SCIENTIFIC REPORTS

Correction: Author Correction

OPEN

RPI-Bind: a structure-based method for accurate identification of RNA-protein binding sites

Received: 24 November 2016
Accepted: 13 March 2017
Published online: 04 April 2017

Jiesi Luo, Liang Liu, Suresh Venkateswaran, Qianqian Song & Xiaobo Zhou

RNA and protein interactions play crucial roles in multiple biological processes, while these interactions are significantly influenced by the structures and sequences of protein and RNA molecules. In this study, we first performed an analysis of RNA-protein interacting complexes, and identified interface properties of sequences and structures, which reveal the diverse nature of the binding sites. With the observations, we built a three-step prediction model, namely RPI-Bind, for the identification of RNA-protein binding regions using the sequences and structures of both proteins and RNAs. The three steps include 1) the prediction of RNA binding regions on protein, 2) the prediction of protein binding regions on RNA, and 3) the prediction of interacting regions on both RNA and protein simultaneously, with the results from steps 1) and 2). Compared with existing methods, most of which employ only sequences, our model significantly improves the prediction accuracy at each of the three steps. Especially, our model outperforms the catRAPID by >20% at the 3rd step. All of these results indicate the importance of structures in RNA-protein interactions, and suggest that the RPI-Bind model is a powerful theoretical framework for studying RNA-protein interactions.

RNA-protein interactions are critical at many regulatory steps of gene expression and stages of organismal development[1–5]. Their interactions may vary according to sequences and structures, and consequently perform distinct functions. For example, tRNAs are bound to aminoacyl-tRNA synthetases for the translation during protein synthesis[6], and nascent RNA coordinates the transition of RNA polymerase (RNAP) II to regulate their own transcription[7]. A large class of long noncoding RNAs (lncRNAs) can bind and modulate the activity of chromatin proteins, and play roles in chromatin modifications[8–13]. In this process, lncRNAs, e.g. the *Xist*, with specific structures can localize chromatin-remodeling complex, such as DNMT3a and possibly also EZH2, to specific target regions whereby stable epigenetic gene silencing can be initiated[14–16], or act as a scaffold, e.g. the Hotair, to bind more than two proteins with their modules and direct them to target loci[17,18]. It is now apparently observed that many lncRNAs are the key regulators of transcriptional and translational output[1,19–21], in addition to other genetic and epigenetic regulators[22–29].

The development of high-throughput experimental methods, such as CLIP-seq and RIP-seq, has greatly advanced the genome-wide studies of RNA-protein interactions[30,31]. Multiple works have been reported to map the full spectrum of RNA interactions of individual RNAs and proteins. For instance, the genome-wide binding of *Xist* and its silencing partners have been profiled[32–34]. Very recently, Hendrickson *et al*. performed experiments with fRIP-Seq to detect widespread binding of mRNA and lncRNA with 24 proteins[35]. However, these experimental methods are always expensive, time-consuming and labor-intensive.

It is necessary to develop computational approaches to efficiently investigate RNA-protein interactions. A few methods have been reported for three purposes: 1) the investigation of associations between proteins and RNAs, such as RPI-Pred[36], RPIseq[37], and lncPro[38]; 2) the prediction of binding sites on either RNAs or proteins, such as the sequenced-based methods: BindN[39], RNABindR[40], RNAProB[41], PPRint[42], RNApin[43], PRINTR[44], RISP[45], PiRaNhA[46], BindN+[47], NAPS[48], PRBR[49], SRCPred[50], Predict_RBP[51], RNABindRPlus[52] and RBRIdent[53]; the structure-based methods: KYG[54], RsiteDB[55], PRIP[56], OPRA[57], DRNA[58], PRNA[59], aaRNA[60], RBRDetector[61] and RBscore[62]; and 3) the residue-nucleotide contacts prediction, such as catRAPID[63]. The catRAPID method is the only available method and different from those are specially designed to determine residue-nucleotide

Center for Bioinformatics and Systems Biology and Department of Radiology, Wake Forest School of Medicine, Winston-Salem, NC, 27157, USA. Jiesi Luo, Liang Liu and Suresh Venkateswaran contributed equally to this work. Correspondence and requests for materials should be addressed to X.Z. (email: xizhou@wakehealth.edu)

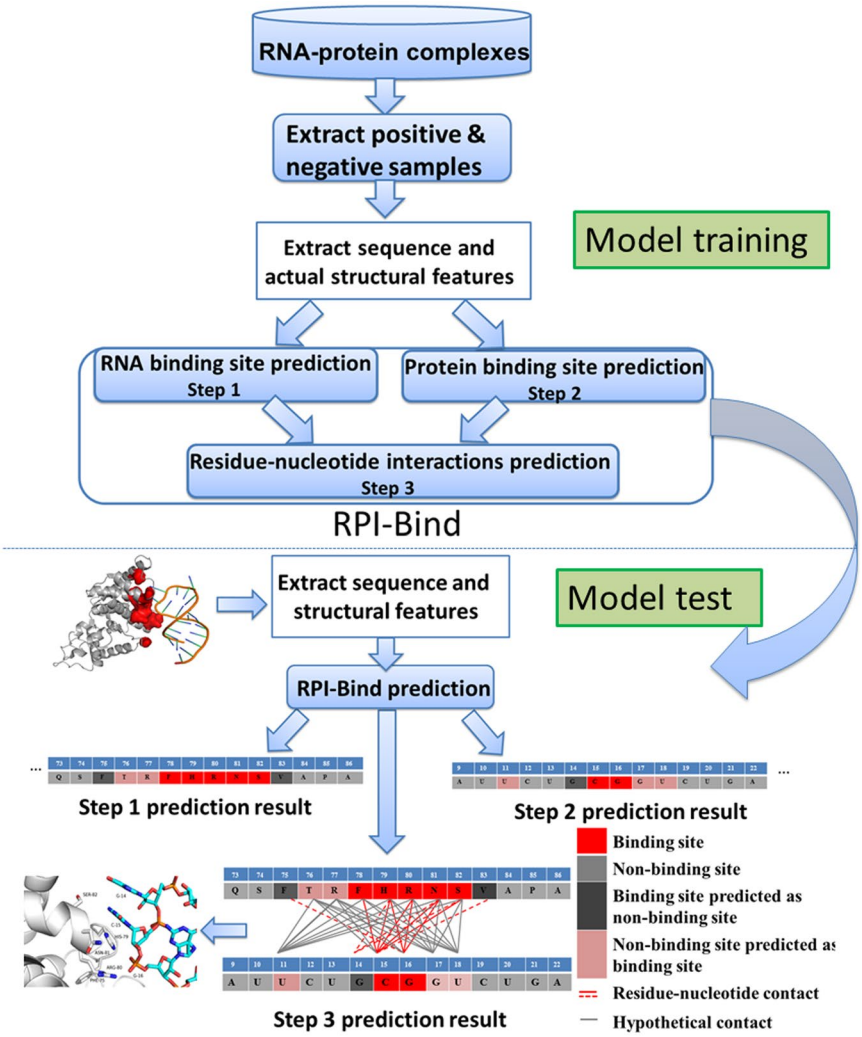

Figure 1. The step-wise work flow of the RPI-Bind prediction method. The whole work flow consists of two steps: training classification models and the applications. The model training process includes various processes, such as construction of the training dataset, feature extraction from sequences and structures in the training data set and development of 'RPI-Bind' method, consisting of three models. The developed models were then applied to solve three problems, including 1) the prediction of RNA binding regions on protein, 2) the prediction of protein binding regions on RNA, and 3) the prediction of interacting regions onboth RNA and protein simultaneously.

interactions of different known DNA-binding domain families[64,65]. The latter cannot be applied to study RNA-protein binding interactions, since RNA is more flexible than DNA and has more complicated structures.

Most of the existing methods only employ the sequences of proteins or RNAs, although some, such as catRAPID, implement other information like van der Waals contacts, hydrogen bonds, electrostatic interactions and stacking interactions across the protein-RNA interfaces. However, it is well known that, in many cases, the structures of molecules, including both proteins and coding/non-coding RNAs, dictates their functions[10,66–72]. As an example, the RPI-Pred, using high-order three-dimensional protein and RNA structures, significantly improves the accuracy in contrast to others using only sequences.

