total number of residue triplet and nucleotide pairs in the protein–RNA pair (p,r). $f_p(x)=N_p(x)/\sum_x N_p(x)$ represents the frequency of the residue triplet x in protein p, where $N_p(x)$ is the number of residue triplet x and $\sum_x N_p(x)$ is the total number of all residue triplets in the protein p. Similarly, $f_r(y)=N_r(y)/\sum_y N_r(y)$ represents the frequency of a nucleotide y, where $N_r(y)$ is the number of nucleotide y and $\sum_y N_r(y)$ is the total number of nucleotides in the RNA r. The interaction propensity of a triplet x and a nucleotide y is calculated on all interacting protein–RNA pairs in the dataset.

2.3 Descriptors for amino acid residues

Given the dataset, we identify the binding SP between all the existing triplets and four RNA nucleotides. A protein sequence of length l residues corresponds to $l-2$ triplets and every triplet will get its corresponding values of interaction propensity with four types of

nucleotides individually. Thus, each residue of $l-2$ centers in the triplets is described by a 4D vector.

In addition to the identified mutual interaction propensity between amino acid triplets and nucleotides, we also encode other properties of amino acids to describe their RNA binding SP. Each amino acid residue in the residue neighbor profile is characterized by six descriptors including physicochemical characteristics, hydrophobic index, relative accessible surface area, secondary structure, sequence conservation score and side-chain environment. The following list is the details of these descriptors for amino acid residues.

- Physicochemical characteristics: the physicochemical features of an amino acid residue are described by three values: number of atoms, number of electrostatic charge and number of potential hydrogen bonds (Li *et al*., 2008). These values are only related to the type of amino acid and do not contain any structural information of the amino acid residue.
- Hydrophobicity: the hydrophobicity of an amino acid residue is described by the hydrophobic index designed in Sweet and Eisenberg (1983).
- Relative accessible surface area: the accessible surface area of an amino acid is calculated by DSSP program (Kabsch and Sander, 1983). Then we calculate the property by dividing the accessible surface area with the accessible surface area of fully exposed amino acid. The accessible surface areas of the fully exposed amino acids are based on Rost and Sander (1994).
- Secondary structure: the secondary structure of an amino acid residue is also calculated by DSSP (Kabsch and Sander, 1983). It is divided into three states: helix, sheet and coil. DSSP secondary structure types I, G and H are considered as helix; types E and B are considered as sheet; types T, S and blank are considered as coil. We use (1, 0, 0), (0, 1, 0) and (0, 0, 1) to represent the candidate residue that belongs to three types of secondary structures, respectively.
- Conservation score: the values of sequence conservation for amino acid residues are obtained by PSI-BLAST (Altschul *et al*., 1997) search of the protein chain sequence in the Uniprot database (UniProt Consortium, 2008). The round of iteration is set to 3. The result of the PSI-BLAST search is a position-specific scoring matrix (PSSM). We extract the diagonal value of each residue as the value of its sequence conservation.
- Side-chain environment: pKa value of an amino acid side chain is an important factor in determining environmental characteristics of a protein. The side-chain pKa values are obtained from Nelson and Cox (2004) representing protein side-chain environmental properties and are widely used (Wang and Brown, 2006).

2.4 Encoding scheme

In this study, we employ the sliding window technique to encode the amino acid residues of proteins. We use the windows of odd number of residues with size s in the encoding scheme. Whether a residue belongs to the RNA binding class or not is determined by the middle residue (itself) and its neighbor $s-1$ residue profile. The feature vector representing the residue in the window is encoded by joining the properties of the s residues. As a result a single candidate, residue is represented in a feature vector of $s\times 7$ descriptors with $s\times 15$ feature elements. We tested various window sizes and choose the best one, i.e. $s=5$ (Table 1). When we evaluate the importance of these features by choosing a proper combination of descriptors, the number of elements in the feature vector is changed correspondingly.

2.5 RF for training and prediction

We formulate the RNA binding residue as a binary classification problem, where a residue is labeled with 1 if it is a RNA binding site or 0 otherwise. RF is a classification algorithm combining ensemble tree-structured classifiers (Breiman, 2001), which is often used when we have a very large training dataset and a very large number of input features. A typical RF model is made up hundreds of decision trees, and the major votes will determine the final prediction. The RF algorithm is implemented by the R randomForest package (Liaw and Wiener, 2002), where we use the default parameters of the package in this article.

