Sequence analysis

aPRBind: protein-RNA interface prediction by combining sequence and I-TASSER model-based structural features learned with convolutional neural networks

Yang Liu[1], Weikang Gong[1], Yanpeng Zhao[1], Xueqing Deng[1], Shan Zhang[1], and Chunhua Li[1,*]

[1]Faculty of Environmental and Life Sciences, Beijing University of Technology, Beijing 100124, China

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Protein-RNA interactions play a critical role in various biological processes. The accurate prediction of RNA-binding residues in proteins has been one of the most challenging and intriguing problems in the field of computational biology. The existing methods still have a relatively low accuracy especially for the sequence based *ab-initio* methods.

Results: In this work, we propose an approach aPRBind, a convolutional neural network (CNN)-based *ab-initio* method for RNA-binding residue prediction. aPRBind is trained with sequence features and structural ones (particularly including residue dynamics information and residue-nucleotide propensity developed by us) that are extracted from the predicted structures by I-TASSER. The analysis of feature contributions indicates the sequence features are most important, followed by dynamics information, and the sequence and structural features are complementary in binding site prediction. The performance comparison of our method with other peer ones on benchmark dataset shows that aPRBind outperforms some state-of-the-art ab-initio methods. Additionally, aPRBind can give a better prediction for the modeled structures with TM-score $\geqslant$ 0.5, and meanwhile since the structural features are not very sensitive to the refined 3-dimensional structures, aPRBind has only a marginal dependence on the accuracy of the structure model, which allows aPRBind to be applied to the RNA-binding site prediction for the modeled or unbound structures.

Availability: The source code is available at https://github.com/ChunhuaLiLab/aPRbind.

Contact: chunhuali@bjut.edu.cn

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

1 Introduction

Protein-RNA interactions play pivotal roles in a wide variety of biological processes, such as gene expression and regulation, protein synthesis and virus assembly (Keene, 2007). The aberrations in protein-RNA recognition may lead to many diseases (Lukong *et al*., 2008). Thus, the reliable identification of RNA-binding residues on proteins is an important and also a challenging problem, which is critical for understanding the recognition mechanism of protein-RNA interactions and is also helpful for complex structure prediction and drug design.

Experimental methods are very costly and time-consuming. Thus, there is a growing demand for the development of computational methods to predict protein-RNA binding sites. In the past decade, many computational approaches have been established. Based on the features they use, these methods can be divided into two categories: sequence-based (Carson *et al*., 2010; El-Manzalawy *et al*., 2016; Kumar *et al*., 2008; Murakami *et al*., 2010; Terribilini *et al*., 2007; Walia *et al*., 2014) and structure-based methods (Chen and Lim, 2008; Kim *et al*., 2006;

Maetschke and Yuan, 2009; Perez-Cano and Fernandez-Recio, 2010; Tang *et al*., 2017; Towfic *et al*., 2010).

For sequence-based methods, the often-used features for a target protein include position-specific scoring matrix (PSSM), amino acid physicochemical properties, predicted solvent accessibility and so on. PSSM, a common representation of sequence evolutionary features, has been widely used in most of the predictors due to the fact that the interface residues, owing to experiencing relatively higher evolutionary pressures, undergo fewer mutations (more conserved) than the other surface ones. Besides the standard PSSM, the two main improved PSSM profiles have been developed and used in RNA-binding residue prediction which are scaled respectively by a sliding window (El-Manzalawy *et al*., 2016; Li *et al*., 2014; Walia *et al*., 2014), and a smooth processing followed by a sliding window (called smoothed PSSM) (Cheng *et al*., 2008). Different from the smoothed PSSM where the smooth processing and sliding window are both based on the sequence neighbors of the target residue, the SNB-PSSM (spatial neighbor based position-specific scoring matrix) proposed in our previous work adopts a spatial neighbor based smooth processing and a structure window scheme to encode the evolutionary information. SNB-PSSM achieves a better prediction of RNA-binding residues than the smoothed PSSM, which can be explained to some extent by our detection that the conserved interface residues often occur clustered together (Yang *et al*., 2020). For structured-based methods, besides sequence features, the often-used structure-derived features include structural and topological features and others. The former mainly includes secondary structure type, geometrical cleft (Chen and Lim, 2008), complex network property (Maetschke and Yuan, 2009), voronoi contacts and structural neighbors (Tang *et al*., 2017), and the latter basically includes solvent accessibility (Maetschke and Yuan, 2009) and electrostatic potential (Chen and Lim, 2008).

