HHS Public Access

Author manuscript

Pac Symp Biocomput. Author manuscript; available in PMC 2016 January 21.

Published in final edited form as:

Pac Symp Biocomput. 2016 ; 21: 445–455.

A MOTIF-BASED METHOD FOR PREDICTING INTERFACIAL RESIDUES IN BOTH THE RNA AND PROTEIN COMPONENTS OF PROTEIN-RNA COMPLEXES

USHA MUPPIRALA[#*],

Genome Informatics Facility, Iowa State University, Ames, Iowa, 50011, USA

BENJAMIN A LEWIS[#],

Department of Computer Science, Truman State University, Kirksville, Missouri, 63501, USA, benlewis@truman.edu

CARLA M. MANN, and

Bioinformatics and Computational Biology Program, Iowa State University, Ames, Iowa, 50011, USA, cmmann@iastate.edu

DRENA DOBBS

Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, 50011, USA, ddobbs@iastate.edu

[#] These authors contributed equally to this work.

Abstract

Efforts to predict interfacial residues in protein-RNA complexes have largely focused on predicting RNA-binding residues in proteins. Computational methods for predicting protein-binding residues in RNA sequences, however, are a problem that has received relatively little attention to date. Although the value of sequence motifs for classifying and annotating protein sequences is well established, sequence motifs have not been widely applied to predicting interfacial residues in macromolecular complexes. Here, we propose a novel sequence motif-based method for "partner-specific" interfacial residue prediction. Given a specific protein-RNA pair, the goal is to simultaneously predict RNA binding residues in the protein sequence and protein-binding residues in the RNA sequence. In 5-fold cross validation experiments, our method, PS-PRIP, achieved 92% Specificity and 61% Sensitivity, with a Matthews correlation coefficient (MCC) of 0.58 in predicting RNA-binding sites in proteins. The method achieved 69% Specificity and 75% Sensitivity, but with a low MCC of 0.13 in predicting protein binding sites in RNAs. Similar performance results were obtained when PS-PRIP was tested on two independent "blind" datasets of experimentally validated protein-RNA interactions, suggesting the method should be widely applicable and valuable for identifying potential interfacial residues in protein-RNA complexes for which structural information is not available. The PS-PRIP webserver and datasets are available at: *http://pridb.gdcb.iastate.edu/PSPRIP/*.

[*] Corresponding author, usha@iastate.edu.

1. Introduction

Despite the important roles of protein-RNA interactions in many biological processes, including transcription, translation, viral replication and pathogen resistance [1-2], the mechanisms and regulation of protein-RNA recognition are not yet fully understood. The Protein Data Bank (PDB) is a valuable resource for studying protein-RNA complexes, but protein-RNA complexes constitute less than 1% of the total structures in the database [3]. Recently, high-throughput (HTP) methods for identifying the *in vivo* targets of specific RNA binding proteins - and the RNA motifs they bind - have provided a wealth of information about the determinants of sequence recognition in protein-RNA complexes [4-6]. Data from both the PDB and HTP experiments have been exploited to develop several computational methods for predicting interfacial residues in protein-RNA complexes [reviewed in 7-10] as well as a few methods for predicting interaction partners in protein-RNA complexes and interaction networks [reviewed in 11-13].

Most computational approaches for predicting interfacial residues have focused on the protein side of the interface. Methods for predicting RNA-binding amino acid residues in proteins fall into two major classes: i) methods that use only sequence information, and ii) methods that take advantage of structural information, when available [8]. Only one published method [14-15] takes into account information regarding the RNA partner; the rest are “non-partner-specific” predictors of interfacial residues. Computational prediction of protein-binding ribonucleotides in RNA is a more difficult problem. The low per-character information content of the 4-ribonucleotide alphabet of unmodified RNA (i.e., ignoring modified ribonucleotides) makes this problem more challenging. One approach to overcoming this limitation is to expand the RNA alphabet by using known or predicted RNA secondary structure [16]. Another approach, taken in the current study, is to exploit short sequence motifs that occur in the interfaces of known protein-RNA complexes.