In this work, we implemented both sequences and structures of RNAs and proteins for the study of RNA-protein binding sites. To represent structures, we used the protein local conformations (PLCs), named protein blocks (PBs)[73], and 12 classes of RNA local conformations (RLCs) from 'BEAR' encoding[74], respectively. Both of them can give more detailed descriptions of RNA and protein structures than other representations[73,74]. We firstly illustrated the preferring properties of PLCs/RLCs for RNA-protein interactions. For instance, the α-helix and β-sheet PLCs, and the stem RLCs are[64] preferred to exist at protein-RNA interfaces. We then developed a three-step RPI-Bind method to identify the binding sites on a given pair of protein and RNA. The three steps include 1) the prediction of RNA binding regions on protein, 2) the prediction of protein binding regions on RNA, and 3) the prediction of interacting regions on both RNA and protein simultaneously (Fig. 1). We showed that in the 1st and 2nd steps, the inclusion of structures of RNAs and proteins can increase the prediction accuracy of RNA binding regions on protein, and RNAs, respectively. More importantly, at the 3rd step, the inclusion of

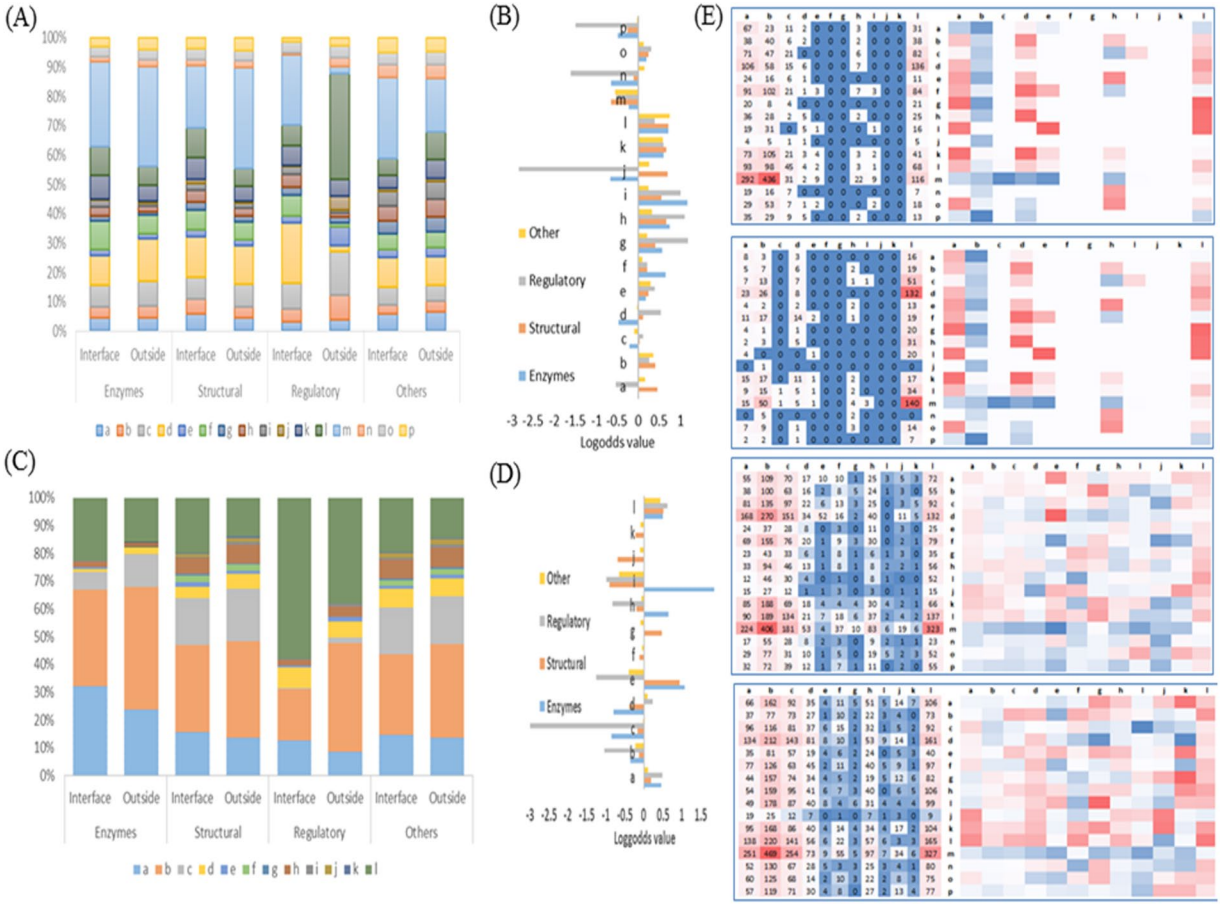

Figure 2. Statistical analysis of protein local conformations (PLCs) and RNA local conformations (RLCs) at and outside the interface for the four types of protein functional classes. (**A**) and (**C**) show the composition percentages of PLCs and RLCs at and outside the interfaces. The corresponding log-adds ratio to represent over and less at and outside the interfaces for PLCs and RLCs are given in (**B**) and (**D**). The mutual interaction propensity matrices between PLCs and RLCs are shown in (**E**). The left side values in (**E**) represent the total number of contacts between PLCs and RLCs and right side boxes represent their corresponding log-odds values. The four classes are shown from top to bottom, are enzymes, structural, regulatory and 'other', respectively.

structures and the predicted results from the first two steps can lead to an accuracy of ~86.9%, significantly higher than the catRAPID (~62%)[63]. We also applied the RPI-Bind method to identify the interacting regions of lncRNA *Xist* and transcription factor YY1, as well as other 20 proteins[35]. The results show great agreements between our predictions and experimental measurements, indicating the RPI-Bind is a powerful theoretical framework for the study of RNA-protein interactions.

Results

Statistical analysis of PLCs and RLCs at RNA-protein interfaces. We extracted 172 non-redundant RNA-protein interacting pairs (Supplemental Table S1) by filtering the pairs from the Nucleic Acid Database (NDB)[75] and the Protein Data Bank (PDB)[76]. We then constructed a database consisting of 28,780 nucleotide-residue contacts, consisting of 9,077 RNA binding sites (on proteins) and 5,692 protein binding sites (on RNAs), respectively. Meanwhile, 9,801 RNA non-binding sites and 3,078 protein non-binding sites were also collected for further analyses. The protein and RNA structures were analyzed with the PDB-2-PB database[77] and the 'BEAR' approach[74] for the PLC and RLC representations, respectively (Supplemental Tables S2 and S3).

We analyzed the PLCs/RLCs compositions, preferences and their mutual interaction propensities at the interfaces of four classes of non-redundant protein-RNA complexes, including enzymes, structural, regulatory and 'others', with each contains 40, 48, 34 and 50 protein-RNA pairs, respectively. By comparing the interface and outside PLCs among the four classes, the most populated PLC at the interface is the d type PLC, representing β-sheet, for the regulatory class (Fig. 2A,B and Supplemental Table S2). Other PLCs show the similar distributions for these four classes of protein-RNA complexes. The m and d type PLCs that represent α-helix and β-sheet are also overpopulated in all four classes, followed by the N-terminal α-helix and β-sheet PLCs (l, f, k, c, a and b types). By contrast, the C-terminal α-helix, β-sheet and coil PLCs (e, f, n, o, p, g, h, l, and j types) show unfavorable at the binding interfaces. Overall, the high preferences of l, k, h and g types were observed. All PLCs do not show much different preferences among the four classes of protein-RNA complexes, except that j, p, and n types have

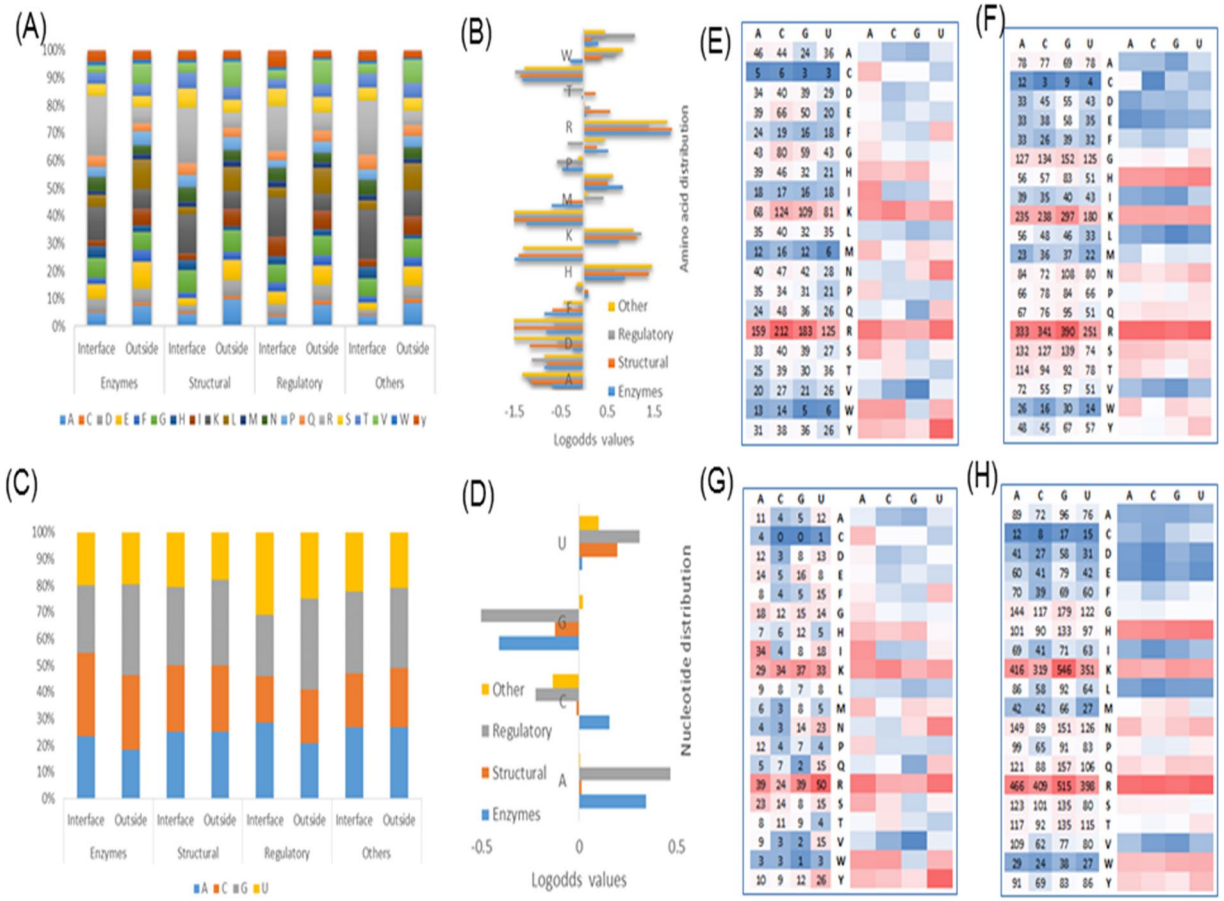

Figure 3. The occurrences of amino acid and nucleotide sequences at and outside the interface for all types of protein functional classes (enzymes, structural, regulatory, and other). The occurrences of amino acid and nucleotide at the interface and outside are shown in (**A**) and (**C**), respectively. In (**B**) and (**D**), the log-odds ratio of amino acid and nucleotide shows the over and less populated amino acid and nucleotides at and outside the interfaces. Further, the mutual interaction propensities (log-odds value) between amino acids and nucleotides are given for all four classes in (**E–H**), respectively. In each figure, the values on the left side represent the total number of contacts between residue and nucleotide and right side boxes represent their corresponding log-odds values.

the lowest preferences in the regulatory class (Fig. 2A,B and Supplemental Table S2). The overall local structure description allows us to understand the protein-RNA binding nature in terms of structural fragments.