In addition to the above features, there are still other ones we need to explore to advance RNA-binding site prediction. As we know, protein-RNA interactions are characterized by both sequence and structural recognition specificities (Chen *et al*., 2004; Jeong *et al*., 2003; Jones *et al*., 2001; Perez-Cano and Fernandez-Recio, 2010). Based on the fact, we extracted a 60×8 residue-nucleotide pairwise propensity potential with secondary structure information considered in a previous work, which shows a good performance in the discrimination of near-native complex structures (Li *et al*., 2012). Combined with physical energy items, this potential can capture at least one docked mode among top 5 on the scoring list which has no lower than 50% native interface residues and nucleotides on its interface for 91.4% protein-RNA interactions (Zhang *et al*., 2017), displaying a good interface prediction ability. Therefore, here we try to apply it as a feature to RNA-binding residue prediction. Additionally, besides sequence and structural features, protein dynamics properties play an important role in protein-protein/ligand specific recognition and interactions, which have been used to predict binding key residues, binding hot spots (Melo *et al*., 2016), and even allosteric sites (Taguchi and Kitao, 2016). We also utilized residue dynamics properties to successfully identify the key association residues for the interaction between snRNA and human U1A protein (Han *et al*., 2019). Thus, the residue dynamics properties are also what we want to apply to RNA-binding site prediction.

In recent years with the development of protein structure prediction methods, the high prediction accuracy in case of no homologous templates available enables us to construct the predicted structure-based binding site prediction method. I-TASSER structure predictor developed by Zhang *et al*., a threading based method, performs pretty well even for the new fold targets, which has been keeping ahead in the last decade of the community-wide CASP (Critical Assessment of protein Structure Prediction) experiments (Yang *et al*., 2015).

In this work, we propose a sequence based ab initio algorithm aPRBind (*ab-initio* Protein-RNA Binding site prediction) to predict RNA-binding residues in proteins, which utilizes the sequence features including SNB-PSSM based evolutionary information, and I-TASSER model-based structural ones including residue dynamics properties and residue-nucleotide propensities. These features are learned by a deep convolutional neural network model.

2 Methods

2.1 Data

In this work, we use the benchmark dataset RB198 (El-Manzalawy *et al*., 2016) as the training set. The data in RB198 were derived by removing the complexes from protein-RNA complexes in PDB that meet any criterion of the following: i) structure resolution worse than 3.5 Å; ii) protein residues < 40 or RNA nucleotides < 5; iii) interface residues < 3; iv) sequence identity > 30% with other chains. The RB198 dataset has 134 complexes with 198 protein chains.

As many protein-RNA interface prediction servers have been compared with each other by EL-Manzalawy *et al*. (El-Manzalawy *et al*., 2016) on the benchmark dataset RB111, we use RB111 as the independent verification dataset to compare our method with other servers.

For the two datasets, an interface residue of a protein is defined as the one that has at least one atom closer than 5 Å to any atom of its partner RNA. All the residues are labeled '1' or '0' depending on whether they are binding or non-binding residues respectively. Totally, there are 7878 binding residues and 43150 non-binding ones in RB198 dataset, and the corresponding numbers are 3305 and 34255 in RB111 dataset respectively.

2.2 Structure construction with I-TASSER

I-TASSER, a threading based protein structure prediction method, developed by Zhang *et al*. (Yang *et al*., 2015) is used to construct protein structures from their sequences. It hierarchically constructs full-length models by iteratively reassembling structure fragments extracted from the threading templates. The parameters are set to default values. Additionally, it should be pointed out that in protein structure construction, all the templates with sequence identity > 30% to the query sequence are excluded from the template library. Finally, the first model with the highest TM-score is selected from the five models as the constructed structure.

2.3 Features extraction

The features we use include sequence and structural ones. All the following features except for SPIDER3-based features and physicochemical properties belong to the structural ones that are extracted from the modeled structures by I-TASSER.

Spatial neighbor based position-specific scoring matrix (SNB-PSSM): Considering the conserved interface residues often occur clustered together in protein tertiary structures (Ahmad *et al*., 2008; Capra and Singh, 2007; Guharoy and Chakrabarti, 2010), we proposed in previous work a new encoding scheme of evolutionary information, i.e., spatial neighbor based PSSM (SNB-PSSM) which is different from the smoothed PSSM (Cheng *et al*., 2008).

In SNB-PSSM method, first, for a protein with N residues, the size of PSSM matrix is 20 × N with the evolutionary information for each position. Then, a spatial neighbor based smooth processing is carried out, making the evolutionary score of a target residue being an average value of the evolutionary scores over the residues whose Cα atoms are within 7.5 Å from that of the target one. Finally, a spatial neighbor based window scheme is adopted where the evolutionary information of a target residue is encoded with the smoothed evolutionary scores of the spatially nearest 25 residue positions to the target one. Thus, for a target residue, its evolutionary information is encoded into a 20 × 25 matrix. This encoding process considers the evolution of the spatial neighbors around a target residue.