The prediction is evaluated by the 5-fold cross-validation. The whole dataset was randomly partitioned into five groups with approximately equal size. To ensure that the training process is completely independent of the test data, the classifier is trained on the four groups and tested on the remaining group and each of them is chosen for assessment one by one. The predictive results are evaluated by different measures, i.e. sensitivity (SN), SP, ACC, F-measure and Matthews correlation coefficient (MCC). Mathematically, they are defined by the following equations:

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

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

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

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

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

where TP, FN, FP, and TN are the numbers of true positive, false negative, false positive and true negative residues in the prediction, respectively. SN and SP are also used to plot the receiver operating characteristic (ROC) curves and AUC is also calculated. F-measure is the harmonic mean of SN and SP. Usually the best F-measure point is chosen as the cutoff for SN and SP in ROC curves. MCC value ranges between 1 (all predictions are correct) and -1 (none are correct).

3 RESULTS AND DISCUSSIONS

3.1 Prediction performance of RF method

The overall performance of our prediction method was evaluated by 5-fold cross-validation experiments as described in Section 2. To identify the SP and SN, we use the ROC curve as shown in Figure 1 to present their interrelationship. With an input window of 5 amino acids, our RF-based classifier achieves an ACC of 84.5%, F-measure of 0.85 and AUC of 0.92. The maximum F-measure point is also shown in Figure 1, which refers to the cutoff point for the best SP and SN of the prediction performance. The detailed measures are listed in Table 1. Table 1 also summarizes the predictive performance using various input window sizes from 1 to 13, i.e. the sliding window length in the coding scheme. The maximum F-measure point is also chosen as the cutoff of SP and SN in every window size.

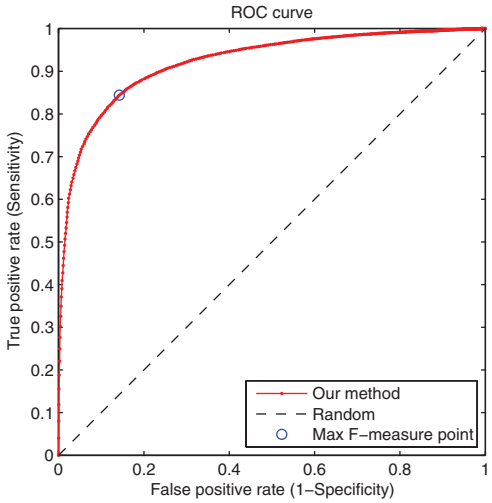

Fig. 1. The ROC curve of predicting performance.

Table 1. The result of 5-fold cross-validation of the RF classifier with different window sizes in the 53 315 residues

Window size	SN (%)	SP (%)	ACC (%)	*F*-measure	AUC
1	72.5	79.5	73.2	0.758	0.844
3	83.1	86.6	83.4	0.848	0.918
5	84.4	85.8	84.5	0.851	0.923
7	84.9	85.0	84.9	0.850	0.922
9	85.0	84.5	84.9	0.850	0.920
11	82.5	86.0	82.8	0.842	0.916
13	83.8	84.0	83.9	0.840	0.915

The performance of *F*-measure and AUC is slightly declined when we increase the window size. The best performance was obtained with the window size of 5 residues. The test provided evidences for the effectiveness of our method to predict the RNA interface residues in proteins. Figures 2a and b show an example of the predicted interface residues with RNA in protein 1R3E:A. Figure 2a presents the actual interface residues of the protein structure in red. Figure 2b uses a different coloring scheme to illustrate the prediction performance on individual residues. Figure 2c gives a part of the predicted versus the actual binding residues (structure is shown in the box panel). Most of the actual interface residues (ACC of 85.57%) are well identified in the protein, and the prediction details are shown in Supplementary Materials.

