Received: 1 May 2018 | Revised: 10 October 2018 | Accepted: 29 November 2018

DOI: 10.1002/prot.25639

WILEY PROTEINS

RESEARCH ARTICLE

Partner-specific prediction of RNA-binding residues in proteins: A critical assessment

Yong Jung[1,2,5] | Yasser EL-Manzalawy[2,4,6] | Drena Dobbs[7,8] | Vasant G. Honavar[1,2,3,4,5,6]

[1]Bioinformatics and Genomics Graduate Program, Pennsylvania State University, University Park, Pennsylvania

[2]Artificial Intelligence Research Laboratory, Pennsylvania State University, University Park, Pennsylvania

[3]Institute for Cyberscience, Pennsylvania State University, University Park, Pennsylvania

[4]Clinical and Translational Sciences Institute, Pennsylvania State University, University Park, Pennsylvania

[5]The Huck Institutes of the Life Sciences, Pennsylvania State University, University Park, Pennsylvania

[6]College of Information Sciences and Technology, Pennsylvania State University, Pennsylvania

[7]Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa

[8]Department of Genetics, Development, and Cell Biology, Iowa State University, Ames, Iowa

Correspondence
Vasant G. Honavar; College of Information Sciences and Technology, Pennsylvania State University, University Park, PA.
Email: vhonavar@ist.psu.edu

Funding information
Indian Institute of Science; Pennsylvania State University; Pennsylvania State University; Huck Institutes of the Life Sciences; National Institutes of Health, Grant/Award Number: NCATS UL1 TR002014-01; National Science Foundation, Grant/Award Number: ACI 1640834

Abstract

RNA-protein interactions play essential roles in regulating gene expression. While some RNA-protein interactions are "specific", that is, the RNA-binding proteins preferentially bind to particular RNA sequence or structural motifs, others are "non-RNA specific." Deciphering the protein-RNA recognition code is essential for comprehending the functional implications of these interactions and for developing new therapies for many diseases. Because of the high cost of experimental determination of protein-RNA interfaces, there is a need for computational methods to identify RNA-binding residues in proteins. While most of the existing computational methods for predicting RNA-binding residues in RNA-binding proteins are oblivious to the characteristics of the partner RNA, there is growing interest in methods for partner-specific prediction of RNA binding sites in proteins. In this work, we assess the performance of two recently published partner-specific protein-RNA interface prediction tools, PS-PRIP, and PRIdictor, along with our own new tools. Specifically, we introduce a novel metric, RNA-specificity metric (RSM), for quantifying the RNA-specificity of the RNA binding residues predicted by such tools. Our results show that the RNA-binding residues predicted by previously published methods are oblivious to the characteristics of the putative RNA binding partner. Moreover, when evaluated using partner-agnostic metrics, RNA partner-specific methods are outperformed by the state-of-the-art partner-agnostic methods. We conjecture that either (a) the protein-RNA complexes in PDB are not representative of the protein-RNA interactions in nature, or (b) the current methods for partner-specific prediction of RNA-binding residues in proteins fail to account for the differences in RNA partner-specific versus partner-agnostic protein-RNA interactions, or both.

KEYWORDS

partner-specific protein-RNA binding, performance evaluation, protein-RNA interactions, protein-RNA Interface prediction, RNA-specificity metric

1 | INTRODUCTION

Protein-RNA interactions play crucial roles in regulating gene expression.[1–4] RNAs function by binding to proteins to form protein-RNA complexes.[5] The large numbers of proteins and RNAs in living cells constitute complex networks of RNA-protein interactions.[3] There is growing evidence that aberrations or dysregulation of the expression of RNA-binding proteins (RBPs) are associated with diseases, including neurodegeneration and cancer.[4,6–8] Hence, deciphering the RNA-protein recognition code is important for understanding the sequence and structural determinants of the specificity and affinity of RNA binding sites of RBPs, which is essential both for comprehending the functional implications of protein-RNA interactions and for developing new therapeutic strategies for many diseases. Despite the

Yong Jung and Yasser EL-Manzalawy should be considered co-first authors

progress in both experimental and computational strategies for understanding protein-RNA interactions, the nature of the protein-RNA recognition code is far from well understood.

While roughly half of all RBPs bind preferentially to a particular RNA sequence or structural motif, the rest appear to bind RNA in a "nonspecific" manner.[9] Sequence-specific RBPs typically recognize their RNA binding partners by forming complex binding surfaces that combine multiple modular RNA-binding motifs or domains (RBDs),[10] examples include Cas9,[11] Pumilio,[12] and zinc finger proteins.[13] Nonspecific RBPs include translation elongation and initiation factors, and proteins involved in RNA degradation.[14,15] Recent studies have called into question the widely used classification of RBPs as specific versus nonspecific[9] and argue for a more nuanced characterization of the RNA partner-specificity of RNA-protein interactions.

Determining the RNA binding sites, that is, the individual amino acids (AAs) and ribonucleotides (rNTs) that form protein-RNA interfaces, is a necessary step in understanding protein-RNA interactions. Experimental methods for identifying protein-RNA interfaces include methods for solving the structures of protein-RNA complexes, for example, x-ray crystallography,[16] nuclear magnetic resonance (NMR) spectroscopy,[17] and individual nucleotide resolution UV crosslinking and immunoprecipitation (iCLIP).[18] However, because biophysical methods for characterizing protein-RNA interfaces are particularly challenging,[19] there is an increasing reliance on computational methods, including in particular, statistical machine learning methods, for predicting RNA-binding sites in RBPs.[20–28] Most of the existing methods for predicting RNA-binding sites of RBPs are partner-agnostic in that they do not take into account the characteristics of the putative RNA binding partner.[22,29,30] Several recent studies have shown that the performance of protein-protein interface prediction can be improved by incorporating information from the interacting partner protein.[31–36] Motivated by the success of such partner-specific protein-protein interface prediction methods, several methods for DNA or RNA partner-specific prediction of protein-DNA/RNA interface residues in DNA-binding proteins (DBPs)/RBPs have been proposed recently.[37–39] However, it is unclear as to whether predicted DNA/RNA-binding sites are indeed DNA/RNA partner-specific.

Against this background, we focus on partner-specific protein-RNA interface prediction methods. We show that the two existing partner-specific methods, PS-PRIP[39] and PRIdictor,[38] with publicly accessible web servers have lower performance compared to some state-of-the-art partner-agnostic protein-RNA interface prediction tools. Moreover, we demonstrate, using an independent test set of 24 protein-RNA interacting chains, that the predictions returned by the "partner-specific" methods are unaffected by changes in the RNA sequence. We introduce a novel metric, RNA-specificity metric (RSM), for quantifying the RNA specificity of the protein-RNA interface predictions. Together with the standard machine learning performance evaluation metrics, RSM offers a useful metric for assessing and comparing RNA partner-specific protein-RNA interface prediction methods. This study underscores the importance of using appropriate test sets and evaluation metrics, and sets the bar for future works on partner-specific prediction of protein-RNA interfaces and by extension, other types of macromolecular interfaces, complexes and interactions.

2 | MATERIALS AND METHODS

2.1 | RNA partner-specific protein-RNA interface predictors

The vast majority of protein-RNA interface prediction methods are not RNA partner-specific, that is, they are partner-agnostic. Such methods predict all RNA interface residues in a query protein. Alternatively, RNA partner-specific interface prediction methods (eg, PRIdictor[38] and PS-PRIP[39]) take as input a protein and RNA pair and return predicted interfaces specific for this interaction (Figure 1).

To the best of our knowledge, there are two sequence-based and one structure-based RNA partner-specific protein-RNA interface predictors, PRIdictor,[38] PS-PRIP[39] and RPI-Bind,[40] respectively. Only PRIdictor and PS-PRIP are accessible as online web servers. Although RPI-Bind provides a set of their source codes, it is for the purpose of reproduction against their own fixed datasets and features. Hence, we excluded RPI-Bind in our analysis due to the difficulty of testing with our benchmark dataset.

PRIdictor uses physicochemical properties, rNT and AA composition of binding partners, positional information, and the interaction propensities of AA triplets to bind specific rNTs to train and evaluate an SVM classifier for predicting RNA-binding residues in the given RBP and the protein-binding residues in the given RNA. The dataset consisted of 542 protein-RNA complexes (formed by 376 proteins with 439 RNA binding partners) extracted from the Protein Data Bank (PDB) in 2013. The dataset was split into a training set (formed by 284 RNA and 246 protein chains) and a test set (formed by 155 RNA and 130 protein chains) such that similarity between any pair of RNA sequences from training and test sets is less than 80% (the redundancy cutoff for the protein similarity was unspecified). An rNT (or AA) involved in at least one of the interactions was classified as a protein-binding (or RNA-binding) site. PRIdictor consists of four support vector machine (SVM)[41] predictors: (a) RP model for predicting protein-binding sites in RNA using both protein and RNA sequences; (b) RaP, a predictor of protein-binding sites in RNA using RNA sequence alone; (c) PR, a predictor of RNA-binding sites in protein using protein and RNA sequences; (d) PaR, a predictor of RNA-binding sites in protein from protein sequence alone.