Here, we report a preliminary large scale analysis of contiguous RNA sequence motifs present in the interfaces of protein-RNA complexes and propose a new “partner-specific” motif-based method to simultaneously predict RNA-binding residues in the protein component and protein-binding ribonucleotides in the RNA component of a given protein-RNA pair.

2. Methods

2.1. Generating interfacial sequence motifs

To generate interfacial sequence motifs with which to scan target protein and RNA sequences, a dataset of 1,408 protein-RNA complex structures deposited in the Protein Data Bank (PDB) as of September 2012 was analyzed to find short strings of amino acids or ribonucleotides, contiguous in the primary sequence and composed entirely of interacting residues in either the protein or RNA chains. The sequences of these interfacial segments were extracted as ‘*n*-mer motifs’, where *n* can vary between 3 and 8. No requirement was made for motifs to be bounded by non-interacting residues; therefore, overlapping motifs were included. Thus, a 5-mer motif necessarily contains two 4-mer motifs and three 3-mer motifs.

2.2. Datasets for interface prediction

To generate datasets for evaluating the utility of motifs for interface prediction, interacting protein and RNA chains were extracted from protein-RNA complexes in the PDB with at least 3.5Å resolution. In one dataset, RPInt327, proteins of length < 25 amino acids and RNAs of length < 100 ribonucleotides were excluded. This dataset was used for training and cross-validation tests. The interaction information (i.e., interfacial residues) for these chains was downloaded from PRIDB [17]. Several additional fully independent datasets were generated to evaluate the performance of the classifier on RNAs of different lengths, e.g., RPInt79 (RNAs > 250 nts) and RPInt83 (RNAs 50-100 nts). The interfacial residues for these chains were computing using contact-chainID [18]. For both datasets, residues in protein and RNA chains were defined as interacting if any heavy atom in one chain lies within a 5Å distance cutoff of any heavy atom in the other chain. Based on BLASTClust results, redundant protein sequences (i.e., with ≥ 30% sequence identity) in complexes with similar RNA sequences (i.e., with ≥ 30% sequence identity) were discarded; RNA sequences in such redundant complexes were also discarded. For RPInt327, this resulted in a non-redundant dataset containing a total of 1,637 interacting protein-RNA pairs. 327 pairs were kept aside for independent evaluation and 5-fold cross-validation was performed on the remaining 1,310 pairs. Datasets RPInt79 and RPInt83 were reserved as a fully independent test datasets and were not used for training or cross-validation in this study.

2.3. Generating a protein-RNA interface motif lookup table

As illustrated in Figure 1, the protein-RNA interface motif lookup table consists of pairs of protein and RNA interfacial sequence motifs that are known to contact one another in a characterized protein-RNA complex. Entries in the lookup table were obtained as follows: First, the protein sequences in the non-redundant dataset of 1,637 protein-RNA pairs were scanned for interfacial sequence motifs (identified as described above) using a sliding window approach. Similarly, RNA sequences were scanned for interfacial sequence motifs (Fig. 1A). Second, each pair of protein-RNA sequences in the training dataset of known protein-RNA complexes was examined to identify cases in which there exists at least one physical contact (<5Å) between a heavy atom in any of the amino acids and any heavy atom in any of the ribonucleotides in a corresponding pair of sequence motifs (Fig. 1B & C). If a physical interaction is detected, that particular protein-RNA sequence motif pair is added to the lookup table (Fig. 1D).

2.4. Motif-based prediction of interfacial residues in both RNA and protein

After generating the protein-RNA interface motif lookup table, prediction of interfacial residues in a query protein-RNA pair is done in a single step. The protein and RNA sequences are scanned simultaneously for the presence of motif pairs in the lookup table. If any motif pair is present, those amino acids and ribonucleotides are marked as “interfacial” in the given query sequences. The remaining residues and ribonucleotides are marked as non-interfacial residues. For example, using the lookup table in Figure 1, if ‘TRTYR’ is found in the query protein and ‘UUAAU’ is found in the query RNA, the corresponding amino acids and ribonucleotides are predicted as interfacial residues.