For RNAs, the major difference lies in the c type RLC, representing stem branch with less number in regulatory class (Fig. 2C and D). The b and l type RLCs, representing stem and unknown regions respectively are highly overpopulated in all four classes of protein-RNA complexes. Followed by stem branch, left internal loop, bulge left, left internal loop branch and bulge left branch (c, d, e, f and g types) are also populated. The Right internal loop, Bulge right, Right internal loop branch, and Bulge right branch (h, i, j and k types) show few or no contacts throughout all four classes (Supplemental Table S3).

The mutual interaction propensity (MIP) matrices of PLCs and RLCs were analyzed for quantitative evaluation of protein/RNA structure preferences. According to the total number of occurrences, the pair of α-helix and stem (m-b pair) appears the most at the interface for the enzymes, structural and 'others' classes (Fig. 2E). While in the regulatory class, α-helix and unknown pair (m-l pair) is highly preferred. Among all PLCs and RLCs, α-helix and β-strand PLCs (m and d types), and loop, stem and unknown RLCs (a, b, and l types) have the most contacts. Overall, the enzymes and regulatory classes share similar interaction propensities: the RLCs, including loop, left internal loop, right internal loop and unknown (a, d, h, and l types), show high propensities for interacting. In both structural and 'others' classes, many pairs have high interaction propensities, but the highest propensities were observed for β-strand – bulge left pair (d-e pair) and coils–bulge right branch pair (g-k pair), respectively (Fig. 2E).

In addition to the PLCs and RLCs, we also analyzed the compositions, preferences and interaction propensities of amino acids and nucleotides from the four classes of protein-RNA complexes (Fig. 3). The positively charged residue, arginine and lysine, and the single aromatic residues, phenylalanine and tyrosine, play key roles in the RNA binding sites, consistent with the previous studies[78,79]. RNA-binding proteins achieve RNA-binding affinity through favorable charge-charge interactions between positively charged Arg and Lys residues and the

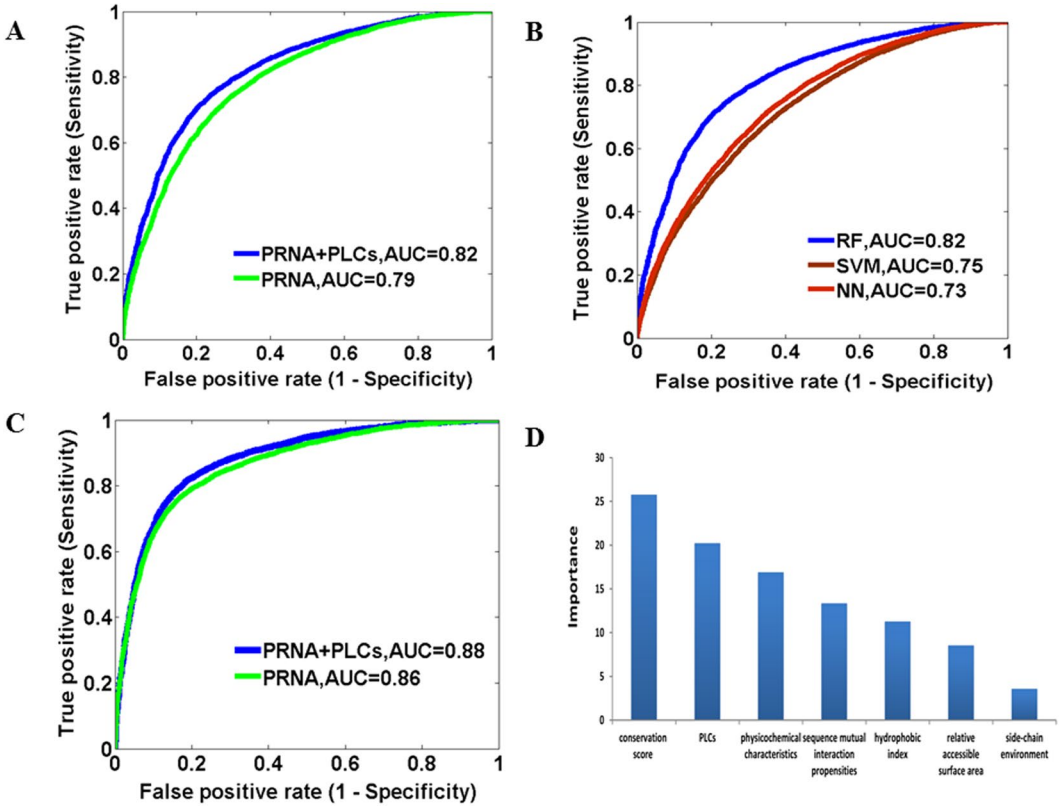

Figure 4. The performance of RNA binding site prediction. (**A**) Comparison of ROC curves for binding site prediction using different features on our constructed database. (**B**) Comparison of ROC curves for binding site prediction using different classifiers. (**C**) Comparison of ROC curves for binding site prediction on an independent dataset. (**D**) The importance and individual contribution ratio of each feature type.

negatively charged RNA phosphate. The observed high propensity of single aromatic residues reflects the frequent stacking interactions between aromatic side chains and nucleic acid bases in a number of protein-RNA complexes. In contrast to the observation of enormous variances in individual residue at the interface, few nucleotide variations were observed among protein binding sites. This could be the reason why very few methods are currently available to predict the binding sites of proteins on RNAs.

PLCs enable highly accurate prediction of RNA binding regions on proteins. We then start to build the first model (1st step) to predict RNA binding regions on proteins for the three-step RPI-Bind (RNA-protein binding region predictor) method (Fig. 1).

We developed a Random Forest (RF)-based machine learning method for the prediction of RNA binding sites on proteins, by firstly using the features of sequence mutual interaction propensities, physicochemical characteristics, hydrophobic index, relative accessible surface area, conservation score and side-chain environment, respectively. These features were chosen because they had been shown to outperform other RNA binding residue prediction methods[59]. Within the constructed database of 9,077 RNA binding sites and 9,801 RNA non-binding sites, we used a standard five-fold cross validation procedure to estimate the performance. The Sensitivity (SN), Specificity (SP), Accuracy (ACC), and Matthew's Correlation Coefficient (MCC) were 66.8%, 74.8%, 71.3% and 0.425, respectively.

We then combined the structure local conformations features of protein and RNA structures (PLCs and RLCs), because it could be better to predict binding residues. A five-residue sliding window method was used, as it outperforms other (3, 7, or 9) size windows. If the center residue or nucleotide in each window was detected to physically interact with each other, then the window was treated as a positive sample, otherwise negative. The performance could be improved to 71.1%, 77.7%, 74.8% and 0.489, respectively. The ROC curves were also adopted to show the prediction accuracy (Fig. 4A).

We also employed other machine learning methods for the model development, including Support Vector Machine (SVM) and Neural Network (NN). The result shows that the RF-based model achieves the best prediction (Fig. 4B). Therefore we selected the RF for the following analyses.