PS-PRIP[39] simultaneously predicts RNA binding residues in the protein sequence and protein binding rNTs in the RNA sequences given a pair of interacting protein-RNA sequences. PS-PRIP used a training set of 1637 interacting protein-RNA pairs, such that each protein sequence was at least 25 AAs in length and each RNA sequence was at least 100 rNTs in length, Protein-RNA complexes with resolution better than 3.5 Å were extracted from PDB in 2015.[39] PS-PRIP makes predictions based on a lookup table of 55 154 protein-RNA interacting motifs comprising 3275 unique 5-mer protein subsequence and 835 unique 5-mer RNA subsequences. A pair of protein and RNA 5-mer subsequences are considered to be a protein-RNA interacting motif if they appear in interacting protein-RNA chains and

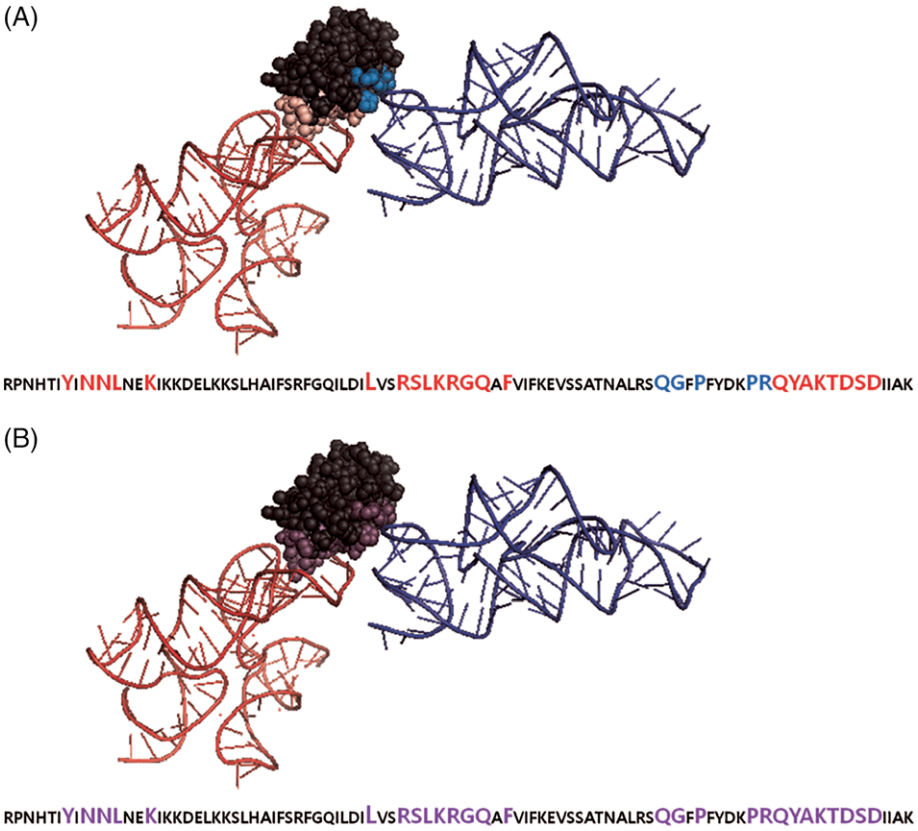

FIGURE 1 Difference between RNA partner-specific and partner-agnostic interface residue predictors. U1 small nuclear ribonucleoprotein A is an example of an RNA binding protein with multiple binding sites. Two protein-RNA complexes are used for this illustration (PDB ID: 4 W90, chains B and C, and 4YB1, chains P and R). A, An RNA partner-specific interface residue predictor takes as input a query protein and one or more putative RNA partners and return predicted binding site for each RNA separately (highlighted using different colors). B, A partner-agnostic interface residue predictor takes as input a query protein and returns all predicted RNA interface residues for that protein

have at least one physical contact (< 5 Å) between a heavy atom in any AA and a heavy atom in any rNT. Given a pair of protein and RNA sequences, PS-PRIP searches for all protein-RNA interacting motifs. If a motif is found, then the corresponding AA residues and rNTs in a query protein-RNA pair are labeled as interfaces.

2.2 | RNA-specificity metrics

Let *M* be an RNA partner-specific protein-RNA binding site predictor. In order to quantify the extent to which the RNA-binding residues in the protein *p* predicted by *M* are specific to a putative interacting RNA partner *r*, we proceed as follows.

Given a pair (*p*, *r*) of putative protein and RNA binding partners, generate $\{r^0, r^1, ..., r^k\}$ such that r^0 corresponds to *r*, the putative RNA partner in the complex (*p*, *r*) and $r^1, ..., r^k$ correspond to alternative putative RNA binding partners as follows:

1. If *M* is an RNA partner-specific protein-RNA interface predictor that does not utilize RNA structural features, $r^1, ..., r^k$ correspond to randomly generated RNA sequences of the same length and rNT composition as *r*.
2. If *M* is an RNA partner-specific protein-RNA interface predictor that utilizes RNA structural features, and the RNA structures $r^1, ..., r^k$ are obtained from a reference set of nonredundant RNA structures (specifically, R213, described below, in our experiments) such that they share the least sequence identity with target RNA *r*.

Let $I_0^p, I_1^p, ..., I_k^p$ denote the sets of predicted interface residues of p with putative RNA partners $r^0, r^1, ..., r^k$ (respectively). Here, each such set I_j^p where $0 \le j \le k$ is a subset of indices $\{1, 2, \cdots |p|\}$ where |p| denotes the length of the protein sequence *p*. RSM is defined as:

$$RSM(M,p,r) = \frac{1}{k}\sum_{i=1}^{k} g\left(I_0^p, I_i^p\right)$$

where

$$g\left(I_0^p, I_i^p\right) = \begin{cases} 1 - \dfrac{2 \times \left|I_0^p \cap I_i^p\right|}{\left|I_0^p\right| + \left|I_i^p\right|} \; if \; \left|I_0^p\right| + \left|I_i^p\right| \neq 0 \\ 0, otherwise \end{cases}$$

RSM assesses the partner-specificity of the RNA-binding residues predicted by an RNA partner-specific protein-RNA interface predictor *M* applied to a protein-RNA pair (*p*, *r*) by quantifying the changes in the predicted RNA-binding residues when the RNA is replaced by another. The RSM score is a real number *x* in the interval [0, 1] and is interpreted as the expected percentage of overlap between the predicted binding sites in *p* for its RNA binding partner and alternative potential binding partners is $(1 - x) \times 100\%$. For example, if an RNA partner-specific predictor has an estimated RSM of 0.02 for a query protein-RNA pair (*p*, *r*), then we expect the overlap between the

predicted interface of *p* with *r*, and different RNAs to be 98%. Loosely speaking, an RSM score of 0 indicates that predictor is not RNA partner-specific (ie, the predicted RNA-binding residues of the protein are unaltered by the changes to the RNA binding partner under consideration) whereas an RSM score of 1 indicates that the predicted RNA-binding residues are maximally altered by the changes to the RNA under consideration.

2.3 | Performance evaluation metrics

We evaluated the performance of different methods for predicting protein-RNA interface residues as well as protein-RNA interfacial pairs using four standard threshold-dependent metrics: sensitivity (*Sn*), specificity (*Sp*), accuracy (*ACC*), Matthew's correlation coefficient (*MCC*), as defined below[42,43]:

$$Sn = \frac{TP}{TP + FN}$$

$$Sp = \frac{TN}{TN + FP}$$

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

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

In the case of prediction for RNA-binding residues in proteins, TP and TN denote correctly predicted interface and non-interface protein residues (respectively), while FP and FN denotes misclassified non-interface and interface residues (respectively). In the case of protein-RNA interfacial pair prediction, TP and TN represent correctly predicted interacting and noninteracting AA-rNT pairs, whereas FP and FN represent misclassified noninteracting and interacting AA-rNT pairs (respectively). We estimate TP, TN, FP, and FN counts for *each protein* or *protein-RNA pair in the test dataset*, yielding the corresponding protein or complex based performance estimates. They provide estimated performance of the predictor on each protein or complex tested. The results can then be summarized in a variety of ways, including average performance over the entire test set. The resulting protein or complex based performance estimates provide much more useful measures of performance of binding site predictors than residue-based estimates, which are obtained from TP, TN, FP, and FN counts summed over the entire dataset of proteins (or protein-RNA complexes) in the test dataset.[44]

We also report two threshold-independent area under curve (AUC) metrics, area under receiver operating characteristic curve (AUC [ROC]) and area under concentrated receiver operating characteristic curve (AUC [CROC]). The receiver operating characteristic (ROC) curve is a two-dimensional plot in which the true positive rate is plotted against the false positive rate. The area under ROC curve, AUC [ROC], is a summary statistic of the ROC curve and it is interpreted as the probability that a randomly chosen positive sample will be ranked higher than a randomly chosen negative sample. Therefore, any AUC [ROC] value higher than 0.5 is considered better than random guessing. AUC [ROC] is not very useful when the data are highly imbalanced.[45,46] An alternative metric, called area under concentrated receiver operating characteristic curve AUC [CROC] has been proposed to assess early retrieval of positive samples.[47] In such settings, the important portion of the ROC curve is magnified by magnification factor, $\propto$, and the area under this magnified curve is reported. In our experiments, we set $\propto = 7$ which sends the point $x = 0.1$ in the ROC curve onto $x = 0.5$. Using $\propto = 7$, the AUC [CROC] for a random guessing classifier is $\frac{1}{\propto} - \frac{e^{-\propto}}{1-e^{-\propto}} \approx \frac{1}{\propto} = 0.14$.