Interface propensity (IP): The residue-nucleotide propensities (60×8) with secondary structure information of proteins and RNAs considered (Li *et al*., 2012), extracted by us from 251 non-redundant protein-RNA complexes, are used as a feature in this work. In the propensities, protein and RNA secondary structures are classified into three and two classes based on their interface propensities respectively. The propensity of a specific residue-nucleotide pair is calculated from its observed probability at interfaces divided by its expected probability. Here, the interface propensity of a residue type with a certain class of secondary structures is represented as an average value of its pairwise propensities for eight kinds of nucleotides.

Residue fluctuation dynamics: The features of residue fluctuation dynamics are calculated from Gaussian network model (GNM) which, based on a harmonic potential, has been proven to be a reliable method for reproducing the intrinsic dynamics of biomacromolecules (Bahar *et al*., 1997). The total internal potential energy of the network of N nodes can be written as

$$V = \frac{1}{2}\gamma\left[\Delta \boldsymbol{R}^T (\boldsymbol{\Gamma} \otimes \boldsymbol{E}) \Delta \boldsymbol{R}\right] \quad (1)$$

where γ is the harmonic force constant of the springs, the column vector $\Delta\boldsymbol{R}$ represents the fluctuation of the N nodes, the superscript T denotes the transpose, $\boldsymbol{E}$ is the 3×3 identity matrix, $\otimes$ is the matrix direct product and $\boldsymbol{\Gamma}$ is the N×N symmetric Kirchhoff matrix. The mean-square fluctuation of the ith residue can be expressed as

$$\left\langle \Delta R_i \cdot \Delta R_i \right\rangle = \frac{3k_B T}{\gamma}\left[\Gamma^{-1}\right]_{ii} \propto \sum_{k=2}^{N} \lambda_k^{-1} \left[u_k\right]_i^2 \quad (2)$$

where k_B is the Boltzmann constant and T is the absolute temperature. In GNM, as slow modes contribute majority to the residue fluctuation (Bahar *et al*., 1998), we calculate the relative residue fluctuations (with $3k_BT/\gamma$ ignored) from the first m (1-6) motional modes as residue fluctuation features denoted as $\boldsymbol{F_1}$ to $\boldsymbol{F_6}$.

SPIDER3-based features: The program SPIDER3 is used to predict protein secondary structure (SS) state and solvent accessibility (SA) from sequence (Heffernan, et al., 2017). In addition, the features computed from the sequence include the main chain torsional angles (φ and ψ), the main chain angles between Cα atoms (θ and τ), and the half-sphere exposure (HSE) within upper and down half spheres (HSEα_up and HSEα_down) defined according to the neighboring Cα-Cα vectors of each considered residue.

Depth and protrusion indices (DPX and CX): The program PSAIA (Mihel *et al*., 2008) is employed to obtain the residue depth and protrusion indices. The depth index (DPX) of an atom is defined as its distance to the closest solvent accessible atom. The protrusion index (CX) is defined by V_{ext}/V_{int}, where V_{int} is the volume occupied by non-hydrogen atoms within a sphere of fixed radius (here 10 Å, the default value) centered on the considered atom, and the remaining volume of the sphere is Vext. For a residue, its DPX and CX are the average values of DPX and CX over its all non-hydrogen atoms respectively.

Topology characteristics: We take use of the Python programming package NetworkX (v1.11) to calculate the topology characteristics of each residue, which include degree, cluster coefficient, closeness centrality, betweenness and degree centrality. Here, Cα atoms are taken as nodes and two nodes are connected by an edge if they are less than 7.5 Å.

Physicochemical properties: The ten physicochemical properties of each residue type are obtained from the AA index database (Kawashima *et al*., 2008), which include number of atoms, number of electrostatic charges, number of potential hydrogen bonds, molecular mass (M-mass), hydrophobicity, hydrophilicity, polarity, polarizability, propensity and average accessible surface area. In addition, amino acids are divided into positively-, negatively- and non-charged three groups based on their charge types, which are represented by 1, -1 and 0 respectively.

2.4 Convolutional neural networks

The convolutional neural network (CNN) model (Defferrard *et al*., 2016) is implemented in Tensorflow (Rampasek and Goldenberg, 2016), which contains two convolutional, two max pooling and two fully connected layers. 2D CNN architecture is adopted. A residue is described by a feature vector of 24×24 dimensions (500 dimensions for SNB-PSSM, 36 for other features and 40 with zero values for nothing). The two convolutional layers have 8 and 16 filters, respectively, and each filter (kernel) is 2×2 size. The max pooling of size 2 applies to both of the convolutional layers. The two fully connected layers have 576 and 64 units consequently. The parameterized rectified linear unit (PReLU) is used as the activation function, and the dropout rate is set to 0.5. The cross entropy loss is minimized for the true label and the sampled negative classes. We use 50k iterations with a mini-batch size of 100.