3.2 Comparison with other methods

In this work, we proposed a novel RF-based method to predict the binding residues in proteins by using the combined features. By considering the two-side information in the interacting partners of amino acids and nucleotides, we show that the proposed interaction propensity feature is able to represent the binding SP of the residues in proteins. The sequence features as well as the structure features of the residues can discriminate the propensity of amino acids to interact with RNA when combined together. The superior performance of the prediction in cross-validation experiments confirms the effectiveness of our method. In this subsection, we further validate our method by comparing our method with other existing methods.

In recent years, several methods have been proposed to predict RNA binding residues in proteins, such as RNABindR (Terribilini *et al*., 2007), BindN (Wang and Brown, 2006), RNAProB (Cheng *et al*., 2008) and PPRint (Kumar *et al*., 2008). For comparison, we tested the prediction performances of these methods. In the 205 protein chains, we constructed a training dataset by randomly selecting 105 protein chains as well as a testing dataset of the rest 100 chains. We trained the RF classifier by using the training dataset and validated the prediction in the testing dataset to compare the performance of the methods. The tradeoff threshold between SN and SP of our method was set to be RF voting score that gave the best performance. Figure 3 shows the ROC curve of the results. The predictions in the testing 100 protein chains by other methods were carried out by using their default parameters. The results are shown in Table 2. RNABindR is a Bayesian classifier for RNA binding sites in proteins. BindN, RNAProB and PPRint are predictors of protein–RNA binding residues based on SVMs. RNABindR was implemented in three different options, i.e. 'optimal prediction (opt)', 'high sensitivity (sn)' and 'high specificity (sp)' prediction. Similarly, BindN was tested with two options of 'expected sensitivity of 80% (sn)' and 'expected specificity of 80% (sp)'. RNABindR and BindN with high SP are obtained at the expense of SN, and vice versa. Clearly, our method consistently outperforms the existing methods in terms of these measures.

Since some prediction scores of the compared methods are not available, their ROC curves cannot be drawn. To remove the possible biases in the comparison, we also compared the underlying machine learning algorithms in these predictors. We implemented several different algorithms, i.e. support vector machine (SVM), naive Bayes (NB) and neural network (NN), using the same procedure as our RF-based method. Figure 3 shows the ROC curves of different classifiers in the testing dataset. The performance details are given in Table 2. In Figure 3, AUC of RF-, SVM-, NN- and NB-based predictors are 0.912, 0.801, 0.782 and 0.713, respectively. Our RF-based method clearly outperformed other classifiers. As to the different datasets, we also tested our methods in several benchmarks. In RNA binding protein datasets RB86 (Kumar *et al*., 2008; Terribilini *et al*., 2006), RB107 (Kumar *et al*., 2008; Wang and Brown, 2006), RB109 (Terribilini *et al*., 2007) and RB149 (Terribilini *et al*., 2007), our method can achieve the AUC of 0.884, 0.885, 0.884 and 0.858, respectively, which also demonstrate the effectiveness of the proposed method. We also carried out more comparison with RNAProB (Cheng *et al*., 2008). The details of these results can be found in the Supplementary Materials.

3.3 Evaluation of feature importance

We combined various features of the residues in addition to the mutual interaction propensity to represent the specific interaction properties of protein residues with RNA nucleotides. Seven descriptors were contained in a hybrid feature vector. To verify their effects on the prediction of binding sites, we tested the performance of the selected features of these descriptors. Table 3 presents the results of prediction performance of the 5-fold cross-validation by subtracting one of the descriptors individually in the scoring scheme. After subtracting each descriptor in describing these