2.4 | Datasets

We retrieved 1590 protein-RNA complexes and 150 protein-DNA-RNA complexes from the Protein Data Bank (PDB)[48] in September 2015. We then selected protein-RNA pairs using the following criteria: (a) complex resolution better than 3.5 Å; (b) protein sequence length ranges from 40 to 500 AA residues and RNA sequence length ranges from 25 to 300 rNTs; (c) the number of interface residues (AAs as well as rNTs) is ≥3. An interface residue is determined using a cut-off distance of 5 Å between any two pairs of atoms; (d) any two protein-RNA pairs have neither their protein chains sharing greater than 25% sequence identity nor their RNA chains sharing sequence identity greater than 40%. These selection criteria resulted in a nonredundant dataset of 172 protein-RNA interacting pairs. Then, we split the dataset into training and test sets based on their release date. Protein-RNA pairs extracted from complexes deposited into PDB before January 1, 2014 were used to train our classifiers (PR122) and the remaining 50 protein-RNA pairs (PR50) served as our independent test set. Two more test sets, PR24 and PR30, were used in our experiments. PR24 was derived from PR122 and PR50 by excluding protein-RNA pairs with RNA chains of length less than 100 rNTs. The reason is that the PS-PRIP server restricts submissions to protein-RNA pairs with at least 100 rNTs. PR30 was derived from PR50 by excluding protein-RNA pairs in which the protein sequence might share high sequence similarity with any protein sequence in PR50. A summary of the four datasets is provided in Table 1 and the PDB chain IDs for

TABLE 1 Protein-RNA datasets used in this study

Dataset	No. interfacial pairs	No. non-interfacial pairs	No. interfacial residues	No. non-interfacial residues
PR122	6429	1 786 901	3328	25 474
PR50	2662	608 602	1391	8664
PR24	1048	406 061	512	2702
PR30	1283	361 984	708	5580

PR122: A dataset for training, which consists of 122 protein-RNA complexes.
PR50: A dataset for independent testing, which consists of 50 protein-RNA complexes.
PR24: A dataset derived from PR122 and PR50 by excluding protein-RNA pairs where RNA length is less than 100 ribonucleotides.
PR30: A dataset derived from PR50 by excluding protein-RNA pairs where the protein sequence shares high sequence similarity (> 25%) with any protein sequence in PR50.

interacting protein-RNA pairs in the four datasets are also provided in Supporting Information Tables S1-S4. Supporting Information Table S5 provides the partitioning of PR122 interacting protein-RNA pairs into five subsets used for fivefold cross-validation experiments.

In addition to the datasets derived from protein-RNA complexes in PDB, we assembled a dataset of RNAs to be evaluated as putative binding partners for proteins to obtain the RSM of RNA binding residue predictors that make use of the structural features of the RNA. Such a dataset is required because unlike in the case of RNA sequences, we cannot artificially generate RNA structures to evaluate as putative alternative binding partners of a protein that is part of protein-RNA complex in PDB by randomly shuffling an RNA structure extracted from the bound complex. First, we identified 2495 PDB structures containing at least one RNA chain, regardless whether it is bound to a protein. Second, we collected the RNA structures (including those extracted from protein-RNA complexes) from this set. Third, we filtered the resulting set of RNA structures based on criteria identical to those used to derive the dataset of 172 protein-RNA pairs described above: (a) the resolution of the corresponding PDB structure is at least 3.5 Å; (b) RNA sequence length ranges from 25 to 300 rNTs; (c) No RNA structure contain RNA chains that share a sequence identity greater than 80% with an RNA chain that is part of another RNA structure included in the dataset. This procedure yielded a nonredundant dataset of 213 RNAs (R213). It is should be emphasized that R213 includes RNA structures extracted from protein-RNA complexes as well as those that are not bound to any protein, permitting us to obtain relatively unbiased RSM estimates (modulo the current coverage of PDB).

2.5 | Feature extraction

Let (p, r) denote a pair of interacting protein (p) and RNA (r) chains where $p = p_1, p_2, \cdots, p_n$ represents the n residues of the protein sequence and $r = r_1, r_2, \cdots, r_m$ represents the m rNTs of the RNA sequence. We extracted all protein-RNA residue pairs (p_i, r_j) where $i \in \{1, 2, \cdots, n\}$ and $j \in \{1, 2, \cdots, m\}$ and labeled them as either interfacial or non-interfacial AA-rNT pairs. We then extracted sequence-based features from each such residue pair (p_i, r_j) from sequence windows centered at p_i and r_j over the corresponding protein and RNA chains, respectively. Using windows of size 15 and 21 for protein and RNA chains (respectively), we extracted five types of features: (a) Protein sequence features, (b) Protein structure features, (c) RNA sequence features, (d) RNA structure feature, and (e) Protein-RNA interface motif feature, as described below. We also extracted several structural features of the corresponding protein residues and RNA nucleotides, also described below.

2.5.1 | Protein sequence features

We used a position-specific scoring matrix (PSSM) to encode each AA residue within a window. Specifically, we ran PSI-BLAST[49] against NCBI nr database to retrieve the sequence homologs of each protein sequence using three iterations of PSI-BLAST with an e-value of 0.001. We used the resulting hits to build the PSSM profile of the protein sequence. Following the procedure used in previous studies,[23,50,51] we normalized the PSSM values to lie in the range [0,1] using the logistic (also called sigmoid) function. Finally, each residue in a sequence window was encoded using a 20-element vector that corresponds to its normalized PSSM profile. Thus, each AA sequence window was encoded using 15×20 numeric features.

2.5.2 | Protein structure features

Two structural features used for each protein residue were protein secondary structure codes and the relative accessible surface area (rASA). Using STRIDE,[52] each residue in protein structure was assigned to one of seven structure category codes; alpha helix, 3-10 helix, PI-helix, extended conformation, isolated bridge, turn, and coil. We encoded the structure codes with numbers from one to seven as a feature. The rASA of each residue was calculated using both its solvent accessible area obtained using STRIDE and the known surface area of the residue.[53]

2.5.3 | RNA sequence features

We extracted RNA tri-nucleotide composition features[54] from RNA sequence windows consisting of 21 rNTs. Occurrences of all 64 possible trinucleotides (4 ×4× 4 rNTs, from AAA to UUU) within an RNA sequence window were counted and used as 64 sequence features. Also, we added total length of an RNA sequence and the frequency of the four rNT types in the RNA sequence as features.

2.5.4 | RNA structure feature

We assigned RNA secondary structure codes to rNT using DSSR[55] and BEAR encoding[56] following the protocol used in the RPI-Bind method.[40] A dot-bracket notation sequence for entire RNA sequence was obtained using DSSR. Based on the resulting dot-bracket notation, we assigned a BEAR RNA structure code to each rNT. The 12 different RNA structure BEAR codes we used were: Unpaired, Loop, Stem, Stem branch, Left internal loop, Left internal loop branch, Right internal loop, Right internal loop branch, Bulge left, Bulge right, Bulge left branch, and Bulge right branch. We encoded the RNA structure BEAR codes with numbers from 1 to 12 as a feature.

2.5.5 | Protein-RNA interface motif feature

Following Muppirala et al.[39], we extracted 494 unique protein-RNA interface motifs from PR122 dataset. Each motif is a pair of interacting AA and rNT 5-mers that includes at least one interfacial protein-RNA pair. The presence or absence of each of these 494 motifs is encoded by a corresponding binary feature with a '1' denoting the presence of the corresponding protein-RNA interface motif, and a '0' denoting its absence within a pair of protein-RNA windows.

Based on the generated feature vectors, we developed two RNA partner-specific protein-RNA interfacial pairs predictors trained on PR122, a sequence-based predictor, PSPRInt-Seq and a structure-based predictor, PSPRInt-Str. For PSPRInt-Seq, we utilized three of the five feature types: protein sequence features, RNA sequence features, and protein-RNA interface motif features. For PSPRInt-Str, we utilized all five feature types.

3 | RESULTS

3.1 | How reliable are partner-specific protein-RNA binding interface prediction methods?

We experimented with a set of widely used features for representing protein and RNA sequences and structures as well as features for representing the interactions between an AA residue and an RNA nucleotide (eg, protein-RNA interaction motifs[39]). Most of the sequence features have been used in the two related studies, PS-PRIP[39] and PRIdictor.[38] Tables 2 and 3 report the performance of three machine learning classifiers, random forest (RF)[57] with 1000 trees, support vector machine (SVM)[41] with the radial basis function (RBF) kernel (SVM-RBF) and naïve bayes (NB)[58] on predicting protein-RNA interfacial pairs using fivefold cross-validation and independent test procedures. It is important to note that ignoring the fact that the PR122 and PR50 datasets have extremely uneven class distribution (ie, millions or hundreds of thousands of non-interfacial pairs and only few thousands of interfacial pairs) might lead to inaccurate conclusions. In such a setting, naïve use of the standard metrics such as *ACC*, *Sn*, *Sp*, AUC can lead to highly misleading conclusions. For example, because more than 99% of the instances correspond to are non-interfacial pairs, classifiers that predict every protein-RNA residue pair as a non-interfacial residue pair can achieve an accuracy exceeding 99%; A classifier that achieves a sensitivity of 0.5 and specificity of 0.9 has a false positive rate (1-specificity) of 0.1 which can be unacceptably high when the dataset contains millions of non-interfacial residues. The statistics summarized in Table 1 show that in order for the RF classifier to correctly identify ~2700 interfacial pairs, one must be willing to tolerate 178 690 false positive predictions; Even AUC has limitations when the data are unbalanced.[45,46] As noted in the Methods section, AUC [CROC] offers a more useful alternative to AUC in this setting.[47] Hence, the only conclusions that we can draw from the results in Tables 2 and 3 are that: (a) All three classifiers outperform random guessing (recall that a random guessing classifier would have MCC, AUC [ROC], and AUC [CROC] equal to 0, 0.5, and 0.14 respectively); (b) The RF classifier outperforms NB and SVM-RBF in terms of MCC, and AUC [CROC] when performance is evaluated using cross-validation; (c) Evaluations using the PR50 independent test set show that RF and SVM-RBF achieve comparable performance and both outperform NB.