2.5 Architecture of aPRBind algorithm

The flowchart of aPRBind is shown in Figure 1. The query protein sequence is submitted to two programs PSI-BLAST and SPIDER3 to generate the PSSM profile and SPIDER3-based features. In addition, the sequence is fed into I-TASSER Suite (Yang *et al*., 2015) to generate the structure model. Then the structural features including SNB-PSSM profile, interface propensities, topology and dynamics properties and so on are calculated based on the constructed structure model. The aPRBind prediction is tuned on RB198 training set by taking use of convolutional neural networks.

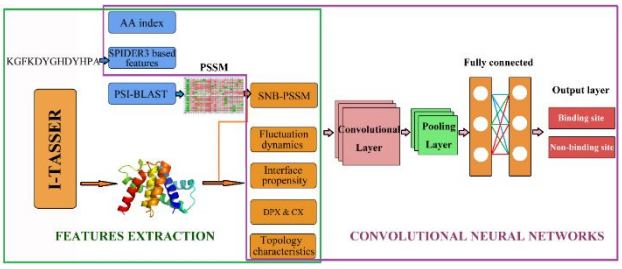

Fig. 1. Flowchart of aPRBind for RNA-binding site prediction. The used features are derived from sequences and I-TASSER models, which are learned with a deep convolutional neural network model.

2.6 Performance evaluation measures

The CNN model is trained through 5-fold cross validation, and the independent dataset test is adopted to examine the effectiveness of the classifier. The predictive performance of the classifier is assessed with the overall accuracy (*ACC*), sensitivity (*SN*), specificity (*SP*), and Matthews correlation coefficient (*MCC*) that are defined as follows

$$ACC = \frac{TP + TN}{TP + FP + TN + FN} \tag{3}$$

$$SN = \frac{TP}{TP + FN} \tag{4}$$

$$SP = \frac{TN}{TN + FP} \tag{5}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TN + FN)(TN + FP)(TP + FN)(TP + FP)}} \tag{6}$$

where the true positive (*TP*), false positive (*FP*), true negative (*TN*) and false negative (*FN*) are obtained by comparing the predicted label for each residue with the actual one.

3 Results

3.1 Structure construction by I-TASSER

I-TASSER program was used to generate atomic structure models for the protein sequences in RB198 and RB111 datasets. The accuracy of the I-TASSER model is estimated by the template modeling score (TM-score) (Xu and Zhang, 2010) that is in range of 0 to 1. Generally, the I-TASSER models with TM-score ≥ 0.5 are considered to be correct folds.

Figure S1 shows the distribution of TM-score values of the modelled structures from 309 protein chains in RB198 and RB111 datasets. From Figure S1, most of the models (82.5%) have a TM-score above 0.5, and the models with a TM-score > 0.8 account for 21.4% of all the models. For the 309 models, their average value of TM-score is 0.65. The results reveal that although all the homologous templates with sequence identity > 30% to the target are excluded from the template library, the majority of the I-TASSER models have the correct folds.

3.2 Analyses of feature contributions

The maximum Relevance Minimum Redundancy (mRMR) (Peng *et al*., 2005) (see supplementary materials) was used to evaluate the importance of the features in RNA-binding residue prediction. As the residue evolutionary information has been proven to be the most discriminative and effective feature in many predictors *(El-Manzalawy et al., 2016; Kumar et al*., 2008; Li *et al*., 2014; Walia *et al.*, 2014), here we evaluated the importance of other features. Table S1 gives the top 20 features obtained by applying mRMR on RBP198 dataset, with their average values over the interface and non-interface residues, along with *P*-values.