To evaluate the performance of our model and the contribution of structures in RNA-protein interactions, we applied the developed model to an independent dataset from the PRNA method[59]. It should be noted that the independent dataset and our non-redundant dataset have partially overlaps. There are 26 common protein chains in the two datasets. So the redundant protein chains were removed from independent dataset and the final dataset contains 3,584 (10.3%) interacting residues and 31,284 (89.7%) non-interacting residues. Our model

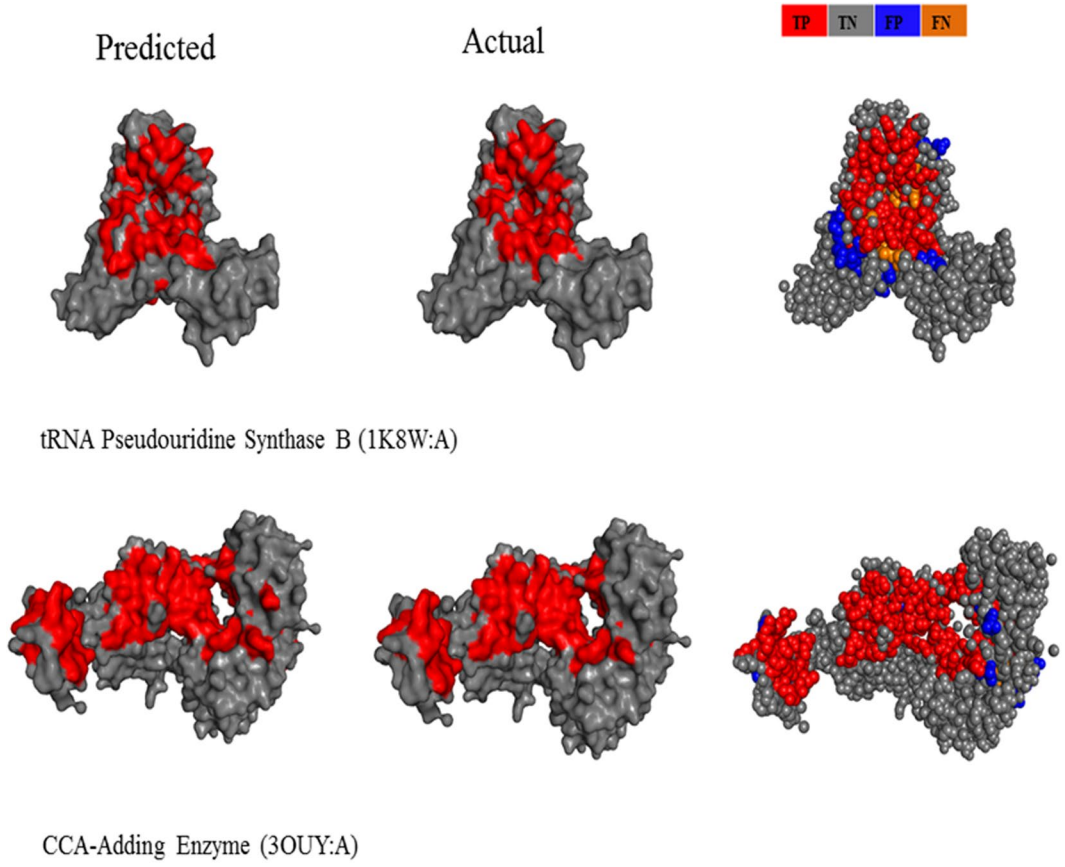

Figure 5. Examples of predicted RNA-protein interacting complexes. The two examples are tRNA Pseudouridine Synthase B and CCA-Adding Enzyme Predicted RNA binding sites are shown in red and predicted non-binding sites in gray (left panels). Actual RNA binding sites in red and actual non-binding sites in gray (middle panels). The performance of prediction for individual residues, with true positives (TP) shown in red, false positives (FP) in blue, false negatives (FN) in orange, and true negatives (TNs) in gray (right panels). Thus, red + orange residues correspond to the actual binding residues; red + blue residues correspond to the predicted binding residues. All structure diagrams were generated using PyMol (http://www.pymol.org).

can achieve predictions with ACC of 81.3%, PPV of 31.9%, NPV of 97.5% and the Area Under ROC curves (AUC) of 0.88 (Fig. 4C), compared to the PRNA with the ACC, PPV, NPV and AUC as 78.5%, 27.3%, 96.2% and 0.86. Furthermore, we test the 1st step model on a recent data set, RBscore_R130[62]. The common protein chains to our non-redundant dataset and independent dataset were removed from RBscore_R130. Our model exhibits high prediction accuracy on this set with 0.8305 AUC, compared with 0.8063 AUC for BindN+[47], 0.82 for RNABindRPlus[52], 0.8383 for aaRNA[60], 0.7501 for PPRint[80] and 0.7479 for KYG[54].

We also used the permutation importance analysis to evaluate the contribution of individual feature type in predictions (Fig. 4D). The importance score of our local conformations feature is 20, which is higher than that of other features except the conservation score. These results indicate that structures (local conformation features) make great contribute to the prediction of RNA binding sites and could improve the prediction performance.

The Fig. 5 shows examples in which the prediction method was tested with two RNA binding proteins, including CCA-Adding Enzyme (PDB 3OUY:A)[81], and tRNA Pseudouridine Synthase B (PDB 1K8W:A). The CCA-adding enzymes are nucleotidyltransferases that catalyze the posttranscriptional addition of the nucleotide sequence CCA onto the 3′ terminus of immature tRNA without using a nucleic acid template, and tRNA Pseudouridine Synthase B catalyzes the isomerization of specific uridines in cellular RNAs to pseudouridines and may function as RNA chaperones. For the former, our method correctly identified 89 of 97 actual interface residues, and for the latter, we predicted 63 out of 83 actual interface residues. The prediction accuracy for each RNA-protein complex is shown in the Supplemental Table S4. These results prove the predictive ability of our model.

RLCs enable highly accurate prediction of protein binding regions on RNAs. We also built a RF-based machine learning method for the prediction of protein binding sites on RNAs, firstly using three compositional features (the 2nd step of our approach, Fig. 1), including mono-, di- and tri-nucleotide composition profile[43]. Similarly, within the constructed database of 5,692 protein binding sites and 3,078 protein non-binding

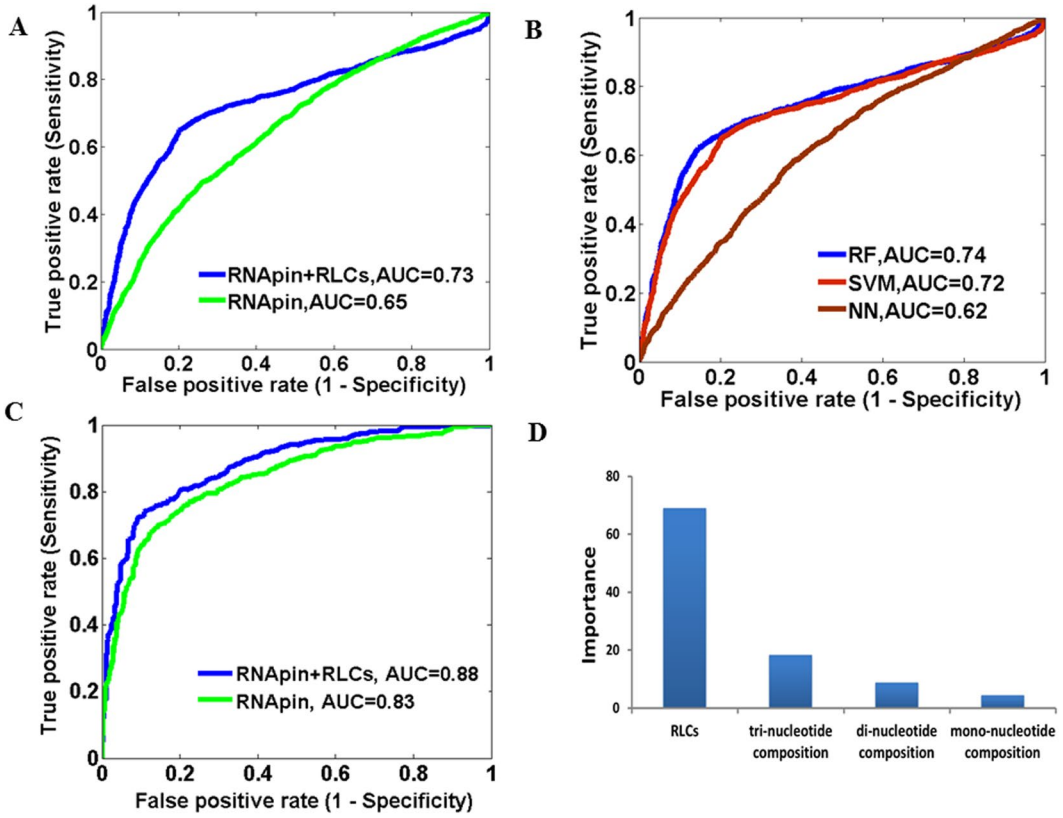

Figure 6. The performance of protein binding site prediction. (**A**) Comparison of ROC curves for binding site prediction using different features on our constructed database. (**B**) Comparison of ROC curves for binding site prediction using different classifiers. (**C**) Comparison of ROC curves for binding site prediction on an independent dataset. (**D**) The importance and individual contribution ratio of each feature type.

sites, we used a standard five-fold cross validation procedure to estimate the performance, that is, SN, SP, ACC and MCC were 72.6%, 54.1%, 63.5% and 0.271, respectively.

When additional structural features were used, the prediction could reach 77.4%, 65.2%, 71.4% and 0.4 for SN, SP, ACC and MCC, respectively (Fig. 6A). Other machine learning methods, such as SVM, also give similar performances (Fig. 6B). Therefore, the five-residue window and the RF were selected for the model building and following analyses. We then applied our method to a 'RNA208' dataset to compare the performance with the RNApin[43], one of the two methods currently available to predict protein interacting nucleotides in RNAs and having better performances[43,82]. We also removed 23 overlapped RNA chains from the 'RNA208' dataset. Our method outperformed the RNApin (77.6% ACC, 55.7% PPV, 94.6% NPV and 0.83 AUC) with ACC of 81.9%, PPV of 59.5%, NPV of 95.0% and AUC of 0.88 on the new dataset (Fig. 6C). More importantly, the importance score of RNA local conformations feature is as high as 70, by applying the permutation importance analysis (Fig. 6D), implying that the structures contain more important information than the sequence only.