The preceding results demonstrate that the performance of predictors of protein-RNA interfacial residue pairs trained using standard machine learning algorithms and the sequence-derived features used in previous studies[38,39] leaves much room for improvement. They also underscore the some of the challenges that need to be addressed, not the least of which has to do with the extremely uneven class distribution in datasets commonly used. Based on these results, we decided to use the RF classifier, PSPRInt-Seq, and PSPRInt-Str, for the rest of our experiments. Despite its poor performance (based on MCC) in predicting protein-RNA interfacial residue pairs, it was of interest to determine whether this classifier can accurately predict the binding residues in the RBP.

3.2 | From interfacial AA residue-rNT pairs to RNA binding residues in RBPs

Although it is claimed that PS-PRIP and PRIdictor models predict interfacial protein-RNA residue pairs, the implemented servers return predicted interfaces in the RNA and protein sides separately. Both studies also assessed the performances of the resulting RNA partner-specific predictors of RNA-binding residues in proteins. In the following experiment, we evaluate three methods for converting predicted interfacial protein-RNA residue pairs into predicted RNA-binding residues in RBPs. Briefly, for each residue in the RNA-binding protein, we aggregate the predicted scores for that particular residue paired with all rNTs in the RNA sequence. This aggregated score is then treated as the predicted score for that AA residue to be an RNA-binding residue in the protein-RNA complex. We evaluated three different methods for aggregating the AA residue-rNT scores into a single "per residue" score: (a) maximum of pair-wise scores; (b) average of all pair-wise scores; and (c) average of top k pair-wise scores for k = 5, 15, and 25. Tables 4 and 5 show that it is reasonable to use any of these aggregation methods to map residue-rNT interaction scores onto RNA-binding residue scores, because no method outperforms the others using MCC, AUC [ROC], and AUC [CROC], with differences in any metric (if any) at most 2%. Hence, our final model for RNA partner-specific predictor of RNA-binding residues in proteins is an RF trained using the PR122 dataset, and using the maximum of all AA residue-rNT scores for a specific AA residue as the predicted score that the residue is an RNA-binding residue within the corresponding protein-RNA complex. These conversions were applied to both PSPRInt-Seq and PSPRInt-Str.

TABLE 2 Performance of different classifiers for predicting interfacial amino acid residue-ribonucleotide pairs using fivefold cross-validation and PR122 dataset

Features	Classifier	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
Sequence-based	RF	0.43	0.91	0.91	0.07	0.77	0.45
	SVM-RBF	0.16	0.98	0.98	0.06	0.74	0.39
	NB	0.50	0.75	0.75	0.04	0.68	0.34
Structure-based	RF	0.47	0.91	0.91	0.08	0.80	0.48
	SVM-RBF	0.16	0.98	0.98	0.06	0.75	0.40
	NB	0.51	0.76	0.75	0.04	0.69	0.35

Abbreviations: ACC, accuracy; AUC, Area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; NB, naïve bayes; RF, random forest; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity; SVM-RBF, support vector machine with radial basis function kernel.

TABLE 3 Performance of different classifiers for predicting interfacial amino acid residue-ribonucleotide pairs using the PR50 independent test set

Features	Classifier	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
Sequence-based	RF	0.48	0.91	0.91	0.09	0.80	0.49
	SVM-RBF	0.35	0.96	0.96	0.10	0.79	0.48
	NB	0.47	0.77	0.77	0.04	0.69	0.30
Structure-based	RF	0.52	0.90	0.90	0.09	0.83	0.52
	SVM-RBF	0.36	0.96	0.96	0.11	0.81	0.50
	NB	0.49	0.76	0.76	0.04	0.69	0.30

Abbreviations: ACC, accuracy; AUC, Area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; NB, naïve bayes; RF, random forest; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity; SVM-RBF, support vector machine with radial basis function kernel.

TABLE 4 Performance comparison of different methods for mapping predicted interfacial amino acid residue-ribonucleotide pairs to RNA-binding residues on protein side using PR50 independent test set

Model	Method	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
PSPRInt-Seq	Max	0.47	0.93	0.87	0.42	0.81	0.42
	Average	0.38	0.96	0.88	0.41	0.80	0.41
	Average top-5	0.46	0.93	0.87	0.42	0.81	0.42
	Average top-15	0.44	0.94	0.87	0.41	0.81	0.42
	Average top-25	0.42	0.95	0.87	0.42	0.81	0.42
PSPRInt-Str	Max	0.55	0.90	0.85	0.42	0.84	0.44
	Average	0.41	0.95	0.88	0.42	0.83	0.43
	Average top-5	0.54	0.91	0.86	0.43	0.84	0.44
	Average top-15	0.50	0.93	0.87	0.43	0.84	0.44
	Average top-25	0.48	0.93	0.87	0.43	0.84	0.43

Abbreviations: ACC, accuracy; AUC, area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity.

It should be emphasized that our primary intention here is not necessarily to develop a novel method for RNA partner-specific prediction of RNA-binding residues in proteins or AA-rNT pairs in the interface of a protein-RNA complex. Instead, our goal is to implement baseline sequence-based and structure-based methods to gain insights into the relative performance of the RNA partner-specific versus partner-agnostic predictors of protein-RNA interfaces. Specifically, we systematically evaluated the predictions made by PSPRInt-Seq and PSPRInt-Str, and the importance of features[57] of the learned RF models implemented in these classifiers to explore the conditions under which RNA partner-specific and RNA partner-agnostic methods for protein-RNA interface prediction can provide inaccurate or potentially misleading conclusions about their performance relative to each other.

3.3 | Comparisons with other RNA partner-agnostic predictors of RNA-binding residues in proteins

Table 6 compares the performances of PSPRInt-Seq and PSPRInt-Str with five RNA partner-agnostic predictors of RNA-binding residues in

TABLE 5 Performance comparison of different methods for mapping predicted interfacial amino acid residue-ribonucleotide pairs to RNA-binding residues on protein side using PR30 independent test set

Model	Method	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
PSPRInt-Seq	Max	0.25	0.93	0.85	0.21	0.74	0.33
	Average	0.18	0.95	0.86	0.19	0.73	0.34
	Average top-5	0.23	0.93	0.85	0.20	0.74	0.33
	Average top-15	0.21	0.94	0.85	0.19	0.74	0.33
	Average top-25	0.21	0.95	0.85	0.19	0.74	0.33
PSPRInt-Str	Max	0.38	0.91	0.85	0.27	0.78	0.35
	Average	0.24	0.96	0.88	0.26	0.78	0.37
	Average top-5	0.36	0.92	0.86	0.28	0.78	0.36
	Average top-15	0.32	0.93	0.87	0.28	0.78	0.36
	Average top-25	0.30	0.94	0.87	0.26	0.78	0.36

Abbreviations: ACC, accuracy; AUC, area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity.

TABLE 6 Performance comparisons of PSPRInt-Seq and PSPRInt-Str with RNA partner-agnostic RNA-binding residue prediction methods using RBPs in PR50 test set

Method	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
PSPRInt-Seq	**0.47**	**0.93**	**0.87**	**0.42**	**0.81**	**0.42**
PSPRInt-Str	**0.55**	**0.90**	**0.85**	**0.42**	**0.84**	**0.44**
RNABindRPlus	0.47	0.94	0.88	0.44	0.83	0.44
FastRNABindR	0.70	0.75	0.74	0.33	0.79	0.39
RNABindR v2.0	0.68	0.71	0.71	0.28	0.77	0.35
BindN+	0.47	0.85	0.80	0.27	0.75	0.33
RBScore	0.60	0.81	0.78	0.33	0.76	0.33

Results for two RNA partner-specific methods are shown in bold, above the line. All other methods (below the line) are RNA partner-agnostic.
Abbreviations: ACC, accuracy; AUC, area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity.

proteins[20,21,25,26,51] using the PR50 test set. Interestingly, both PSPRInt-Seq and PSPRInt-Str outperform most of the other methods based on all performance metrics examined except Sensitivity. Moreover, a substantial improvement in MCC is observed. However, we note that this comparison is not entirely fair. Although PR50 is nonredundant with respect to the set of interacting protein-RNA pairs in PR122, our training dataset, PR50, and PR122 need not necessarily be nonredundant with respect to proteins. Recall that two protein-RNA pairs are considered redundant if the similarity between the two protein sequences is greater than 25% and the similarity between the RNA sequences is greater than 40%. Thus, two protein-RNA pairs that have the same protein sequence but different RNAs are considered to be nonredundant. Consequently, a protein belonging to a protein-RNA pair included in PR50 can have a high degree of sequence similarity with one or more proteins belonging to protein-RNA pairs included in PR122, as long as the corresponding RNAs have sufficiently dissimilar sequences. If we eliminate redundancy with respect to both protein and RNA, we end up with the PR30 dataset, a subset of PR50 which only includes only those protein-RNA pairs containing proteins with sequence identity no greater than 25% with respect to any protein belonging to a protein-RNA pair included in PR122. Table 7 compares the performance of PSPRInt-Seq and PSPRInt-Str with five web servers that implement partner-agnostic predictors of RNA-binding residues in proteins using the PR30 dataset. The results of this comparison, unlike those obtained using the PR50 dataset, show that PSPRInt-Seq and PSPRInt-Str no longer substantially outperform the other prediction methods.