2.5. Performance evaluation

We used the following measures to evaluate the performance of motif-based prediction of interfacial residues on both proteins and RNAs. True Positive (TP) refers to the number of interfacial residues correctly identified as such by the method. False Positive (FP) refers to the number of non-interfacial residues misclassified as interfacial residues. False Negative (FN) refers to the number of interfacial residues misclassified as non-interfacial residues. True Negative (TN) refers to the number of non-interfacial residues correctly identified as such by the method. Note that here our definition of Sensitivity (true positive rate) is the same as Recall. We compute both Specificity (true negative rate), here as defined as in medical statistics literature, and Precision, which is referred to as Specificity in the machine learning literature [19].

$$Sensitivity\,(recall) = \frac{TP}{TP+FN} \quad (1)$$

$$Specificity = \frac{TN}{TN+FP} \quad (2)$$

$$Precision = \frac{TP}{TP+FP} \quad (3)$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\,(TP+FN)\,(TN+FP)\,(TN+FN)}} \quad (4)$$

3. Results

3.1. Motif-based partner-specific prediction of interfacial residues

To evaluate whether an interface motif lookup table can be used to predict interfacial residues in specific protein-RNA pairs, we first performed preliminary experiments in which we tested the effect of varying the length of protein motifs from 4 to 6 amino acids, and the length of RNA motifs from 4 to 8 ribonucleotides (see *Methods*). As expected, using shorter motifs resulted in a larger number of false positive predictions, whereas using longer motifs resulted in larger number of false negative predictions. Based on these results, we determined that a protein motif of length 5 provides a good balance between prediction specificity and sensitivity. Although there are 205 different potential combinations of amino acid 5-mers, only 0.3% (11,269) of the theoretically possible 5-mer motifs were observed in interfaces extracted from known protein-RNA complexes (1,408 complexes, comprising 17,385 protein chains) in the PRIDB database [17].

To predict RNA-binding residues in the protein component of a given protein-RNA pair, we used a protein motif size of length 5 and varied the RNA motif lengths from 4 to 6. Table 1 summarizes results obtained using a 5-fold cross validation approach, in which 80% of the data was used to generate the protein-RNA motif lookup table and predictions were made on the remaining 20% of the data. There is little difference in the *Specificity* or *Matthews*

correlation coefficient (MCC) using RNA motifs of length 4 and 5. Although using an RNA motif of length 6 resulted in higher *Specificity* (0.94), it resulted in lower *Sensitivity* and *MCC*. Using an RNA 4-mer resulted in higher *Sensitivity* (0.65) compared with using 5- and 6-mers.

To predict which ribonucleotides in the RNA component of a given protein-RNA pair participate in protein binding, we again used a protein motif size of length 5 and varied the RNA motif lengths from 4 to 6. Table 2 summarizes the prediction results obtained in 5-fold cross-validation experiments. Again, as the RNA motif size is increased, the *Specificity* increased, but with the expected decrease in *Sensitivity*. A high *Specificity* of 0.91 is obtained using an RNA motif length of 6, but the corresponding *MCC* is much lower than that obtained for RNA binding site prediction (Table 1].

3.2. Performance evaluation on independent test sets

To more rigorously test the performance of the method, we evaluated it on several independent datasets of known protein-RNA pairs (See *Methods*). As summarized in Table 3, on the RP327 dataset (which contains 327 protein-RNA pairs), using protein and RNA motifs of length 5, we obtained 92% *Specificity* and 64% *Sensitivity* in predicting RNA-binding residues. In predicting protein-binding ribonucleotides, the *Specificity* was 67% and *Sensitivity* was 79%. Thus, performance on the independent test set was comparable to that obtained in cross-validation experiments. This suggests that our proposed “partner-specific” method for predicting protein-RNA interfaces using sequence motifs, which we call PS-PRIP, should be generally applicable.

To investigate the influence of RNA length on performance, we also evaluated the classifier on several additional independent datasets of complexes containing RNAs of different lengths (Table 4 and data not shown). Although PS-PRIP performs very well on complexes containing RNAs longer than 250 nts, performance is poor on complexes containing shorter RNAs.