The Fig. 7 shows examples of our predictions. Our model correctly identified 19 of 24 actual interface nucleotides for the 7S.S SRP RNA[83]. The SRP (signal recognition particle) is a ribonucleoprotein, which associates with ribosomes to mediate co-translational targeting of membrane and secretory proteins to biological membranes. The S domain of SRP interacts with 7S.S RNA for SRP assembly in Eukarya and Archaea. Our model also successfully predicted 18 of 25 actual interface nucleotides for T-RNA, 20 of 27 for fragment of 23S rRNA and 12 of 20 for dsRNA, respectively. These results indicate that our model is reliable for identifying protein binding sites on RNAs (Supplemental Table S4).

PLCs and RLCs enable residue-nucleotide interaction prediction at RNA-protein interface. The successful prediction of RNA binding sites on proteins and protein binding sites on RNAs in the above two steps of RPI-Bind indicates the importance of structures in the determination of RNA-protein interactions. We further used the structure and sequence information to build the 3rd model of RPI-Bind for the identification of residue-nucleotide interactions at the protein-RNA interfaces, that is, interacted residues in the protein chains and nucleotides in the RNA chains. There are a total of 28,780 residue-nucleotide contacts in our database, here, 70% of them were randomly selected to construct the training set, and the remaining was put into the test set to evaluate the performance of our final model. Each residue-nucleotide interaction was divided into fragments by moving a window of 5 successive residues or nucleotides along the proteins and RNAs. Each window was encoded by structure and sequence features used in the 1st and 2nd steps. We then trained our RPI-Bind model with the combinational features by employing the RF classifier. Our model can achieve predictions with SN of 70.0%, SP

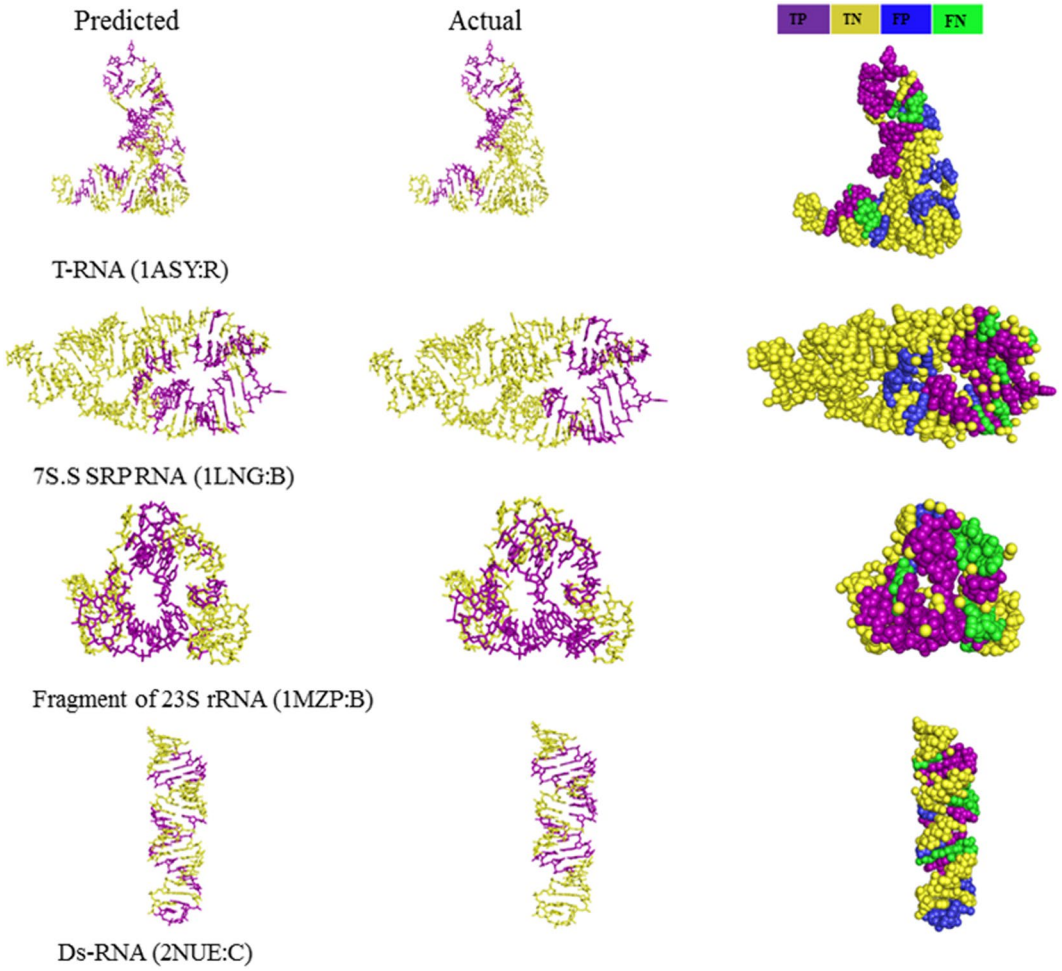

Figure 7. Examples of predicted RNA-protein interacting complexes. Examples of prediction results for four different RNA are shown from top to bottom, are T-RNA, 7S.S SRP RNA, Fragment of 23S rRNA and dsRNA, respectively. Predicted protein binding sites are shown in purple and predicted non-binding sites in yellow (left panels). Actual RNA binding sites in purple and actual non-binding sites in yellow (middle panels). The performance of prediction for individual nucleotides, with true positives (TP) shown in purple, false positives (FP) in blue, false negatives (FN) in green, and true negatives (TNs) in yellow (right panels). Thus, purple + green nucleotides correspond to the actual binding nucleotides; purple + blue nucleotides correspond to the predicted binding nucleotides. All structure diagrams were generated using PyMol (http://www.pymol.org).

of 89.2%, ACC of 79.5%, and MCC of 0.60, evaluated by five-fold cross validation. Then, we further evaluate the practical prediction ability of our model on test set. The SN, SP, ACC and MCC are 73.4%, 83.8%, 79.7% and 0.57, respectively.

Furthermore, we can take the advantage of our predictions from the RNA binding sites on proteins (1st step, Fig. 1) and the protein binding sites on RNAs (2nd step, Fig. 1). We selected the successfully predicted 4,738 (out of 6,334, from step 1) interacting residues and 2,608 (out of 3,984, from step 2) interacting nucleotides, which form 11,211 true residue-nucleotide contacts from our training set. We firstly used five-fold cross validation to estimate the performance of new training set. The average SN, SP, ACC and MCC are 84.0%, 90.0%, 86.9% and 0.76, respectively. The performance of the new training set was also evaluated on test set, the SN, SP, ACC and MCC are 82.4%, 95.1%, 91.4% and 0.79, respectively. These results show that the inclusion of structures and the predicted results from the first two steps significantly increase the prediction accuracy of residue-nucleotide contacts. Our method significantly outperforms the catRAPID method, the prediction accuracy of which is close to 62%.

Test of two individual RNA-protein interacting complexes (PDB id: 1I6U and 3IAB) also indicates that our model is reliable for residue and nucleotide interaction prediction (Fig. 8A and C). Our 1st step model of RPI-Bind correctly predicted 28 of 40 and 34 of 46 RNA interacting residues for the 1I6U and 3IAB, respectively. Then the 2nd step model correctly predicted 11 of 14 and 10 of 16 protein interacting nucleotides, respectively. For the residue-nucleotide contacts prediction, our 3rd step model correctly predicted all actual interactions. On the other hand, 2% of non-contacts was wrongly predicted as contacts (false positives), if the lowest score of actual interactions was set as a cutoff (Fig. 8B and D). In contrast, the catRAPID failed to identify the actual residue-nucleotide