In summary, the results of evaluating the performance of RNA partner-specific predictors of RNA-binding residues in proteins using a nonredundant test dataset (PR30, Table 7) contradicts the conclusions from an assessment using a test dataset (PR50, Table 6) that, as in previously published studies of RNA partner-specific predictors of RNA binding residues in proteins,[38,39] does not ensure that the proteins included in the protein-RNA pairs in the test dataset are nonredundant with respect to those in the training dataset. Hence, we conjecture that the purported superior performance of RNA partner-specific methods, which has been attributed to their use of RNA-derived features, AA-rNT pairs,[38] or protein-RNA interfacial motifs[39] can be simply explained instead by the redundancy of the proteins belonging to protein-RNA pairs in the test dataset with respect to those in the training dataset. We conclude that the performance of the existing RNA partner-specific predictors of RNA binding residues in proteins is *at best* comparable to that of their partner-agnostic counterparts.

3.4 | Do features of the RNA binding partner improve PSPRInt predictions?

We proceeded to test whether our RNA partner-specific predictors, PSPRInt-Seq, and PSPRInt-Str, derive significant performance gains from the features of the putative RNA partner, relative to their counterparts that do not take into account features of the RNA, that is, those that use only features derived from the RBP. Specifically, we conjecture that the added RNA features do not provide enough useful information to boost the performance of methods such as PSPRInt-

TABLE 7 Performance comparisons of PSPRInt-Seq and PSPRInt-Str with RNA partner-agnostic RNA-binding residue prediction methods using RBPs in PR30 test set

Methods	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
PSPRInt-Seq	**0.25**	**0.93**	**0.85**	**0.21**	**0.74**	**0.33**
PSPRInt-Str	**0.38**	**0.91**	**0.85**	**0.27**	**0.78**	**0.35**
RNABindRPlus	0.34	0.94	0.88	0.32	0.79	0.40
FastRNABindR	0.58	0.73	0.72	0.25	0.73	0.36
RNABindR v2.0	0.57	0.70	0.70	0.22	0.72	0.35
BindN+	0.40	0.83	0.78	0.21	0.71	0.30
RBScore	0.50	0.79	0.76	0.27	0.73	0.31

Results for two RNA partner-specific methods are shown in bold, above the line. All other methods (below the line) are RNA partner-agnostic.
Abbreviations: ACC, accuracy; AUC, area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity.

Seq and PSPRInt-Str, relative to their partner-agnostic counterparts. To test this conjecture, we examined the feature importance scores[57] extracted from the respective RF models. Supporting Information - Tables S6 and S7 list all the features used to encode data for training PSPRInt-Seq and PSPRInt-Str, respectively, ordered by their respective feature importance scores. In the case of PSPRInt-Seq, the top ranked feature happens to be the AA-rNT interaction motif feature, followed by the PSSM value of the target AA residue (located at the center of the 15 AA window) and those of its flanking sequence neighbors. Surprisingly, RNA trinucleotide features, derived from the putative RNA partner, are ranked at the bottom of the list. Thus, the RNA-derived features do not appear to contribute much to the predictions made by the RF classifier implemented in PSPRInt-Seq. In the case of PSPRInt-Str, interestingly, the rASA of the target residue is ranked at the top of the list, with the secondary structure of the RNA binding partner also appearing among the top ranked features (among the top 3). This suggests that the structural features of both the protein and its putative RNA binding partner tend to be informative in discriminating protein-RNA interfacial residue pairs from non-interfacial residue pairs. However, surprisingly, incorporating these structural features did not lead to significant performance gains over RNA partner-agnostic methods. We note further that the AA-rNT interaction motif feature and the PSSM values of the AA residues are still ranked highly, as in the case of PSPRInt-Seq.

3.5 | How sensitive are the RNA partner-specific predictions of RNA-binding residues in proteins to changes in the RNA partner?

The preceding approach to assessing the contribution of the RNA features used by RNA partner-specific protein-RNA interface predictors is applicable only when we have access to the internal structure of the respective predictors (ie, their code, eg, the classification models with the list of features used and their feature importance scores). Because PS-PRIP[39] and PRIdictor[38] are accessible only as online web servers, it is not possible to use this approach. Hence, we devised a novel approach to assess the role-played by the RNA-derived features indirectly by quantifying how the predicted interface changes depending on the putative RNA binding partner. For each protein-RNA interacting chain in the PR24 test set, we queried each of the four RNA partner-specific predictors 11 times, once using the actual RNA partner of the protein and 10 times with its putative RNA binding partner replaced with its sequence variant (in the case of sequence-based predictors) or structure variant (in the case of structure based predictors) in order to compute an RNA-Specificity Metric, RSM, for each predictor (see the RNA-Specificity Metrics subsection in the Materials and Method Section for details concerning the choice of the sequence and structure variants of the putative RNA binding partner). The key intuition here is that if a predictor is indeed RNA partner-specific, then the predicted RNA binding residues in the query protein or the predicted protein-RNA interfacial residue pairs should vary significantly when we replace the actual RNA binding partner with its sequence or structure variants. More "RNA partner-specific" predictions should have higher RSM values (closer to 1), and predictions that are not affected by changing the RNA partner should have lower RSM values (closer to 0).

Table 8 shows the RSM scores for predictors computed for each protein-RNA pair in PR24 dataset. PS-PRIP has the highest average RSM score of 0.138. However, its RSM score is zero for 13 out of the 24 test cases. PRIdictor has the average RSM score of 0.056, with RSM score of zero for seven cases. The average RSM scores for PSPRInt-Seq and PSPRInt-Str are 0.050 and 0.069, respectively. Recall that an RSM score of zero means that in all pair-wise comparisons between the predicted interface with the actual RNA binding partner and predicted interfaces associated with its RNA sequence or structure variants, the model returns exactly the same predictions (ie, the prediction is independent of the putative RNA binding partner in question). The detailed RSM assessing results of PS-PRIP, PRIdictor, PSPRInt-Seq, and PSPRInt-Str are provided in Supporting Information Text.

Table 9 compares the performance of RNA partner-specific and partner-agnostic RNA-binding residue predictors on the PR24 test set. For a fair comparison, we considered only protein-side predictions of the RNA partner-specific methods. In the cases of PS-PRIP and PRIdictor, we used results of their protein-side predictions provided by the web servers. In the cases of PSPRInt-Seq and PSPRInt-Str, we used the aggregated prediction scores for single AA residues from the corresponding residue-rNT pairs (see the *From interfacial AA residue-rNT pairs to RNA binding residues in RBPs* subsection in the Results for details). At the time of the experiments reported in this section, BindN+[26] web server was no longer accessible, so could not be included in the comparison. Even though ignoring BindN+, PSPRInt-Seq and PSPRInt-Str (because of the overlap between PR24 and PR122) from the comparison, we observed that the RNA partner-agnostic predictors substantially outperform their RNA partner-specific counterparts PS-PRIP and PRIdictor, in terms of MCC.

In summary, results in Tables 8 and 9 show that: (a) the predictions of RNA-binding residues in RBPs produced by PS-PRIP and PRIdictor have a low degree of RNA partner-specificity; and (b) the performance of "RNA partner-specific" predictors in identifying RNA-binding residues in RBPs is not superior to that of RNA partner-agnostic predictors.

3.6 | Effect of increasing the size of the test set on estimated RSM scores

A major limitation in Table 8 results, is that the estimated RSM scores for the four RNA-specific protein-RNA interface prediction methods is obtained using only 24 protein-RNA pairs. To address this limitation, we generated a new test set covering all protein-RNA complexes deposited in PDB between January 2014 and July 2018. Following the same data preprocessing procedure described in the Methods Section, the final dataset (called PR38) consists of 38 nonredundant protein-RNA pairs and the length of each RNA chain is at least 100 nucleotides. Supporting Information Table S8 reports the RSM scores for protein-RNA pairs on PR38 dataset estimated for four RNA-specific protein-RNA interface prediction methods. Interestingly, the results are in agreement with the results in Table 8 obtained using the smaller test set. We conclude that the poor RSM scores of all the methods considered in this study is not an artifact of using a small test set.