From Table S1, the average value of relative solvent accessibility (RSA) over the interface residues is higher than that over the non-interface ones by 45.1%, which implies that the residues with high RSA values are likely to appear at the interface. The average value of P(H) over the interface residues is lower than that over the non-interface ones by 20.5%. Our previous work found that the residues in alpha-helix do not prefer to appear at RNA-binding interfaces, consistent with the current observation to some extent. (Li *et al*., 2012). The charge type of amino acids is ranked the third. As nucleotides are negatively charged, electrostatic interactions play an important role in protein-RNA interactions with the positively charged amino acids preferring to bind to RNA molecules (Li *et al*., 2012). Here, we also calculated the interface (against non-interface) propensity of each kind of amino acids in RB198 and RB111 datasets. The results are shown in Figure S2. From Figure S2, amino acids Arg, His and Lys have the highest propensity values of 1.94, 1.41 and 1.38, respectively, which demonstrates the positively charged amino acids prefer to bind to RNA molecules. In addition, the relative residue fluctuations contributed by the first *m* (1-6) motional modes ($\boldsymbol{F_1 \sim F_6}$) are all ranked within top 14. The average relative residue fluctuation of interfaces is higher than that of non-interfaces by at least 146.7%, which hints that generally, the interface residues have relative high flexibility. The average HSEα_up of interface residues is lower than that of non-interface ones by 16.4%. The average IP of interface residues is higher than that of non-interface ones, which is consistent with the definition of IP itself. The average depth index (DPX) is ranked top 10. On the average, the interface residues are prone to having low DPX values than non-interface ones. To sum up, besides the sequence-based features, the structural and dynamical features play important roles in RNA-binding residue prediction.

Table 1. Average results of 5-fold cross validation experiments of CNN models with different combinations of features considered on RB198.

Combinations of features	*SN*	*SP*	*ACC*	*MCC*
SNB-PSSM+CX/DPX	0.44	0.83	0.64	0.29
SNB-PSSM+Topology	0.50	0.79	0.65	0.30
SNB-PSSM+IP	0.47	0.83	0.65	0.32
SNB-PSSM+Dynamics	0.50	0.81	0.66	0.33
SNB-PSSM+AA	0.49	0.82	0.66	0.33
SNB-PSSM+SPIDER3	0.51	0.81	0.66	0.34
SNB-PSSM+AA+SPIDER3 +CX/DPX	0.56	0.79	0.68	0.36
SNB-PSSM+AA+SPIDER3 +Topology	0.54	0.83	0.69	0.39
SNB-PSSM+AA+SPIDER3 +IP	0.58	0.80	0.69	0.39
SNB-PSSM+AA+SPIDER3 +Dynamics	0.59	0.80	0.70	0.40
SNB-PSSM+AA+SPIDER3 +IP+CX/DPX+Topology	0.61	0.78	0.70	0.40
SNB-PSSM+AA+SPIDER3 +IP+CX/DPX+Dynamics	0.62	0.81	0.72	0.44
All features	0.65	0.82	0.74	0.48

Furtherly, in order to detect the contributions of the features to the CNN model, we performed predictions using the CNN models with different combinations of features considered and compared the results. As the evolutionary information has been proven to be the most discriminative and effective feature in many predictors (El-Manzalawy *et al*., 2016; Kumar *et al*., 2008; Li *et al*., 2014; Walia *et al*., 2014), the SNB-PSSM based feature was taken into account in all the models. The other 36 features are classified into six types: two types of sequence-based features (including AA index and SPIDER3-based features) and four types of structure-based features (including CX/DPX, IP, topology and dynamics

features). Table 1 shows the average results of 5-fold cross validation experiments of different CNN models on RB198. For the combinations of SNB-PSSM based feature with different single types of features, the one with SPIDER3-based features considered presents the best effectiveness in terms of *MCC* value, followed by the ones with AA index, dynamics, IP, topology and CX/DPX features considered respectively. Thus, the sequence-based features are relatively more important, and in addition the dynamics features derived from structures are also critical, which is consistent with the results from the single feature analysis by mRMR. When the sequence-based features are combined into the four types of structure-based features respectively, the further advances are observed with the combination where dynamics features are considered attaining the best result (*MCC* of 0.40), which suggests their complementarity in binding site prediction. Additionally, the comparison between the performances with all the features but topology and dynamics features respectively considered hints that the dynamics features are relatively more important than the topology ones. Finally, the performance with all the features considered achieves the best result with *SN*, *SP*, *ACC* and *MCC* of 0.65, 0.82, 0.74 and 0.48, respectively.

As a whole, the above results show that the sequence-based features are most important to the prediction made by CNN model, and the dynamics features first applied on protein-RNA binding site prediction to our knowledge are also crucial. The sequence and structure-based features are complementary in binding site prediction.