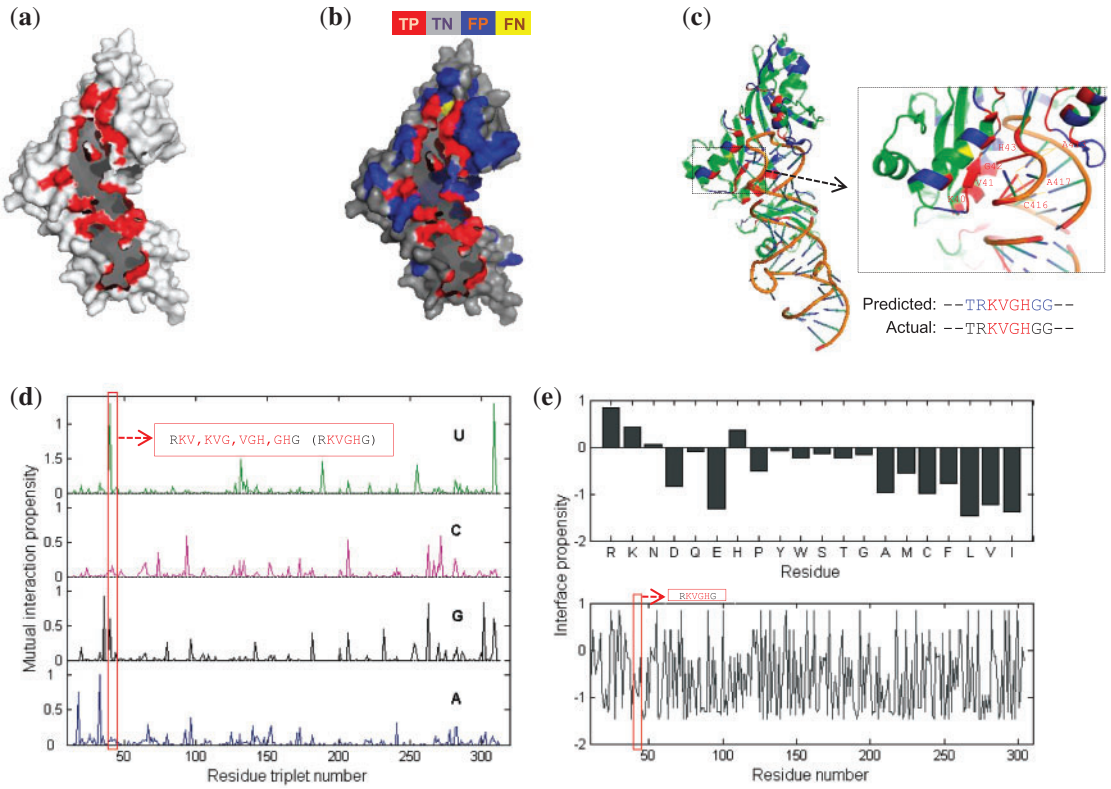

Fig. 2. An example of predicting RNA binding sites. (**a**) Actual interface residues with RNA in protein 1R3E:A. (**b**) Predictions are mapped onto the original structure where different prediction catalogs are represented by different colors. (**c**) Structure of the protein–RNA complex with an example of prediction in the zoomed part. (**d**) Mutual interaction propensity between the triplets and nucleotides in the protein. Triplets are listed by sliding residues through the protein sequence. The box part corresponds to the values of residues in the zoomed part of (c). (**e**) Upper panel shows the interface propensity of each amino acid type in the dataset. It is defined as the proportion of an amino acid in interaction sites divided by the proportion of the residue in the dataset (see more in Supplementary Materials). Lower panel shows the interface propensity of binding with RNA for the residues in the protein. The box part corresponds to the values of the zoomed sites.

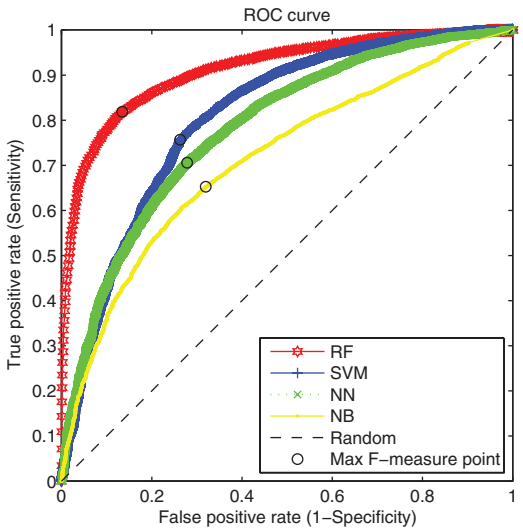

Fig. 3. The ROC performance of several classifiers.