TABLE 8 RSM scores for protein-RNA pairs in the PR24 dataset determined for four RNA partner-specific RNA-binding residue prediction methods

Interacting protein-RNA chain	PS-PRIP	PRIdictor	PSPRInt-Seq	PSPRInt-Str
1MFQ_B-1MFQ_A	**0.000**	0.063	0.053	0.063
1MFQ_C-1MFQ_A	**0.000**	**0.000**	0.057	0.037
1NWY_M-1NWY_9	0.288	0.400	0.029	**0.000**
1U6B_A-1U6B_B	**0.000**	0.007	0.058	0.055
1VQN_Q-1VQN_9	0.193	0.004	0.114	0.135
1W2B_V-1W2B_9	0.251	0.024	0.073	0.063
2OTJ_D-2OTJ_9	0.070	0.034	0.038	0.029
2ZJR_D-2ZJR_Y	0.206	**0.000**	0.032	0.058
3DLL_J-3DLL_Z	0.076	**0.000**	0.038	0.070
3DLL_S-3DLL_Z	**0.000**	0.114	0.014	0.053
3G71_H-3G71_9	0.129	0.054	0.054	0.053
3G8T_A-3G8T_P	**0.000**	0.005	0.091	0.116
3HHN_B-3HHN_C	**0.000**	0.032	0.072	0.056
3I56_N-3I56_9	0.033	0.014	0.057	0.080
3IVK_H-3IVK_M	**0.000**	**0.000**	0.047	0.130
3NDB_B-3NDB_M	0.367	0.167	0.070	0.085
3V7E_A-3V7E_C	**0.000**	**0.000**	0.029	0.066
4IO9_W-4IO9_Y	0.697	0.132	0.077	0.062
4LCK_A-4LCK_C	**0.000**	**0.000**	0.003	0.032
4P3E_B-4P3E_A	**0.000**	0.108	0.038	0.033
4P3E_C-4P3E_A	**0.000**	**0.000**	0.047	0.089
4UYJ_D-4UYJ_S	1.000	0.074	0.024	0.125
4UYK_A-4UYK_R	**0.000**	0.083	0.055	0.093
4W90_B-4W90_C	**0.000**	0.017	0.033	0.082
Average	0.138	0.056	0.050	0.069
STDEV	0.248	0.088	0.025	0.034

Abbreviation: STDEV, standard deviation.

4 | DISCUSSION

Identifying the individual AAs and rNTs that form interfaces in protein-RNA complexes is a crucial step in understanding the mechanisms of recognition in protein-RNA interactions. Computational approaches, including statistical machine learning methods, for predicting RNA-binding sites in RBPs[20–28] are increasingly valuable because biophysical characterization of protein-RNA complexes is

TABLE 9 Performance comparisons of RNA partner-specific RNA-binding residue predictors with RNA partner-agnostic RNA-binding residue predictors using RBPs in PR24 test set

Methods	Sn	Sp	ACC	MCC	AUC [ROC]	AUC [CROC]
PSPRInt-Seq	**0.90**	**0.91**	**0.90**	**0.69**	**0.97**	**0.57**
PSPRInt-Str	**0.90**	**0.89**	**0.88**	**0.66**	**0.97**	**0.57**
PRIdictor	**0.17**	**0.97**	**0.84**	**0.18**	**NA**	**NA**
PS-PRIP	**0.21**	**0.92**	**0.78**	**0.10**	**NA**	**NA**
RNABindRPlus	0.79	0.81	0.80	0.54	0.88	0.36
FastRNABindR	0.85	0.57	0.62	0.35	0.80	0.33
RNABindR v2.0	0.83	0.54	0.58	0.30	0.74	0.29
BindN+	NA	NA	NA	NA	NA	NA
RBScore	0.7	0.75	0.74	0.40	0.79	0.37

Results for two RNA partner-specific methods are shown in bold, above the line. All other methods (below the line) are RNA partner-agnostic.
Abbreviations: ACC, accuracy; AUC, area under curve; CROC, concentrated receiver operating characteristic; MCC, Matthew's correlation coefficient; ROC, receiver operating characteristic; Sn, sensitivity; Sp, specificity.
NA AUC results because corresponding methods return only predicted binary labels. NA results for BindN+ because the server was no longer accessible at the time of running this experiment.
The performance metric for PSPRInt-Seq and PSPRInt-Str assessed using the PR24 dataset should be interpreted with caution because PR24 is derived in part from PR122 (dataset used to train PSPRInt-Seq and PSPRInt-Str) and PR50. Therefore, PR24 is not an independent test set for the PSPRInts and the superior performance of the PSPRInts here may be due to the 19 protein-RNA pairs that they have in common.

difficult.[19] Most existing methods for predicting RNA-binding residues in RBPs use only the sequence and/or structural features of the RBP, without considering the sequence or structure of its putative RNA partner(s) even though RBPs such as Cas9 proteins,[11] Pumilio[12] and zinc finger proteins[13] are known to contain modular RNA-binding motifs or domains that recognize specific RNA sequence and/or structural features of their binding partners[10]; Base-specific hydrogen bonding,[59] electrostatic interactions, and geometric factors influence binding specificity.[60] Reliable computational tools for predicting RNA binding residues in RBPs can both complement and help narrow the focus of biophysical and molecular genetic methods for identifying features critical for recognition in protein-RNA interactions.

In this study, we have focused on computational methods for RNA partner-specific prediction of RNA-binding residues or protein-RNA interfacial pairs in RBPs. We rigorously assessed the extent to which existing methods for RNA partner-specific prediction of RNA-binding residues in RBPs[37–39] are indeed RNA partner-specific. We devised a novel metric, the RSM, for quantifying the RNA partner-specificity of the RNA-binding residues predicted by such tools. Our results show that the RNA-binding residues predicted by these methods are, in fact, almost oblivious to the characteristics of the putative RNA binding partner. Moreover, when evaluated using RNA partner-agnostic metrics, RNA partner-specific methods are outperformed by the state-of-the-art partner-agnostic methods. Our results underscore the importance of rigorously assessing the presumed RNA partner-specificity of protein-RNA interface predictors. The importance of ensuring that the training and test data are nonredundant cannot be overstated. Our results also highlight the importance of "looking under the hood" of black-box predictive models trained using machine learning to carefully examine the presumed contribution of RNA-derived features (as well as other features) in the resulting predictions.[61] To facilitate the interpretation of black-box predictions, it is essential to employ machine learning algorithms that produce interpretable models and provide information regarding which features contribute most to the accurate predictions.

Partner-specific protein-RNA interface prediction methods are designed to predict pairs of contacting AAs and rNTs, that is, residues on both sides of the interface. In this study, we focused only on predicting residues on the protein side for two reasons: (a) the reported performance of PRIdictor and PS-PRIP on predicting rNTs on the RNA side of the interface is very poor (MCC of 0.19 and 0.13 for PRIdictor and PS-PRIP, respectively); (b) the lack of existing tools that effectively use RNA sequence or structural information for predicting interfaces. We note, however, that the RNA-specificity metric introduced here could be easily adapted to assess the protein partner-specificity of the predicted protein-binding rNTs.

It is worth noting that most of the available structures of protein-RNA complexes in the PDB have been solved for proteins with high affinity for their bound RNA partner(s).[62] Furthermore, ribosomal RNA structures are greatly overrepresented among the protein-RNA complexes in PDB, making up almost 50% of the total. At present, it is unclear which among the RBPs represented in protein-RNA complexes in the PDB,[61] or even what fraction of them, are RNA partner-specific.[63]

Recent advances in high throughput techniques for characterizing protein-RNA complexes have provided unprecedented opportunities to identify partners and the interfaces in protein-RNA complexes in vivo.[64,65] In addition, it is now possible to estimate binding affinities for a large number of possible sequence variants for most RBPs in vitro, and to derive binding models.[9,66] Affinity distributions obtained from such experiments can provide an unbiased picture of protein binding to unstructured RNA or to specific RNA structures. Although it is believed that roughly half of all RBPs, and hence half of the corresponding protein-RNA interactions, are highly sequence-specific,[9,11] affinity distributions of RBPs measured in vitro and RNA binding patterns of numerous RBPs measured in vivo have called into question the classification of RBPs as "specific" versus "nonspecific".[9,66] For example, some proteins, for example, the C5 subunit of RNase P, can display both specific and nonspecific binding modes.[9] Studies that map RNA-protein interactions on a transcriptome-wide scale have shown that certain RBPs often bind to RNA sites that vary considerably in sequence and/or structure.[67] Additional sources of complication include potential cooperative interactions among RNA binding motifs and domains, the kinetics of the reactions that precede and follow binding, the presence of non-Watson-Crick base pairing in "loop" regions and the other tertiary interactions in RNA,[68–70] the diversity, complexity, and flexibility of the secondary structures formed by RNA,[71–73] the fact most RNAs exist as an ensemble of tertiary structures in solution,[74,75] and results showing that the RNA-binding sites of many RBPs correspond to intrinsic disordered regions (IDRs) in the unbound proteins.[76] Taken together, these findings argue for a more nuanced characterization of the "specificity" of RBPs and protein-RNA interactions.

In light of the preceding discussion, it is probably safe to conclude that the defining characteristics of RNA partner-specificity in protein-RNA interactions are largely not understood. Where does this leave current attempt to develop tools for predicting RNA-binding sites in proteins or protein-binding sites in RNA, which are bound by specific proteins and vice versa?