Table 2. Comparison of aPRBind with SVM and RF models and some existing protein-RNA interface prediction servers on RB111

Methods	*ACC*	*SN*	*SP*	*MCC*
aPRBind	0.86	0.48	0.90	0.32
SVM	0.86	0.46	0.89	0.29
RF	0.85	0.45	0.88	0.27
FastRNABindR	0.75	0.61	0.76	0.24
RNABindR v2	0.72	0.63	0.73	0.22
BindN+	0.84	0.43	0.87	0.24
PPRInt	0.76	0.48	0.79	0.18

3.3 Performance of aPRBind on the independent test set and comparison with existing prediction servers

We tested the trained CNN model with all features considered on the independent test set RB111, with the results shown in Table 2. Meanwhile, in order to examine the effect of the CNN deep leaning method on the prediction performance, we tested the trained support vector machine (SVM) and random forest (RF) models via the same method on RB111. In addition, Table 2 gives the test results of the four sequence-based protein-RNA interface prediction servers (FastRNABindR *(El-Manzalawy et al*., 2016), RNABindR v2 (Walia, et al., 2012), BindN+ (Wang *et al*., 2010) and PPRInt (Kumar *et al*., 2008)) on RB111 dataset. From Table 2, the SVM model is a little better than the RF model by all of the evaluation metrics. Our method aPRBind attains *ACC*, *SN*, *SP* and *MCC* of 0.86, 0.48, 0.90 and 0.32 respectively, which consistently outperforms SVM and RF models with *MCC* increasing by 10.3% and 18.5% respectively. Compared with the four prediction servers FastRNABindR, RNABindR v2, BindN+ and PPRInt, aPRBind achieves the highest *ACC*, *SP*, and *MCC* values, with *MCC* increasing by 33.3%, 45.5%, 33.3% and 77.8% respectively.

To sum up, among the seven sequence-based *ab-initio* methods, aPRBind shows a better performance in interface residue prediction, which we think is mainly due to the consideration of the structure-based features (such as dynamics features, interface propensity and SNB-PSSM based evolutionary information) and the usage of the convolutional neural network algorithm.

3.4 Impact of accuracy of I-TASSER models on binding site prediction

Since the structure-based features used in aPRBind are extracted from I-TASSER models, we want to know whether the quality of the predicted models has an effect on its prediction performance. Commonly, the I-TASSER models with TM-score ≥ 0.5 are considered to be correct folds. Thus in order to detect the issue, we split all the models into three groups with TM-score < 0.5, 0.5 ≤ TM-score < 0.7, and TM-score ≥ 0.7 respectively. Table 3 shows the performances of aPRBind on these groups. From Table 3, the performance on the group with TM-score < 0.5 is worse than those on the groups with TM-score ≥ 0.5. And in particular, for the models with TM-score < 0.3 (2XLK:A and 3T5N:A), their *MCC* values are lower, 0.18 and 0.19 respectively. Comparing the performances on the latter two groups with TM-score ≥ 0.5, we can see that the one on the group with TM-score ≥ 0.7 is a little better. Furtherly, we performed aPRBind on all the experimental structures from RB111, with the result also shown in Table 3. It can be seen that the performance of aPRBind on experimental structures is almost the same with that on the I-TASSER models with TM-score ≥ 0.7, with only *ACC* value increasing by 0.01.

From the above results, it can be concluded that the structural features based on the correct folds are helpful for aPRBind to perform binding site predictions. In addition, aPRBind has a good robustness against the structural variations as long as the residue positions are approximately correct, which is mainly because the structural features used in aPRBind are at a coarse-grained level, not very sensitive to the refined 3-dimensional structures.

Table 3. Comparison of aPRBind with SVM and RF models and some existing protein-RNA interface prediction servers on RB111

Structures (numbers)	*ACC*	*SN*	*SP*	*MCC*
TM-score < 0.5 (17)	0.86	0.48	0.90	0.32
0.5 ≤ TM-score < 0.7 (39)	0.72	0.63	0.73	0.22
TM-score ≥ 0.7 (55)	0.84	0.43	0.87	0.24
Experimental structures (111)	0.76	0.48	0.79	0.18

3.5 Case study

Figure 2 shows the prediction results by aPRBind on the two protein examples whose I-TASSER models have a TM-score < 0.5 and > 0.5 respectively, and also gives the prediction results on their experimental structures for comparison. The first example is the nucleocapsid protein tetramer (PDB ID: 4H5P:A) (Raymond *et al*., 2012). The TM-score of its I-TASSER model is 0.38. On the I-TASSER model, aPRBind attains *ACC, SN, SP* and *MCC* of 0.87, 0.68, 0.90 and 0.54 respectively, and its corresponding results on the experimental structure are 0.87, 0.65, 0.91 and 0.52 respectively (see Figure 2 (A) and (B)).

The second example is a RBP, human Dicer Platform-PAZ-Connector Helix cassette (PDB ID: 4NGD:A) (Tian *et al*., 2014). The TM-score of

its I-TASSER model is 0.53. On the I-TASSER model, aPRBind attains *ACC, SN, SP* and *MCC* of 0.91, 0.68, 0.93 and 0.52 respectively, and its corresponding results on the experimental structure are 0.91, 0.77, 0.92 and 0.55, respectively (see Figure 2(C) and (D)).