3.3. Comparison with other interface prediction methods

Only one other published study has addressed the prediction of interfacial residues in protein and RNA components of protein-RNA complexes simultaneously. The catRAPID method proposed by Bellucci *et al.* [20] divides the protein and RNA sequences into a number of fragments and calculates interaction propensities between each pair of protein-RNA fragments. Because binding site prediction on a per residue basis was not reported, we could not directly compare our method with catRAPID.

A method for predicting protein-binding sites in RNAs was reported by Choi and Han [14-15]. We have not been able to make direct performance comparisons with this method because neither the test dataset nor a working webserver is available, and we did not attempt to re-implement it in order to provide a direct comparison with our method. In an earlier report, Choi and Han also proposed a partner-specific RNA binding site prediction method, in which the RNA sequence is encoded as the sum of the normalized positions of each nucleotide (A, C, G and U) in the sequence [14]. When we examined the dataset used in that

study, we noticed that all except one RNA sequence was less than 100 nucleotides in length, and approximately half of the dataset consists of very short RNAs (< 15 nts). Because the minimum length of the RNA used in our training dataset is 100 nt, and, as discussed in the next section, our method is not suitable for small RNAs, we did not compare PS-PRIP with Choi and Han's method. Choi and Han reported prediction performance of 91% specificity and 60.7% sensitivity with a CC of 0.24 on a dataset of 267 interacting protein-RNA pairs [14].

We were able to compare the performance of our *partner-specific* PS-PRIP method with existing *non-partner* specific sequence-based methods for predicting RNA-binding residues in proteins. Walia *et al.* [8] performed a systematic comparison of existing methods for predicting RNA-binding residues and showed that PSSM-based methods had the best performance among published sequence-based approaches. Thus, we directly compared the performance of PS-PRIP with RNABindRPlus [21], which combines homology-based predictions with predictions from an optimized SVM classifier that uses a PSSM-based approach. Because homology-based methods exploit existing structures and interfaces, and our independent test set was extracted from the PDB, we expected the homology-based method to perform very well. Homology-based methods fail, however, when the query sequence has no homologs in the PDB. We also compared our method with the SVM component of RNABindRPlus and the results are also shown in Table 6. PS-PRIP has better performance in terms of *Specificity* (0.92), but lower *Sensitivity* (0.64) compared to RNABindRPlus. RNABindRPlus had the highest *MCC* (0.71); the MCCs for the other two methods were similar (0.59 vs 0.61). A larger difference is seen in the *Precision* (or positive prediction rate) of the two methods: PS-PRIP has higher *Precision* (0.80) than RNABindRPlus (0.76), when evaluated on this dataset.

4. Discussion

This study suggests that specific subsets of short contiguous interfacial motifs are over-represented relative to others within the sequences of both protein and RNA components of known protein-RNA complexes. A large number of interfacial amino acid motifs occur only once in the dataset analyzed here. This may be a consequence of the criteria for generating the short RNA-binding motifs in this study: all residues in an interfacial motif must be contiguous in sequence and must interact with at least one atom in a ribonucleotide within a 5 Å distance cutoff. It is striking that a simple lookup table of motif pairs, identified in a training set of protein-RNA complexes, can be used to accurately predict interfacial residues in an independent set of complexes. Although we have not yet directly calculated the interface propensities of these motifs (i.e., the over-representation of these motifs in interfacial versus non-interfacial regions of the protein and RNA sequences), it should be possible to improve prediction of interfacial residues by focusing on motifs with high interface propensity.

The interface prediction results reported here demonstrate that an ribonucleotide motif of length 5, while not informative on its own, can be highly informative when used in combination with an amino acid motif of length 5. From the non-redundant dataset of protein-RNA complexes used in this study, we generated a lookup table of 55,154 protein-

RNA motif pairs, comprising 3,275 unique protein motifs and 835 unique RNA motifs. Using a non-redundant dataset is the appropriate way to evaluate and compare interface prediction methods, but doing so is expected to exclude some informative motif combinations. Thus, we created a motif lookup table *without* discarding redundant motifs. As expected, many additional protein-RNA motif pairs were identified: a total of 88,994 protein-RNA motif pairs, comprising 4,035 protein motifs and 893 RNA motifs.