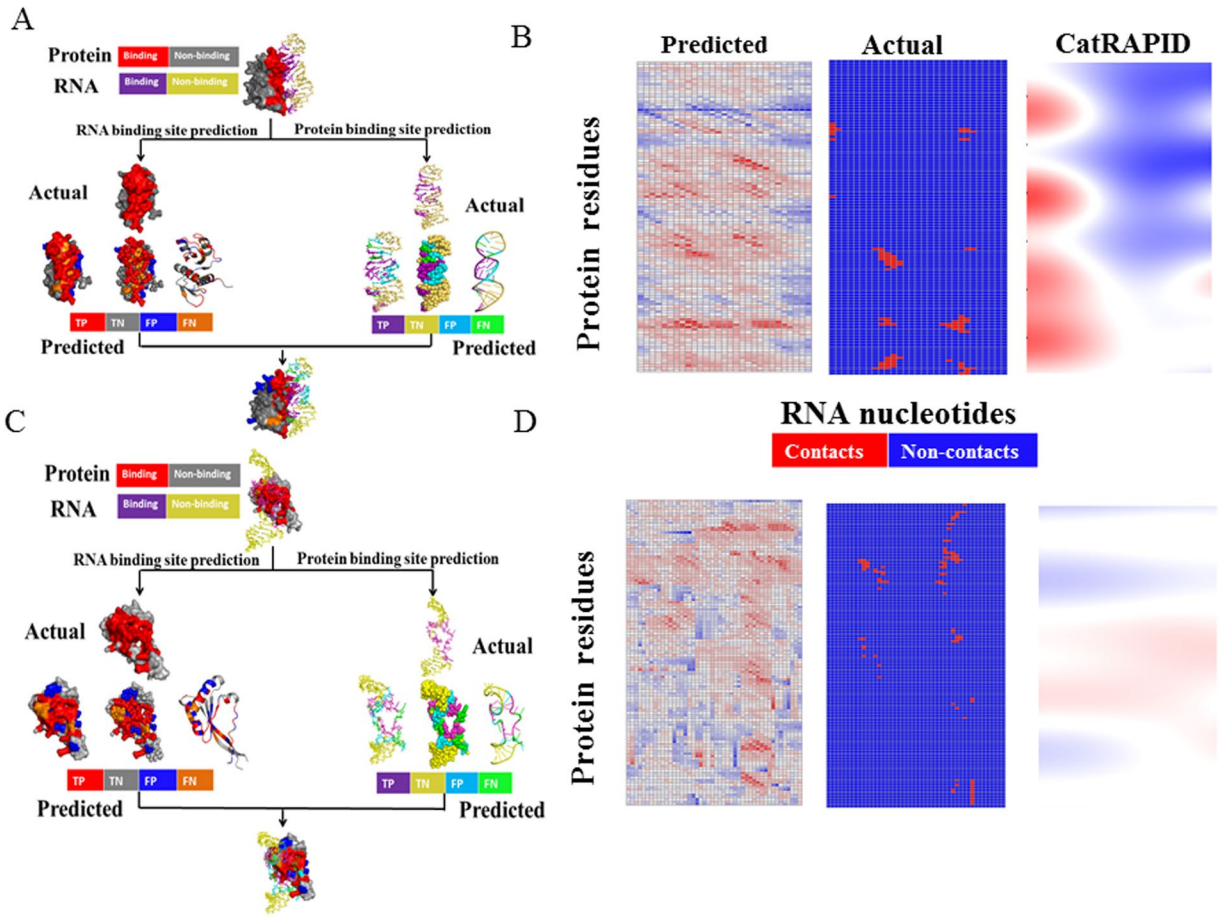

Figure 8. Two example of protein-RNA complexes (PDB id: 1I6U and 3IAB). (**A**) and (**C**) Protein and RNA binding sites prediction results of 1I6U and 3IAB, respectively. The results are mapped onto the original structure where different prediction catalogs are represented by different colors; (**B**) and (**D**) Comparison of residue-nucleotide contacts prediction results by our 3rd step model and the catRAPID method (http://service.tartaglialab.com/page/catrapid_group).

contacts and the majority of residues and nucleotides in the protein chain and RNA chain were predicted as binding sites. All of these results indicate the outperformance of our model in RNA-protein binding prediction, and the crucial roles of RNA and protein structures in their interactions.

Predicting protein associations with lncRNAs. LncRNAs play essential roles in a variety of biological process[84], and are implicated in serial steps of cancer development[20]. LncRNAs mainly perform the biological functions by interacting with different proteins[13]. Our previous RPI-Pred method[36] can accurately predict protein-lncRNA interaction pairs or identify the binding partners of a given protein or RNA. On the basis of RPI-Pred, our three-step RPI-Bind model can further predict the interaction region between protein and lncRNA, and therefore can be useful for the study of lncRNA functions.

The first application of our model is to investigate the interaction between lncRNA *Xist* and a transcription factor YY1. Their specific contacts are required to load *Xist* onto X chromosome[85]. Due to the lack of solved structures, we applied PB-kPRED[86] and RNAfold[87] for the local conformation prediction for YY1 and *Xist*, respectively. As well, the *Xist* is very large, so in order to obtain better lncRNA structures, functionally important segments, A, F, B, C, and E, were selected, and analyzed separately. Our 1st step model predicted three regions of YY1 (the sequence from 60–80, 183–210 and 303–320) are interacted with *Xist*, while the 2nd step model predicted the B, C and E segments of *Xist* are more likely to associate with YY1. With our 3rd step model, we successfully identified that the sequence from 60–80, 183–210 and 303–320 of YY1, and the B and C segments of the *Xist* are really contacted (Supplemental Fig. S1), in consistence with experimental evidence[85].

The second application of our model is to identify the binding regions between *Xist* and other 20 chromatin-associated proteins, such as EZH2 and CHD4, which have been recently determined with the fCLiP-seq technique[35]. Our goal is to predict potential regions that associate with a number of proteins involved in epigenetic regulation on the *Xist*. The results shown by the heat-map is very intuitive and straight-forward (Supplemental Fig. S2). Our model predicts that *Xist* segment E (6990–9467 nt) binds strongly to EZH2, CHD4, SUZ12 and HNRNPU, in agreement with experimental evidence. EZH2 also shows high interaction propensity to segment D (5550–5730). Neither segment A nor F is predicted to be in contact with any of the 20

chromatin-associated proteins except SUZ12. On the other hand, no or low interactions were found between four proteins (HDAC1, CHD1, HUR and CBX3) and any functional segments of *Xist*. For those *Xist* non-binding proteins (RBBP5, CBP, CHD7, DNMT1, ADAR, CTCF, PCAF, NUP98, WDR5, LSD1, IMP1), our method correctly predicted that they do not bind to any functional segment of *Xist*. Our model also predicts some unknown regions of *Xist*, such as *Xist* 3′terminus, showing high propensity to contact most of proteins, which implies that these regions might be new functional segments. In summary, our model provides highly accurate predictions of RNA-protein interactions, in agreement with experimental evidence (Supplemental Fig. S2).

Discussion

RNA-protein interactions are of importance for the function of RNAs, especially lncRNAs[8–13]. Numerous methods have been proposed for the identification of these interactions; however, two main limitations exist for those methods: 1) most of them are only able to identify the interacting regions of either RNA on protein or protein on RNA[40–43,88]. The model for residue-nucleotide contacts prediction is catRAPID, which unfortunately cannot provide highly accurate identifications. 2) Most of them only use the sequences of proteins or RNAs, but ignore their structures, which, however, are much known to affect their functions and interactions[10,66–72]. In this work, we extracted the general characteristics of protein and RNA binding. To do so, the best approach is to observe the difference between binding and non-binding sites from known structures. However, observation from one single pair of protein-RNA binding structures is always bias, and may not represent the general characteristics of binding/non-binding sites. In contrast, the comparison between the collection of residues including both binding and non-binding ones leads to the detection of these features. Therefore, we first illustrated the structure preferences of binding and non-binding sites at the protein-RNA complexes, and then developed a three-step RPI-Bind model for the prediction of RNA-protein interface by employing the sequence and structure information of RNAs and proteins. Tests show that our RPI-Bind model outperforms other existing models at each of the three levels, including 1) the prediction of RNA binding regions on protein, 2) the prediction of protein binding regions on RNA, and 3) the prediction of interacting regions on both RNA and protein simultaneously. Of note, at the third step, the inclusion of the predicted results from the first two steps can further improve the prediction accuracy, which is significantly better than the catRAPID (86.9% vs 62%). More tests on individual RNA and protein interactions, especially lncRNA-protein interactions, further illustrate the prediction ability of our RPI-Bind model, which, on the other hand, depict the importance of structures in RNA-protein interactions. Our model is freely available at http://ctsb.is.wfubmc.edu/publications/RPI-Bind-Pred.php.

Materials and Methods

We constructed a non-redundant protein-RNA interacting dataset based on the 1,342 protein-RNA complexes from the Nucleic Acid Database (NDB)[75] and their corresponding 3D structures in the Protein Data Bank (PDB)[76]. After a series of data processing steps (Supplemental Materials), we obtained 172 non-redundant protein-RNA pairs (Supplementary Table 1). In these protein-RNA pairs, a total of 28,780 residue-nucleotide contacts, consisting of 9,077 RNA binding sites and 5,692 protein binding sites were obtained based on the distances between interacting residues and nucleotides. As well, 9,801 RNA non-binding sites and 3,078 protein non-binding sites were randomly paired to obtain a total of 180,000 pairs as the negative dataset (Supplemental Materials).

The structures of proteins and RNAs were represented by protein local conformations (PLCs) as 16 types of the protein blocks (PBs), and 12 types of RNA local conformations (RLCs) (Supplemental Materials, and Supplemental Tables S2 and S3). We then investigated the PLCs/RLCs compositions, preferences, and their mutual interaction propensities at the interfaces of protein-RNA complexes (Supplemental Materials).

With these observations, we presented a three-step RPI-Bind (RNA-protein binding region predictor) method for the prediction of binding sites on both RNAs and proteins (Fig. 1) as follows:

1) The prediction of RNA binding regions on protein. We extracted sequence and structure features for RNA binding sites to develop the prediction method. The involved features include mutual interaction propensities, physicochemical characteristics, hydrophobic index, relative accessible surface area, conservation score, and side-chain environment, as well as the PLCs descriptors (triplet-log-odds values) (Supplemental Materials). We employed the sliding window approach to decode the amino acid residues of proteins. Whether a residue belongs to the interactions or not is determined by the middle residue. The feature vector representing the residue in the window is encoded by the properties of the included residues. We compared the performance of different window sizes (3, 5, 7 and 9), and the best prediction performance was obtained with a window of 5 residues. We should note that the mutation of protein leads to two different protein sequences, and maybe two different protein structures, at least for local structure. Therefore, the input to the model will be different, which will result in different binding site identification.
2) The prediction of protein binding regions on RNA. The involved features include mono-, di- and tri-nucleotide sequence compositions, and RLCs descriptors (triplet-log-odds values) (Supplemental Materials). We also compared the performance of four window sizes (3, 5, 7 and 9). The five, seven and nine-window size have similar performance, but better than three-window size. We then set the window size as 5.
3) The prediction of interacting regions on both RNA and protein simultaneously. Our goal is to predict all interacting information in a framework. So the involved dataset and features include all used in the 1st step and 2nd step. We presented two models. One uses the combinational features to develop model, and the original dataset consisting of 28,780 residue-nucleotide contacts for training and testing the model. Another also uses the combinational features, but the dataset was constructed with the positive dataset, obtained from the successfully predicted interacting residues and interacting nucleotides at the 1st and 2nd steps, respectively. The original negative dataset consisting of 180,000 pairs was formed by pairing every

binding site and neighbor non-binding sites around its interacting partners. We used this strategy to construct negative dataset, because these false contacts are more similar to true contacts. Distinguishing true protein-RNA contacts from these false contacts is more practical.