From the two examples, we can find that although the structures are not well constructed by I-TASSER, aPRBind obtains the relatively satisfactory results. For 4H5P:A, the low TM-score is mainly due to the not good prediction for the protruding helix. Thus, the prediction result on its I-TASSER model is similar or a little better than that on the experimental structure. For 4NGD:A, the secondary structures of the interface are not well predicted by I-TASSER, but the arrangement is similar to the experimental one. Thus, the prediction result on the modeled one are still satisfactory.

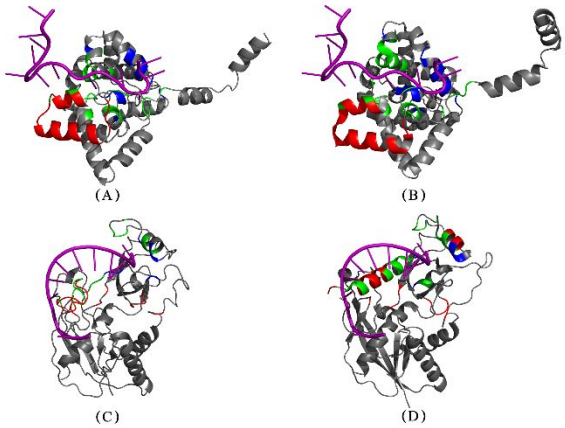

Fig. 2. Predictions of RNA-binding residues with aPRBind on the modeled structures by I-TASSER and experimentally solved ones respectively. (A) and (B) the case of PDB ID: 4H5P:A with the results *SN* = 0.68, *SP* = 0.90, *ACC* = 0.87, *MCC* = 0.54 for the I-TASSER model (TM-score = 0.38) and *SN* = 0.65, *SP* = 0.91, *ACC* = 0.87, *MCC* = 0.52 for the experimental structure respectively. (C) and (D) the case of PDB ID: 4NGD:A with the results *SN* = 0.68, *SP* = 0.93, *ACC* = 0.91, *MCC* = 0.52 for the I-TASSER model (TM-score = 0.53) and *SN* = 0.77, *SP* = 0.92, *ACC* = 0.91, *MCC* = 0.55 for the experimental structure respectively. The protein structures are shown in gray cartoon. The *TP*, *FP* and *FN* results are shown in green, red and blue, respectively.

4 Conclusions

We propose a sequence-based ab initio approach aPRBind (*ab-initio* Protein-RNA Binding site prediction) to predict RNA-binding residues in proteins, where the sequence features and structural features extracted from I-TASSER models are learned by a deep convolutional neural network model. In structure prediction, all the homologous templates with sequence identity > 30% to the target are excluded from the template library in order to meet the general situation. The result shows that although the above process is performed, the majority of the I-TASSER models (82.6%) have a correct fold with TM-score no less than 0.5. The analysis on features' contributions indicates that the sequence features are most important to the prediction, the dynamics features are also crucial, and the sequence and structure-based features are complementary in binding site prediction. The performance of aPRBind on the independent test set illustrates that our method can give a better prediction for the correctly modeled folds. And in addition aPRBind has a good robustness against the structural variations as long as the residue positions are approximately correct, which is mainly because the structural features used in aPRBind are at a coarse-grained level, not very sensitive to the refined 3-dimensional structures. Our method outperforms the SVM and RF models, and some classic sequence-based prediction servers FastRNABindR, RNABindR v2, BindN+ and PPRInt. This work is helpful for strengthening our understanding of protein-RNA recognition and interactions, and can be used for protein-RNA docking prediction and binding hot spot exploration.

Acknowledgements

We are grateful to Dr. Yang Zhang for helping us with the usage of the I-TASSER servers.

Funding

This work was supported by the National Natural Science Foundation of China [31971180, 11474013].

Conflict of Interest: none declared.

References

Ahmad, S. *et al*. (2008) Protein-DNA interactions: structural, thermodynamic and clustering patterns of conserved residues in DNA-binding proteins, *Nucleic Acids Res.*, **36**, 5922-5932.

Bahar, I. *et al*. (1998) Vibrational dynamics of folded proteins: Significance of slow and fast motions in relation to function and stability, *Phys. Rev. Lett.*, **80**, 2733-2736.

Bahar, I. *et al*. (1997) Direct evaluation of thermal fluctuations in proteins using a single-parameter harmonic potential, *Fold Des.*, **2**, 173-181.

Capra, J.A. and Singh, M. (2007) Predicting functionally important residues from sequence conservation, *Bioinformatics*, **23**, 1875-1882.