Our results indicate that binding partner information, which has been largely ignored for predicting interfacial residues in protein-RNA complexes, can be valuable for making "partner-specific" interface predictions. Figures 2 and 3 illustrate this with an example. In the *E. coli* ribosome, the 16S rRNA in the small subunit interacts with various protein components of the 30S subunit, using different binding sites. Interaction of S4 and S11 proteins with a segment of the 16S ribosomal RNA (PDB 4GAS) is shown in Figure 2. In this structure, the majority of 16S rRNA nucleotides that bind the S4 protein are located in the region between 400 – 440 nt. In contrast, the region between 670 – 720 nt of 16S rRNA contains most of the S11 protein-binding residues. Whereas a *non*-partner specific method would not be able to distinguish between these, Figure 3 shows that PS-PRIP makes distinct binding site predictions for the S4 and S11 proteins. In the 16S rRNA sequence between 386 – 437 nt, many S4 binding residues are correctly predicted. In the same region (where S11 does *not* bind), a few residues are incorrectly predicted as interacting with S11 in complexes that contain short RNAs (< 100 nts). Short RNAs, which often correspond to interface-containing fragments of the much longer RNAs present in native complexes, are common in structurally characterized protein-RNA complexes in the PDB. Thus, the likelihood that every ribonucleotide in such an RNA is an interfacial residue is very high compared to the situation for longer RNAs, in which only a small fraction of the ribonucleotides directly contact the bound protein(s). Because of this, we excluded RNAs <100 nts in length for generating motifs (see *Methods*), which results in a bias in our training set for RNA-protein pairs derived from ribosomes. In our experiments, PS-PRIP performed well on RNAs >100 nts in length (Table 3), but poorly when tested on RNAs < 100 nts (Table 4). Thus, PS-PRIP can be used to predict protein-binding sites in mRNAs, rRNAs, long non-coding RNAs and many short ncRNAs, but predictions on RNAs less than 100 nts are likely to be unreliable. Current work is directed at generating "custom" classifiers trained on datasets containing RNAs of variable length to obtain optimal performance on RNAs of different lengths and different functional classes (e.g., non-ribosomal ncRNAs, including sRNAs, sncRNAs, etc.)

In future work, we plan to evaluate the effect of incorporating predicted RNA secondary structure in the RNA sequence representation, which is expected to lead to better performance in predicting protein-binding residues in RNA [16]. In addition, we plan to test whether exploiting the extensively characterized resource of structural motifs in RNAs [22-23], can provide further improvement.

5. Conclusions

We have developed a new method for predicting partner-specific interfacial residues in protein-RNA complexes using short sequence motifs. PS-PRIP can simultaneously predict interfacial residues in both the protein and RNA components of a complex, albeit with much

greater reliability for the protein component. An RNA motif of length 5, in combination with a protein motif of length 5, can be used to predict interfacial residues with high specificity (0.92 for RNA-binding residues in proteins; 0.67 for protein-binding residues in RNA), indicating that PS-PRIP can be a valuable tool for experimentalists who wish to target interfaces in specific protein-RNA complexes or to perturb specific interactions in protein-RNA interaction networks. A PS PRIP webserver and all training and test datasets used in this study are freely available online at: *http://pridb.gdcb.iastate.edu/PSPRIP/*.