One possible direction is to leverage the known RNA-binding domains and motifs of RNA-binding proteins to help improve the performance of RNA partner-specific predictors of RNA-binding residues in RBPs.[77–80] Intrinsically disordered regions (IDRs) that connect multiple domains in RBPs may play a critical role in mediating the RNA sequence specificity of the interaction.[81] Recently, IDR have been successfully used for predicting RNA and DNA-binding proteins on a large scale.[82] This finding, together with the availability of increasing number of computational tools for predicting IDRs from protein sequence,[83–85] suggests a promising direction for developing improved protein-RNA binding site prediction models that incorporate information derived from IDRs. Another promising direction is to leverage high throughput data from protein-RNA binding assays and affinity distributions, to generate the RNA binding models for RBPs.[6,86,87] It should also be possible to incorporate a more nuanced notion of RNA partner-specificity into the training of RNA-binding residues in RBPs. A third promising direction is to leverage RNA binding models for RBPs with structures of protein-RNA complexes to improve the reliability of computational predictions of protein-RNA interfaces, interactions, and complexes (eg, using protein-RNA

docking[88–92] or template-based modeling[93,94]). It is also of interest to consider complexes formed by proteins with one or more other RNA, DNA, or protein partners because several RBPs bind not only to RNA but also to other macromolecules. Finally, because many proteins can bind to different classes of RNAs (eg, miRNA and mRNA)[95] and RBPs can be predicted more accurately for some RNA types (eg, rRNA[96]) than others, exploring the relationship between RNA functional types and RNA partner-specificity in protein-RNA interactions is another interesting research direction.

5 | CONCLUSIONS

In this study, we analyzed several aspects in developing RNA partner-specific protein-RNA interface prediction tools. We conducted experiments to: (a) demonstrate how challenging the problem is; (b) analyze the effect of different aggregation methods in mapping RNA partner-specific predictions into partner-agnostic predictions on the protein side of the interface; (c) highlight inaccurate interpretations of results that might lead to misleading conclusions and claims in that area. We also evaluated the performance of the two existing interface prediction methods, PS-PRIP and PRIdictor, which are publically available through online web servers. In addition to widely used standard performance evaluation metrics, we employed a novel metric the RNA-Specificity Metric (RSM), for evaluating how much prediction results depend on the sequence or structure of the partner RNA. Our results revealed that the predictive performance of RNA partner-specific methods is no better than that of partner-agnostic methods, and that predictions of the existing "partner-specific" methods are actually oblivious to the partner RNAs. Our results highlight the importance of taking steps to eliminate redundancy between training and test datasets and of determining which features actually contribute to the accuracy of machine learning-based predictors. Failure to do so can lead to inaccurate interpretations of results and misleading conclusions. RSM is a threshold-dependent metric, as are accuracy, sensitivity, and specificity. Therefore, it can be sensitive to choice of the threshold used to assign binary labels. Our ongoing work aims to develop a novel RSM score that is threshold-independent (ie, could be estimated directly from predicted probabilities as opposed to predicted interface/non-interface labels). Our future work also aims to adapt an RSM-like metric for partner-specific protein-protein interface residue prediction and assessing the sensitivity of existing tools to changes in the putative interacting partner. We believe that this work identifies important factors to be considered in developing partner-specific protein-RNA interface prediction methods and by extension, partner-specific methods for predicting protein-protein, protein-DNA and other types of macromolecular interfaces, complexes, and interactions.

ACKNOWLEDGMENT

This work was supported in part by grants from the National Science Foundation (ACI 1640834) and the National Institutes of Health (NCATS UL1 TR002014-01), and the Center for Big Data Analytics and Discovery Informatics which is co-sponsored by the Institute for Cyberscience, the Huck Institutes of the Life Sciences, and the Social Science Research Institute at the Pennsylvania State University, and the Edward Frymoyer Endowed Professorship at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science sponsored by the Pratiksha Trust at the Indian Institute of Science (both held by Vasant G. Honavar).

CONFLICT OF INTERESTS

The authors declare no conflict of interest.

ORCID

Yong Jung https://orcid.org/0000-0002-8493-4390
Vasant G. Honavar https://orcid.org/0000-0001-5399-3489

REFERENCES

1. Fabian MR, Sonenberg N, Filipowicz W. Regulation of mRNA translation and stability by microRNAs. *Annu Rev Biochem*. 2010;79:351-379.
2. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol*. 2008;6:e255.
3. Licatalosi DD, Darnell RB. RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*. 2010;11:75-87.
4. Lukong KE, Chang K-w, Khandjian EW, Richard S. RNA-binding proteins in human genetic disease. *Trends Genet*. 2008;24:416-425.
5. Moore MJ. From birth to death: the complex lives of eukaryotic mRNAs. *Science*. 2005;309:1514-1518.
6. Marchese D, de Groot NS, Lorenzo Gotor N, Livi CM, Tartaglia GG. Advances in the characterization of RNA-binding proteins. *Wiley Interdiscip Rev RNA*. 2016;7:793-810.
7. Kutluay SB, Zang T, Blanco-Melo D, et al. Global changes in the RNA binding specificity of HIV-1 gag regulate virion genesis. *Cell*. 2014;159:1096-1109.
8. Castello A, Fischer B, Hentze MW, Preiss T. RNA-binding proteins in Mendelian disease. *Trends Genet*. 2013;29:318-327.
9. Jankowsky E, Harris ME. Specificity and nonspecificity in RNA-protein interactions. *Nat Rev Mol Cell Biol*. 2015;16:533-544.
10. Lunde BM, Moore C, Varani G. RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*. 2007;8:479-490.
11. Ban T, Zhu J-K, Melcher K, Xu HE. Structural mechanisms of RNA recognition: sequence-specific and non-specific RNA-binding proteins and the Cas9-RNA-DNA complex. *Cell Mol Life Sci*. 2015;72:1045-1058.
12. Zhang C, Muench DG. A nucleolar PUF RNA-binding protein with specificity for a unique RNA sequence. *J Biol Chem*. 2015;290:30108-30118.
13. Friesen WJ, Darby MK. Specific RNA binding proteins constructed from zinc fingers. *Nat Struct Mol Biol*. 1998;5:543-546.
14. Parker R, Song H. The enzymes and control of eukaryotic mRNA turnover. *Nat Struct Mol Biol*. 2004;11:121-127.
15. Aitken CE, Lorsch JR. A mechanistic overview of translation initiation in eukaryotes. *Nat Struct Mol Biol*. 2012;19:568-576.
16. Ke A, Doudna JA. Crystallization of RNA and RNA–protein complexes. *Methods*. 2004;34:408-414.
17. Wu H, Finger LD, Feigon J. Structure determination of protein-RNA complexes by NMR. *Methods Enzymol*. 2005;394:525-545.
18. König J, Zarnack K, Rot G, et al. iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*. 2010;17:909-915.
19. Jones S. Protein–RNA interactions: structural biology and computational modeling techniques. *Biophys Rev*. 2016;8:359-367.
20. Terribilini M, Sander JD, Lee J-H, et al. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res*. 2007;35:W578-W584.

21. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, 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*. 2014;9:e97725.
22. Miao Z, Westhof E. A large-scale assessment of nucleic acids binding site prediction programs. *PLoS Comput Biol*. 2015;11:e1004639.
23. Yang X, Wang J, Sun J, Liu R. SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. *PLoS One*. 2015;10:e0133260.
24. Muppirala U, Lewis B, Dobbs D. Computational tools for investigating RNA-protein interaction partners. *J Com Sci Comput Biol*. 2013;6:182.
25. Miao Z, Westhof E. Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. *Nucleic Acids Res*. 2015;43:5340-5351.
26. Wang L, Huang C, Yang MQ, Yang JY. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Syst Biol*. 2010;4:S3.
27. Pérez-Cano L, Fernández-Recio J. Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. *Proteins*. 2010;78:25-35.
28. Zhao H, Yang Y, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Res*. 2011;39:3017-3025.
29. Yan J, Kurgan L. DRNApred, fast sequence-based method that accurately predicts and discriminates DNA-and RNA-binding residues. *Nucleic Acids Res*. 2017;45:e84-e84.
30. Yan J, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues. *Brief Bioinform*. 2015;17:88-105.
31. Ahmad S, Mizuguchi K. Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS One*. 2011;6:e29104.
32. Minhas A, ul Amir F, Geiss BJ, Ben-Hur A. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins*. 2014;82:1142-1155.
33. Xue LC, Dobbs D, Honavar V. HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC bioinformatics*. 2011;12:244.
34. Xue LC, Dobbs D, Bonvin AM, Honavar V. Computational prediction of protein interfaces: a review of data driven methods. *FEBS Lett*. 2015;589:3516-3526.
35. Esmaielbeiki R, Krawczyk K, Knapp B, Nebel J-C, Deane CM. Progress and challenges in predicting protein interfaces. *Brief Bioinform*. 2016;17:117-131.
36. Sela-Culang I, Ofran Y, Peters B. Antibody specific epitope prediction—emergence of a new paradigm. *Curr Opin Virol*. 2015;11:98-102.
37. Wong K-C, Li Y, Peng C, Moses AM, Zhang Z. Computational learning on specificity-determining residue-nucleotide interactions. *Nucleic Acids Res*. 2015;43:10180-10189.
38. Tuvshinjargal N, Lee W, Park B, Han K. PRIdictor: protein–RNA interaction predictor. *Biosystems*. 2016;139:17-22.
39. Muppirala U, Lewis B, Mann C, Dobbs D (2016) A motif-based method for predicting interfacial residues in both the RNA and protein components of protein-RNA complexes. Pacific Symposium on Biocomputing Pacific Symposium on Biocomputing: NIH Public Access. pp. 445.
40. Luo J, Liu L, Venkateswaran S, Song Q, Zhou X. RPI-bind: a structure-based method for accurate identification of RNA-protein binding sites. *Sci Rep*. 2017;7:614.
41. Vapnik V. *The Nature of Statistical Learning Theory*. Berlin: Springer Science & Business Media; 2013.
42. Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*. 2000;16:412-424.
43. Fawcett T. An introduction to ROC analysis. *Pattern Recogn Lett*. 2006;27:861-874.
44. Walia RR, Caragea C, Lewis BA, et al. Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC bioinformatics*. 2012;13:89.
45. Berrar D, Flach P. Caveats and pitfalls of ROC analysis in clinical microarray research (and how to avoid them). *BMC bioinformatics*. 2012;13(1):83-97.
46. Yu T. ROCS: receiver operating characteristic surface for class-skewed high-throughput data. *PLoS One*. 2012;7:e40598.
47. Swamidass SJ, Azencott C-A, Daily K, Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval. *Bioinformatics*. 2010;26:1348-1356.
48. Berman HM, Westbrook J, Feng Z, et al. The protein data bank. *Nucleic Acids Res*. 2000;28:235-242.
49. Altschul SF, Madden TL, Schäffer AA, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*. 1997;25:3389-3402.
50. Wang Y, Xue Z, Shen G, Xu J. PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*. 2008;35:295-302.
51. EL-Manzalawy Y, Abbas M, Malluhi Q, Honavar V. FastRNABindR: fast and accurate prediction of protein-RNA Interface residues. *PLoS One*. 2016;11:e0158445.
52. Heinig M, Frishman D. STRIDE: a web server for secondary structure assignment from known atomic coordinates of proteins. *Nucleic Acids Res*. 2004;32:W500-W502.
53. Chothia C. The nature of the accessible and buried surfaces in proteins. *J Mol Biol*. 1976;105:1-12.
54. Panwar B, Raghava GP. Identification of protein-interacting nucleotides in a RNA sequence using composition profile of tri-nucleotides. *Genomics*. 2015;105:197-203.
55. Lu X-J, Bussemaker HJ, Olson WK. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Res*. 2015;43:e142-e142.
56. Mattei E, Ausiello G, Ferrè F, Helmer-Citterich M. A novel approach to represent and compare RNA secondary structures. *Nucleic Acids Res*. 2014;42:6146-6157.
57. Breiman L. Random forests. *Mach Learn*. 2001;45:5-32.
58. Michalski RS, Carbonell JG, Mitchell TM. *Machine Learning: An Artificial Intelligence Approach*. Berlin: Springer Science & Business Media; 2013.
59. Auweter SD, Oberstrass FC, Allain FH-T. Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res*. 2006;34:4943-4959.
60. Helder S, Blythe AJ, Bond CS, Mackay JP. Determinants of affinity and specificity in RNA-binding proteins. *Curr Opin Struct Biol*. 2016;38:83-91.
61. Zhang J, Ma Z, Kurgan L. Comprehensive review and empirical analysis of hallmarks of DNA-, RNA-and protein-binding residues in protein chains. *Brief Bioinform*. 2017;bbx168.
62. Yang X, Li H, Huang Y, Liu S. The dataset for protein–RNA binding affinity. *Protein Sci*. 2013;22:1808-1811.
63. Bhattacharya R, Rose PW, Burley SK, Prlić A. Impact of genetic variation on three dimensional structure and function of proteins. *PLoS One*. 2017;12:e0171355.
64. Cook KB, Hughes TR, Morris QD. High-throughput characterization of protein–RNA interactions. *Brief Funct Genomics*. 2014;14:74-89.
65. Sutandy FR, Hsiao FS-H, Chen C-S. High throughput platform to explore RNA–protein interactomes. *Crit Rev Biotechnol*. 2016;36:11-19.
66. Guenther U-P, Yandek LE, Niland CN, et al. Hidden specificity in an apparently nonspecific RNA-binding protein. *Nature*. 2013;502:385-388.
67. Milek M, Wyler E, Landthaler M. *Transcriptome-Wide Analysis of Protein–RNA Interactions Using High-Throughput Sequencing. Seminars in cell & developmental biology*. 2012;23:206-212. New York: Academic Press.
68. Jones S, Daley DT, Luscombe NM, Berman HM, Thornton JM. Protein–RNA interactions: a structural analysis. *Nucleic Acids Res*. 2001;29:943-954.
69. Nagai K. RNA—protein complexes. *Curr Opin Struct Biol*. 1996;6:53-61.
70. Hermann T, Westhof E. Non-Watson-crick base pairs in RNA-protein recognition. *Chem Biol*. 1999;6:R335-R343.
71. Fernandez M, Kumagai Y, Standley DM, Sarai A, Mizuguchi K, Ahmad S. Prediction of dinucleotide-specific RNA-binding sites in proteins. *Bmc Bioinformatics*. 2011;12:S5.
72. Morozova N, Allers J, Myers J, Shamoo Y. Protein–RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*. 2006;22:2746-2752.
73. Zheng S, Robertson TA, Varani G. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *FEBS J*. 2007;274:6378-6391.

74. Ehresmann C, Baudin F, Mougel M, Romby P, Ebel J-P, Ehresmann B. Probing the structure of RNAs in solution. *Nucleic Acids Res*. 1987;15:9109-9128.
75. Tan Z-J, Chen S-J. Salt contribution to RNA tertiary structure folding stability. *Biophys J*. 2011;101:176-187.
76. Yu J-F, Dou X-H, Sha Y-J, et al. DisBind: a database of classified functional binding sites in disordered and structured regions of intrinsically disordered proteins. *BMC bioinformatics*. 2017;18:206.
77. Gerstberger S, Hafner M, Tuschl T. A census of human RNA-binding proteins. *Nat Rev Genet*. 2014;15:829-845.
78. Anantharaman V, Koonin EV, Aravind L. Comparative genomics and evolution of proteins involved in RNA metabolism. *Nucleic Acids Res*. 2002;30:1427-1464.
79. Cléry A, Blatter M, Allain FH. RNA recognition motifs: boring? Not quite. *Curr Opin Struct Biol*. 2008;18:290-298.
80. Fairman-Williams ME, Guenther U-P, Jankowsky E. SF1 and SF2 helicases: family matters. *Curr Opin Struct Biol*. 2010;20:313-324.
81. Dyson HJ. Roles of intrinsic disorder in protein–nucleic acid interactions. *Mol BioSyst*. 2012;8:97-104.
82. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. *Nucleic Acids Res*. 2015;43:e121-e121.
83. Jones DT, Cozzetto D. DISOPRED3: precise disordered region predictions with annotated protein-binding activity. *Bioinformatics*. 2014;31:857-863.
84. Hanson J, Yang Y, Paliwal K, Zhou Y. Improving protein disorder prediction by deep bidirectional long short-term memory recurrent neural networks. *Bioinformatics*. 2016;33:685-692.
85. Meng F, Kurgan L. DFLpred: high-throughput prediction of disordered flexible linker regions in protein sequences. *Bioinformatics*. 2016;32:i341-i350.
86. Hafner M, Landthaler M, Burger L, et al. Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*. 2010;141:129-141.
87. Tuszynska I, Bujnicki JM. DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. *BMC bioinformatics*. 2011;12:348.
88. Si J, Cui J, Cheng J, Wu R. Computational prediction of RNA-binding proteins and binding sites. *Int J Mol Sci*. 2015;16:26303-26317.
89. Fornes O, Garcia-Garcia J, Bonet J, Oliva B. On the use of knowledge-based potentials for the evaluation of. *Adv Protein Chem Struct Biol*. 2014;94:77.
90. Tuszynska I, Matelska D, Magnus M, et al. Computational modeling of protein–RNA complex structures. *Methods*. 2014;65:310-319.
91. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. Computational methods for prediction of protein–RNA interactions. *J Struct Biol*. 2012;179:261-268.
92. Madan B, Kasprzak JM, Tuszynska I, Magnus M, Szczepaniak K, et al. Modeling of protein–RNA complex structures using computational docking methods. *Computational Design of Ligand Binding Proteins*; New York, NY: Humana Press. 2016:353-372.
93. Peng J, Xu J. A multiple-template approach to protein threading. *Proteins*. 2011;79:1930-1939.
94. Zhang Y. Template-based modeling and free modeling by I-TASSER in CASP7. *Proteins*. 2007;69:108-117.
95. Zhao H, Yang Y, Zhou Y. Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Mol BioSyst*. 2013;9:2417-2425.
96. Yu X, Cao J, Cai Y, Shi T, Li Y. Predicting rRNA-, RNA-, and DNA-binding proteins from primary structure with support vector machines. *J Theor Biol*. 2006;240:175-184.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Jung Y, EL-Manzalawy Y, Dobbs D, Honavar VG. Partner-specific prediction of RNA-binding residues in proteins: A critical assessment. *Proteins*. 2019;87:198–211. https://doi.org/10.1002/prot.25639