Machine learning methods, such as the Random Forest (RF), Support Vector Machine (SVM) and Neural Network (NN), were employed for model building at the three steps, and the performance was evaluated with Sensitivity (SN), Specificity (SP), Accuracy (ACC), the Area Under ROC curves (AUC), and Matthew's Correlation Coefficient (MCC) (Supplemental Materials).

References

1. Lee, J. T. Epigenetic regulation by long noncoding RNAs. *Science* **338**, 1435–1439, doi:10.1126/science.1231776 (2012).
2. Eddy, S. R. Non-coding RNA genes and the modern RNA world. *Nature reviews. Genetics* **2**, 919–929, doi:10.1038/35103511 (2001).
3. Huttenhofer, A., Schattner, P. & Polacek, N. Non-coding RNAs: hope or hype? *Trends in genetics: TIG* **21**, 289–297, doi:10.1016/j.tig.2005.03.007 (2005).
4. Hirota, K. *et al*. Stepwise chromatin remodelling by a cascade of transcription initiation of non-coding RNAs. *Nature* **456**, 130–134, doi:10.1038/nature07348 (2008).
5. Morris, K. V. *Non-coding RNAs and epigenetic regulation of gene expression: Drivers of natural selection*. (Horizon Scientific Press, 2012).
6. Cusack, S. Aminoacyl-tRNA synthetases. *Curr Opin Struct Biol* **7**, 881–889 (1997).
7. Ji, X. *et al*. SR proteins collaborate with 7SK and promoter-associated nascent RNA to release paused polymerase. *Cell* **153**, 855–868, doi:10.1016/j.cell.2013.04.028 (2013).
8. Moran, V. A., Perera, R. J. & Khalil, A. M. Emerging functional and mechanistic paradigms of mammalian long non-coding RNAs. *Nucleic acids research* **40**, 6391–6400, doi:10.1093/nar/gks296 (2012).
9. Koziol, M. J. & Rinn, J. L. RNA traffic control of chromatin complexes. *Curr Opin Genet Dev* **20**, 142–148, doi:10.1016/j.gde.2010.03.003 (2010).
10. Mercer, T. R. & Mattick, J. S. Structure and function of long noncoding RNAs in epigenetic regulation. *Nat Struct Mol Biol* **20**, 300–307, doi:10.1038/nsmb.2480 (2013).
11. Kelley, R. L. & Kuroda, M. I. Noncoding RNA genes in dosage compensation and imprinting. *Cell* **103**, 9–12 (2000).
12. Rinn, J. L. & Chang, H. Y. Genome regulation by long noncoding RNAs. *Annual review of biochemistry* **81**, 145–166, doi:10.1146/annurev-biochem-051410-092902 (2012).
13. Wang, K. C. & Chang, H. Y. Molecular mechanisms of long noncoding RNAs. *Mol Cell* **43**, 904–914, doi:10.1016/j.molcel.2011.08.018 (2011).
14. Hung, T. *et al*. Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature genetics* **43**, 621–629, doi:10.1038/ng.848 (2011).
15. Gupta, R. A. *et al*. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature* **464**, 1071–1076, doi:10.1038/nature08975 (2010).
16. Groen, J. N., Capraro, D. & Morris, K. V. The emerging role of pseudogene expressed non-coding RNAs in cellular functions. *The international journal of biochemistry & cell biology* **54**, 350–355, doi:10.1016/j.biocel.2014.05.008 (2014).
17. Kung, J. T., Colognori, D. & Lee, J. T. Long noncoding RNAs: past, present, and future. *Genetics* **193**, 651–669, doi:10.1534/genetics.112.146704 (2013).
18. Tsai, M. C. *et al*. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* **329**, 689–693, doi:10.1126/science.1192002 (2010).
19. Xing, Z. *et al*. lncRNA directs cooperative epigenetic regulation downstream of chemokine signals. *Cell* **159**, 1110–1125, doi:10.1016/j.cell.2014.10.013 (2014).
20. Yang, G., Lu, X. & Yuan, L. LncRNA: a link between RNA and cancer. *Biochimica et biophysica acta* **1839**, 1097–1109, doi:10.1016/j.bbagrm.2014.08.012 (2014).
21. Cao, J. The functional role of long non-coding RNAs and epigenetics. *Biol Proced Online* **16**, 11, doi:10.1186/1480-9222-16-11 (2014).
22. Yip, K. Y. *et al*. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol* **13**, R48, doi:10.1186/gb-2012-13-9-r48 (2012).
23. Encode Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74, doi:10.1038/nature11247 (2012).
24. Cheng, C. *et al*. Understanding transcriptional regulation by integrative analysis of transcription factor binding data. *Genome research* **22**, 1658–1667, doi:10.1101/gr.136838.111 (2012).
25. Chen, H. *et al*. An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape. *Scientific reports* **5**, 8465, doi:10.1038/srep08465 (2015).
26. Liu, L., Jin, G. & Zhou, X. Modeling the relationship of epigenetic modifications to transcription factor binding. *Nucleic acids research* **43**, 3873–3885, doi:10.1093/nar/gkv255 (2015).
27. Chen, X. *et al*. Analysis of DNA methylation and gene expression in radiation-resistant head and neck tumors. *Epigenetics* **10**, 545–561, doi:10.1080/15592294.2015.1048953 (2015).
28. Liu, L. *et al*. Mutated genes and driver pathways involved in myelodysplastic syndromes-a transcriptome sequencing based approach. *Mol Biosyst* **11**, 2158–2166, doi:10.1039/c4mb00663a (2015).
29. Liu, L., Zhao, W. & Zhou, X. Modeling co-occupancy of transcription factors using chromatin features. *Nucleic acids research* **44**, e49, doi:10.1093/nar/gkv1281 (2016).
30. Licatalosi, D. D. *et al*. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature* **456**, 464–469, doi:10.1038/nature07488 (2008).
31. Konig, J. *et al*. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol* **17**, 909–915, doi:10.1038/nsmb.1838 (2010).
32. Kaneko, S., Son, J., Shen, S. S., Reinberg, D. & Bonasio, R. PRC2 binds active promoters and contacts nascent RNAs in embryonic stem cells. *Nat Struct Mol Biol* **20**, 1258–1264, doi:10.1038/nsmb.2700 (2013).
33. Khalil, A. M. *et al*. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 11667–11672, doi:10.1073/pnas.0904715106 (2009).
34. Zhao, J. *et al*. Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Mol Cell* **40**, 939–953, doi:10.1016/j.molcel.2010.12.011 (2010).
35. Hendrickson, D. G., Kelley, D. R., Tenen, D., Bernstein, B. & Rinn, J. L. Widespread RNA binding by chromatin-associated proteins. *Genome Biol* **17**, 28, doi:10.1186/s13059-016-0878-3 (2016).
36. Suresh, V., Liu, L., Adjeroh, D. & Zhou, X. RPI-Pred: predicting ncRNA-protein interaction using sequence and structural information. *Nucleic acids research* **43**, 1370–1379, doi:10.1093/nar/gkv020 (2015).