Table 2. Comparison of the prediction performances

Method	SN (%)	SP (%)	ACC (%)	*F*-measure	MCC
Our method	81.9	86.8	82.4	0.843	0.488
RNABindR (opt)	34.2	93.8	87.3	0.501	0.300
RNABindR (sn)	83.7	49.5	62.3	0.623	0.208
RNABindR (sp)	17.1	98.6	89.7	0.292	0.281
BindN (sn)	77.3	52.9	55.5	0.628	0.188
BindN (sp)	51.0	79.6	76.5	0.622	0.225
RNAProB	74.0	65.6	73.1	0.696	0.267
PPRint	78.9	74.1	74.6	0.764	0.355
SVM based	75.7	73.7	75.5	0.747	0.335
NN based	70.6	72.2	70.7	0.714	0.280
NB based	65.3	68.1	65.6	0.667	0.211

residues, we found that the ACC of prediction was declined than that of using all descriptors. For instance, when we deleted the interaction propensity, ACC, *F*-measure and AUC of the prediction became 75.8%, 0.751 and 0.828 comparing to 84.5%, 0.859 and 0.923 with all descriptors, respectively.

From Table 3, we can also identify the importance of these features for the prediction and the contribution of these properties

Table 3. Predictive results by subtracting descriptor(s)

Without the following feature	SN (%)	SP (%)	ACC (%)	*F*-measure	AUC
Mutual interaction propensity	75.9	74.3	75.8	0.751	0.828
Physicochemical characteristics	82.9	87.1	83.3	0.849	0.920
Hydrophobicity	83.7	86.4	84.0	0.850	0.920
PSSM value	82.2	85.9	82.6	0.840	0.912
Accessible surface	82.9	83.2	82.9	0.840	0.905
Secondary structure	83.5	86.9	83.9	0.851	0.920
Side chain pKa value	83.4	86.7	83.8	0.850	0.920
Structure features	80.9	85.6	81.4	0.832	0.905
Sequence features	82.0	87.5	82.6	0.847	0.917
With all features	84.4	85.8	84.5	0.851	0.923

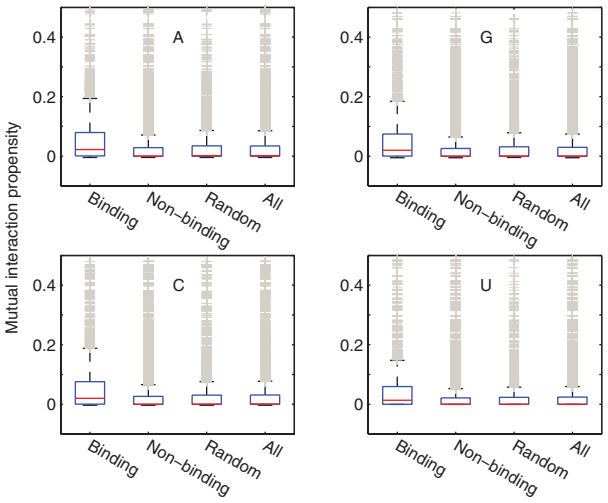

Fig. 4. Box plots of mutual interaction propensity between different residues.

encoded in the feature vector, respectively. The mutual interaction propensity of the residues and its nucleotide partners can provide more confidence of the binding events, and actually has increased the prediction performance significantly (∼10%). Figure 2d presents the mutual interaction propensity value between the triplets and nucleotides for protein 1R3E:A. The selected red box corresponds to the binding sites shown in Figure 2c. For comparison, Figure 2e gives the interface propensity information of different amino acid types (upper panel) and the corresponding interface propensity value of the residues in protein 1R3E:A. Similarly, the box corresponds to the zoomed sites shown in Figure 2c. The interface propensity (Terribilini *et al*., 2006) of each amino acid type shown in the upper panel was defined as the ratio of the percentage of one type residue in the interfaces in the percentage of the type residue in our entire dataset. From Figure 2d and e, we found that more specific information on the interacting sites in the protein were extracted by interaction propensity between residues of protein and RNA. When we use the interface propensity instead of the mutual interaction propensity in the prediction, our method only achieved the ACC of 73.5%, *F*-measure of 0.748 and AUC of 0.827. If we only use the interface propensity as the descriptor in the encoding scheme, we got the prediction of ACC of 65.3%, *F*-measure of 0.650 and AUC of 0.710 (compare to ACC of 80.8, *F*-measure of 0.831 and AUC of 0.906 when only using interaction propensity in feature vector). Clearly, the prediction performance was highly improved when we developed the mutual interaction propensity as the descriptor of the RNA binding information. We also tested the prediction performances by alternative definition of mutual interaction propensity. The details of these results are shown in the Supplementary Materials.