Carson, M.B. *et al*. (2010) NAPS: a residue-level nucleic acid-binding prediction server, *Nucleic Acids Res.*, **38**, W431-W435.

Chen, Y. *et al*. (2004) A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys, *Nucleic Acids Res.*, **32**, 5147-5162.

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.

Defferrard, M. *et al*. (2016) Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering, *30th Conference on Neural Information Processing Systems (NIPS)*. *Neural Information Processing Systems (NIPS)*.

El-Manzalawy, Y. *et al*. (2016) FastRNABindR: Fast and Accurate Prediction of Protein-RNA Interface Residues, *PloS one*, **11**, e158445.

Guharoy, M. and Chakrabarti, P. (2010) Conserved residue clusters at protein-protein interfaces and their use in binding site identification, *BMC Bioinformatics*, **11**, 286.

Han, Z. *et al*. (2019) Interpreting the Dynamics of Binding Interactions of snRNA and U1A Using a Coarse-Grained Model, *Biophys. J*, **116**, 1625-1636.

Heffernan, R. *et al*. (2017) Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility, *Bioinformatics*, **33**, 2842-2849.

Jeong, E. *et al*. (2003) Discovering the interaction propensities of amino acids and nucleotides from protein-RNA complexes, *Mol. Cells*, **16**, 161-167.

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

Kawashima, S. *et al*. (2008) AAindex: amino acid index database, progress report 2008, *Nucleic Acids Res.*, **36**, D202-D205.

Keene, J.D. (2007) RNA regulons: coordination of post-transcriptional events, *Nat. Rev. Genet.*, **8**, 533-543.

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, C.H. *et al*. (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys, *Proteins*, **80**, 14-24.

Li, S. *et al*. (2014) Quantifying sequence and structural features of protein-RNA interactions, *Nucleic Acids Res.*, **42**, 10086-10098.

Lukong, K.E. *et al*. (2008) RNA-binding proteins in human genetic disease, *Trends Genet.*, **24**, 416-425.

Maetschke, S.R. and Yuan, Z. (2009) Exploiting structural and topological information to improve prediction of RNA-protein binding sites, *BMC Bioinformatics*, **10**, 341.

Melo, R. *et al*. (2016) A Machine Learning Approach for Hot-Spot Detection at Protein-Protein Interfaces, *Int. J. Mol. Sci.*, **17**.

Mihel, J. *et al*. (2008) PSAIA - protein structure and interaction analyzer, *BMC Struct. Biol.*, **8**, 21.

Murakami, Y. *et al*. (2010) PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences, *Nucleic Acids Res.*, **38**, W412-W416.

Peng, H. *et al*. (2005) Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy, *IEEE Trans Pattern Anal Mach Intell*, **27**, 1226-1238.

Perez-Cano, L. and Fernandez-Recio, J. (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins, *Proteins*, **78**, 25-35.

Rampasek, L. and Goldenberg, A. (2016) TensorFlow: Biology's Gateway to Deep Learning? *Cell Syst.*, **2**, 12-14.

Raymond, D.D. *et al*. (2012) Phleboviruses encapsidate their genomes by sequestering RNA bases, *Proc. Natl. Acad. Sci. USA*, **109**, 19208-19213.

Taguchi, J. and Kitao, A. (2016) Dynamic profile analysis to characterize dynamics-driven allosteric sites in enzymes, *Biophys. Physicobiol.*, **13**, 117-126.

Tang, Y. *et al*. (2017) A boosting approach for prediction of protein-RNA binding residues, *BMC Bioinformatics*, **18**, 465.

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

Tian, Y. *et al*. (2014) A phosphate-binding pocket within the platform-PAZ-connector helix cassette of human Dicer, *Mol. Cell*, **53**, 606-616.

Towfic, F. *et al*. (2010) Struct-NB: predicting protein-RNA binding sites using structural features, *Int. J. Data Min. Bioinform.*, **4**, 21-43.

Walia, R.R. *et al*. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art, *BMC Bioinformatics*, **13**, 89.

Walia, R.R. *et al*. (2014) 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**, e97725.

Wang, L. *et al*. (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features, *BMC Syst. Biol.*, **4 Suppl 1**, S3.

Xu, J. and Zhang, Y. (2010) How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, **26**, 889-895.

Yang, J. *et al*. (2015) The I-TASSER Suite: protein structure and function prediction, *Nat. Methods*, **12**, 7-8.

Yang, Z. *et al*. (2020) Analyses on clustering of the conserved residues at protein-RNA interfaces and its application in binding site identification, *BMC Bioinformatics*, **21**, 57.

Zhang, Z. *et al*. (2017) A combinatorial scoring function for protein-RNA docking, *Proteins*, **85**, 741-752.