37. Muppirala, U. K., Honavar, V. G. & Dobbs, D. Predicting RNA-protein interactions using only sequence information. *Bmc Bioinformatics* **12**, 489, doi:10.1186/1471-2105-12-489 (2011).
38. Lu, Q. *et al*. Computational prediction of associations between long non-coding RNAs and proteins. *Bmc Genomics* **14**, 651, doi:10.1186/1471-2164-14-651 (2013).
39. Wang, L. J. & Brown, S. J. BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Res* **34**, W243–W248, doi:10.1093/nar/gkl298 (2006).
40. Terribilini, M. *et al*. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic acids research* **35**, W578–584, doi:10.1093/nar/gkm294 (2007).
41. Cheng, C. W., Su, E. C., Hwang, J. K., Sung, T. Y. & Hsu, W. L. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *Bmc Bioinformatics* **9**(Suppl 12), S6, doi:10.1186/1471-2105-9-S12-S6 (2008).
42. Kumar, M., Gromiha, M. M. & Raghava, G. P. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* **71**, 189–194, doi:10.1002/prot.21677 (2008).
43. Panwar, B. & Raghava, G. P. S. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics* **105**, 197–203, doi:10.1016/j.ygeno.2015.01.005 (2015).
44. Wang, Y., Xue, Z., Shen, G. & Xu, J. PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids* **35**, 295–302, doi:10.1007/s00726-007-0634-9 (2008).
45. Tong, J., Jiang, P. & Lu, Z. H. RISP: A web-based server for prediction of RNA-binding sites in proteins. *Comput Meth Prog Bio* **90**, 148–153, doi:10.1016/j.cmpb.2007.12.003 (2008).
46. Murakami, Y., Spriggs, R. V., Nakamura, H. & Jones, S. PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res* **38**, W412–W416, doi:10.1093/nar/gkq474 (2010).
47. Wang, L. J., Huang, C. Y., Yang, M. Q. & Yang, J. Y. BindN plus for accurate prediction of DNA and RNA-binding residues from protein sequence features. *Bmc Syst Biol* **4**, doi:Artn S310.1186/1752-0509-4-S1-S3 (2010).
48. Carson, M. B., Langlois, R. & Lu, H. NAPS: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Res* **38**, W431–W435, doi:10.1093/nar/gkq361 (2010).
49. Ma, X. *et al*. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins* **79**, 1230–1239, doi:10.1002/prot.22958 (2011).
50. Fernandez, M. *et al*. Prediction of dinucleotide-specific RNA-binding sites in proteins. *Bmc Bioinformatics* **12**, doi:Artn S510.1186/1471-2105-12-S13-S5 (2011).
51. Wang, C. C., Fang, Y. P., Xiao, J. M. & Li, M. L. Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids* **40**, 239–248, doi:10.1007/s00726-010-0639-7 (2011).
52. Walia, R. R. *et al*. 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* **9**, doi:ARTN e9772510.1371/journal.pone.0097725 (2014).
53. Xiong, D. P., Zeng, J. Y. & Gong, H. P. RBRIdent: An algorithm for improved identification of RNA-binding residues in proteins from primary sequences. *Proteins* **83**, 1068–1077, doi:10.1002/prot.24806 (2015).
54. Kim, O. T. P., Yura, K. & Go, N. Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res* **34**, 6450–6460, doi:10.1093/nar/gkl819 (2006).
55. Shulman-Peleg, A., Shatsky, M., Nussinov, R. & Wolfson, H. J. Prediction of interacting single-stranded RNA bases by protein-binding patterns. *Journal of molecular biology* **379**, 299–316, doi:10.1016/j.jmb.2008.03.043 (2008).
56. Maetschke, S. R. & Yuan, Z. Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *Bmc Bioinformatics* **10**, doi:Artn 34110.1186/1471-2105-10-341 (2009).
57. Perez-Cano, L. & Fernandez-Recio, J. Optimal Protein-RNA Area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. *Proteins* **78**, 25–35, doi:10.1002/prot.22527 (2010).
58. Zhao, H. Y., Yang, Y. D. & Zhou, Y. Q. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res* **39**, 3017–3025, doi:10.1093/nar/gkq1266 (2011).
59. Liu, Z. P., Wu, L. Y., Wang, Y., Zhang, X. S. & Chen, L. N. Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics* **26**, 1616–1622, doi:10.1093/bioinformatics/btq253 (2010).
60. Li, S. L., Yamashita, K., Amada, K. M. & Standley, D. M. Quantifying sequence and structural features of protein-RNA interactions. *Nucleic Acids Res* **42**, 10086–10098, doi:10.1093/nar/gku681 (2014).
61. Yang, X. X., Deng, Z. L. & Liu, R. RBRDetector: Improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. *Proteins* **82**, 2455–2471, doi:10.1002/prot.24610 (2014).
62. Miao, Z. C. & Westhof, E. Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res* **43**, 5340–5351, doi:10.1093/nar/gkv446 (2015).
63. Bellucci, M., Agostini, F., Masin, M. & Tartaglia, G. G. Predicting protein associations with long noncoding RNAs. *Nat Methods* **8**, 444–445, doi:10.1038/nmeth.1611 (2011).
64. Wong, K. C., Li, Y., Peng, C. B., Moses, A. M. & Zhang, Z. L. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res* **43**, 10180–10189, doi:10.1093/nar/gkv1134 (2015).
65. Tuvshinjargal, N., Lee, W., Park, B. & Han, K. PRIdictor: Protein-RNA Interaction predictor. *Biosystems* **139**, 17–22, doi:10.1016/j.biosystems.2015.10.004 (2016).
66. Lee, K., Varma, S., SantaLucia, J. Jr. & Cunningham, P. R. *In vivo* determination of RNA structure-function relationships: analysis of the 790 loop in ribosomal RNA. *J Mol Biol* **269**, 732–743, doi:10.1006/jmbi.1997.1092 (1997).
67. Liu, L. & Chen, S. J. Computing the conformational entropy for RNA folds. *J Chem Phys* **132**, 235104, doi:10.1063/1.3447385 (2010).
68. Liu, L. & Chen, S. J. Coarse-grained prediction of RNA loop structures. *PLoS One* **7**, e48460, doi:10.1371/journal.pone.0048460 (2012).
69. Robertson, H. D. Life before DNA. *Science* **264**, 1479–1480, doi:10.1126/science.264.5164.1479 (1994).
70. Montange, R. K. & Batey, R. T. Riboswitches: Emerging themes in RNA structure and function. *Annu Rev Biophys* **37**, 117–133, doi:10.1146/annurev.biophys.37.032807.130000 (2008).
71. Chen, S. J. RNA folding: conformational statistics, folding kinetics, and ion electrostatics. *Annu Rev Biophys* **37**, 197–214, doi:10.1146/annurev.biophys.37.032807.125957 (2008).
72. Mortimer, S. A., Kidwell, M. A. & Doudna, J. A. Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* **15**, 469–479, doi:10.1038/Nrg3681 (2014).
73. de Brevern, A. G., Etchebest, C. & Hazout, S. Bayesian probabilistic approach for predicting backbone structures in terms of protein blocks. *Proteins* **41**, 271–287 (2000).
74. Mattei, E., Ausiello, G., Ferre, F. & Helmer-Citterich, M. A novel approach to represent and compare RNA secondary structures. *Nucleic acids research* **42**, 6146–6157, doi:10.1093/nar/gku283 (2014).
75. Coimbatore Narayanan, B. *et al*. The Nucleic Acid Database: new features and capabilities. *Nucleic acids research* **42**, D114–122, doi:10.1093/nar/gkt980 (2014).
76. Rose, P. W. *et al*. The RCSB Protein Data Bank: redesigned web site and web services. *Nucleic acids research* **39**, D392–401, doi:10.1093/nar/gkq1021 (2011).
77. Suresh, V., Ganesan, K. & Parthasarathy, S. PDB-2-PB: a curated online protein block sequence database. *J Appl Crystallogr* **45**, 127–129, doi:10.1107/S0021889811052356 (2012).

78. Bahadur, R. P., Zacharias, M. & Janin, J. Dissecting protein-RNA recognition sites. *Nucleic acids research* **36**, 2705–2716, doi:10.1093/nar/gkn102 (2008).
79. Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M. & Thornton, J. M. Protein-RNA interactions: a structural analysis. *Nucleic acids research* **29**, 943–954, doi:10.1093/Nar/29.4.943 (2001).
80. Kumar, M., Gromiha, A. M. & Raghava, G. P. S. Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins* **71**, 189–194, doi:10.1002/prot.21677 (2008).
81. Pan, B. C., Xiong, Y. & Steitz, T. A. How the CCA-Adding Enzyme Selects Adenine over Cytosine at Position 76 of tRNA. *Science* **330**, 937–940, doi:10.1126/science.1194985 (2010).
82. Choi, S. & Han, K. Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Comput Biol Med* **43**, 1687–1697, doi:10.1016/j.compbiomed.2013.08.011 (2013).
83. Hainzl, T., Huang, S. & Sauer-Eriksson, A. E. Structure of the SRP19 RNA complex and implications for signal recognition particle assembly. *Nature* **417**, 767–771, doi:10.1038/nature00768 (2002).
84. Rinn, J. L. & Chang, H. Y. Genome Regulation by Long Noncoding RNAs. *Annu Rev Biochem* **81**, 145–166, doi:10.1146/annurev-biochem-051410-092902 (2012).
85. Jeon, Y. & Lee, J. T. YY1 Tethers Xist RNA to the Inactive X Nucleation Center. *Cell* **146**, 119–133, doi:10.1016/j.cell.2011.06.026 (2011).
86. Offmann, B., Tyagi, M. & de Brevern, A. G. Local protein structures. *Curr Bioinform* **2**, 165–202, doi:10.2174/157489307781662105 (2007).
87. Gruber, A. R., Lorenz, R., Bernhart, S. H., Neubock, R. & Hofacker, I. L. The Vienna RNA websuite. *Nucleic acids research* **36**, W70–74, doi:10.1093/nar/gkn188 (2008).
88. Wang, L., Huang, C., Yang, M. Q. & Yang, J. Y. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *Bmc Syst Biol* **4**(Suppl 1), S3, doi:10.1186/1752-0509-4-S1-S3 (2010).

Acknowledgements

We would like to acknowledge the members of Center for Bioinformatics and Systems Biology at Wake Forest School of Medicine, especially Dr. Hua Tan, for valuable discussions and advices. This work was supported by National Institutes of Health [1U01CA166886] (to X. Zhou), and partially supported by NSFC No. 61373105. Funding for open access charge: National Institutes of Health [1U01CA166886].

Author Contributions

L.L. and X.Z. designed the experiments. J.L. and S.V. performed the analyses. L.L., J.L. and X.Z. interpreted the results, and drafted the manuscript. Q.S. provided useful suggestions, and helped in software implement. All authors read and approved the final manuscript.

Additional Information

Supplementary information accompanies this paper at doi:10.1038/s41598-017-00795-4

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.