We also tested the prediction performance of our model with different combinations of features, which can be found in Supplementary Materials. Some of the descriptors, i.e. 'Accessible surface' and 'Secondary structure', can be calculated only after the protein structure information is available. If we category them into structure descriptors and the others are sequence descriptors, Table 3 also gives the results of the prediction which combines sequence features and that of structure information. When we use the scheme of combining mutual interaction propensity with other features based only on sequence (without 'Structure features'), our RF-based method can achieve a predictive ACC of 81.4%, *F*-measure of 0.832 and AUC value of 0.905. In contrast, the method without 'Sequence features' achieves an ACC of 82.6%, *F*-measure of 0.847 and AUC of 0.917. The results indicate that the combination of all these descriptors, i.e. interaction propensity, sequence features and structure features has more representative power so that more information can be extracted by the predictor for better differentiating protein–RNA binding residues from non-binding ones.

3.4 Patterns of RNA binding sites

We predicted the RNA binding residues in proteins from integrated features underlying the residues. The results have already shown the importance of residue properties in the identification of binding SP. The test also indicated that the combination of the descriptors can achieve better prediction performance. All these evidences demonstrate that the combined features capture the specificities of RNA binding sites, i.e. they indicate the signatures of the potential places where the binding events take place. As to the individual features of these binding sites, we can compare them with that of non-binding sites to detect their SP. Figure 4 shows the statistics of encoded mutual interaction propensity of different residues to the four RNA nucleotides in their feature vectors individually. 'Binding' represents the actual interface residues with RNA. 'Non-binding' is those residues that are not the interacting residues. When we randomly sample the same size residues as interface residues, the mutual interaction propensity statistics is shown in 'Random' panel. 'All' shows the values of all residue in the dataset. We found that the values of mutual interaction propensity between actual binding residues and four type of nucleotides are higher than that of the non-binding ones, same-size random sampling residues and all the residues in the dataset, respectively. Also it is interesting to point out that we can see from Figure 4 that the residues interacting U are harder to predict than A, C and G based on mutual interaction propensity.

When we mine the common features of these binding residues, we can identify the particular features of these residues with their prioritization contacting with RNA. These features can be used to construct the markers for discriminating RNA binding sites from non-binding ones. Structure motifs of binding RNA in the proteins would be important in the drug design (Liu *et al*., 2008). Here, we learned these features and predicted the RNA binding residues in proteins by RF classifiers. The results demonstrate that our method can identify the potential binding motifs in proteins. These features can also be used to identify the functional structure motifs, such as the functional motif (Terribilini *et al*., 2006) shown in Figure 2c. In the future, we will analyze these detailed features by relating them with the specific local structures in proteins. These identified candidate RNA binding motifs also provide fundamental structure patterns for binding RNA. Our method can be further extended to these interesting and intriguing research topics in the future.

4 CONCLUSION

Knowledge regarding how proteins interact with each other and with other molecules is essential in the understanding of cellular processes (Weigt *et al*., 2009). Here, we proposed a novel method to predict RNA binding residues in proteins using a RF-based method by the integrated features. We encoded the residue features into a vector so as to represent the SP of binding features of residues in proteins. Our method is mainly based on the mutual interaction propensity defined between the interacting residues of protein and RNA. In addition, we also combined various sequence- and structure-derived features together to represent the interaction propensity of the amino acid and its partner nucleotides. The learning and validation processes in the protein–RNA complexes confirm the effectiveness of our method. The comparison with existing methods and other classifiers further show the advantages of the proposed method. The importance and contribution of these features were also evaluated. The identified binding features underlying the protein will have implications of protein–RNA binding. In the same manner, our method can be extended to identify the structure motifs in the RNA binding proteins as well as the RNA binding SP in protein structures in future.

Funding: The Chief Scientist Program of Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences (grant no. 2009CSP002); National Natural Science Foundation of China (grant no. 10631070 and 60873205); Ministry of Science and Technology of China (grant no. 2006CB503905).

Conflict of Interest: none declared.

REFERENCES

Allers,J. and Shamoo,Y. (2001) Structure-based analysis of protein-RNA interactions using the program ENTANGLE. *J. Mol. Biol.*, **311**, 75–86.

Altschul,S.F. *et al*. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Berman,H.M. *et al*. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Berman,H.M. *et al*. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys. J.*, **63**, 751–759.

Breiman,L. (2001) Random Forests. *Mach. Learn.*, **45**, 5–32.

Chen,Y.C. and Lim,C. (2008) Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res.*, **36**, e29.

Cheng,C.W. *et al*. (2008) Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, **9** (Suppl. 12), S6.

Doherty,E.A. *et al*. (2001) A universal mode of helix packing in RNA. *Nat. Struct. Biol.*, **8**, 339–343.

Ellis,J.J. *et al.* (2007) Protein-RNA interactions: structural analysis and functional classes. *Proteins*, **66**, 903–911.

Glisovic,T. *et al*. (2008) RNA-binding proteins and post-transcriptional gene regulation. *FEBS Lett.*, **582**, 1977–1986.

Hall,K.B. (2002) RNA-protein interactions. *Curr. Opin. Struct. Biol.*, **12**, 283–288.

Jeong,E. *et al.* (2004) A neural network method for identification of RNA-interacting residues in protein. *Genome Inform.*, **15**, 105–116.

Jones,S. *et al*. (2001) Protein-RNA interactions: a structural analysis. *Nucleic Acids Res.*, **29**, 943–954.

Kabsch,W. and Sander,C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Kim,H. *et al*. (2003) Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett.*, **552**, 231–239.

Kim,O.T. *et al.* (2006) Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res.*, **34**, 6450–6460.

Kumar,M. *et al.* (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, **71**, 189–194.

Li,N. *et al.* (2008) Prediction of protein-protein binding site by using core interface residue and support vector machine. *BMC Bioinformatics*, **9**, 553.

Liaw,A. and Wiener,M. (2002) Classification and regression by random forest. *R News*, **2**, 18–22.

Liu,Z.P. *et al*. (2008) Bridging protein local structures and protein functions. *Amino Acids*, **35**, 627–650.

Lunde,B.M. *et al.* (2007) RNA-binding proteins: modular design for efficient function. *Nat. Rev. Mol. Cell Biol.*, **8**, 479–490.

Morozova,N. *et al*. (2006) Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, **22**, 2746–2752.

Nelson,D.L. and Cox,M.M. (2004) Amino acids, peptides, and proteins. In *Lehninger Principles of Biochemistry,* 4th edn. W.H. Freeman Publisher, New York, pp. 75–115.

Rost,B. and Sander,C. (1994) Conservation and prediction of solvent accessibility in protein families. *Proteins*, **20**, 216–226.

Shulman-Peleg,A. *et al.* (2008) Prediction of interacting single-stranded RNA bases by protein-binding patterns. *J. Mol. Biol.*, **379**, 299–316.

Spriggs,R.V. *et al*. (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, **25**, 1492–1497.

Sweet,R.M. and Eisenberg,D. (1983) Correlation of sequence hydrophobicities measures similarity in three dimensional protein structure. *J. Mol. Biol.*, **171**, 479–488.

Terribilini,M. *et al*. (2006) Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, **12**, 1450–1462.

Terribilini,M. *et al*. (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res.*, **35**, W578–W584.

The UniProt Consortium. (2008) The Universal Protein Resource (Uniprot). *Nucleic Acids Res.*, **36**, D190–D195.

Wang,L. and Brown,S.J. (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res.*, **34**, W243–W248.

Wang,L. *et al*. (2007) Nearest-neighbor effects on backbone alpha and beta carbon chemical shifts in proteins. *J. Biomol. NMR*, **39**, 247–257.

Weigt,M. *et al*. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proc. Natl Acad. Sci. USA*, **106**, 67–72.