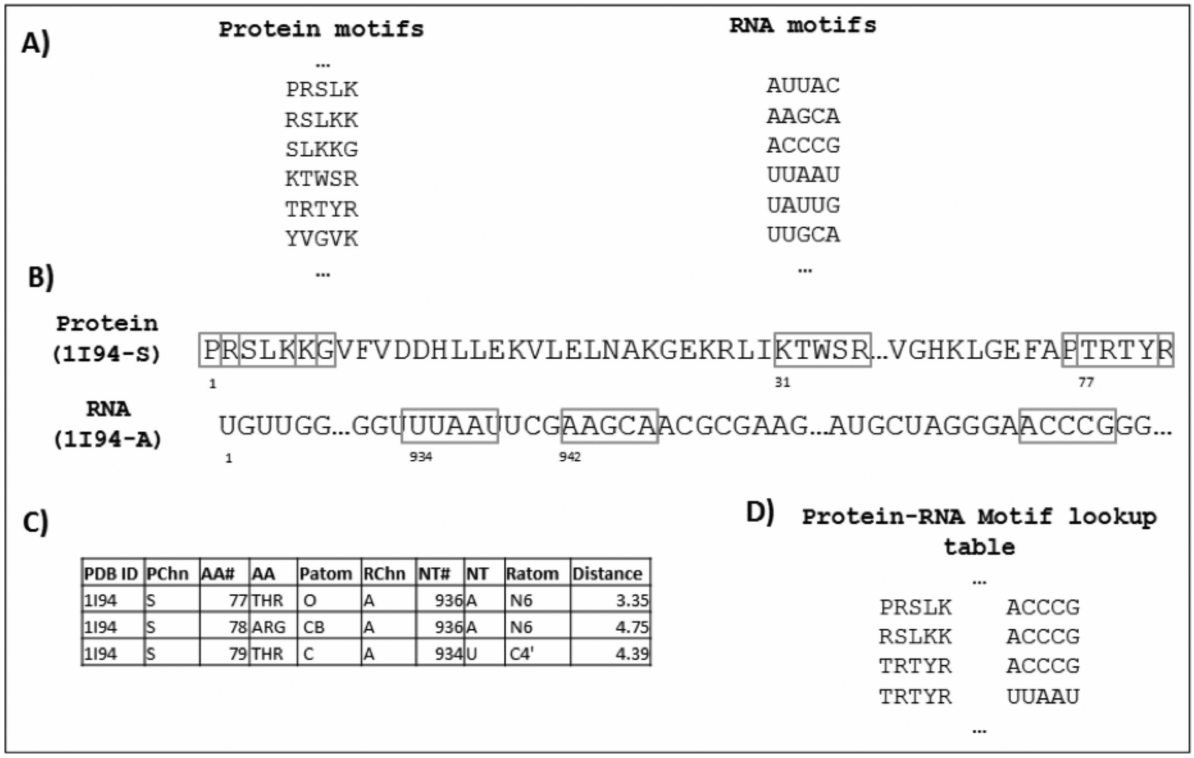
Acknowledgements

Authors wish to thank Rasna Walia for valuable discussions and NIH R01 GM066387 for funding to support portions of this work.

References

1. Licatalosi DD, Darnell RB. Nat. Rev. Genet. 2010; 11:75. [PubMed: 20019688]
2. Re A, Joshi T, Kulberkyte E, Morris Q, Workman CT. Methods Mol. Biol. 2014; 1097:491. [PubMed: 24639174]
3. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. Nucleic Acids Res. 2000; 8:235. [PubMed: 10592235]
4. McHugh CA, Russell P, Guttman M. Genome Biol. 2014; 15:203. [PubMed: 24467948]
5. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, Li X, Gueroussov S, Albu M, Zheng H, Yang A, Na H, Irimia M, Matzat LH, Dale RK, Smith SA, Yarosh CA, Kelly SM, Nabet B, Mecenas D, Li W, Laishram RS, Qiao M, Lipshitz HD, Piano F, Corbett AH, Carstens RP, Frey BJ, Anderson RA, Lynch KW, Penalva LOF, Lei EP, Fraser AG, Blencowe BJ, Morris QD, Hughes TR. Nature. 2013; 499:172. [PubMed: 23846655]
6. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR. Nucleic Acids Res. 2011; 39:D301. [PubMed: 21036867]
7. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM. J. Struct. Biol. 2012; 179:261. [PubMed: 22019768]
8. Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D, Honavar V. BMC Bioinformatics. 2012; 13:89. [PubMed: 22574904]
9. Zhao H, Yang Y, Zhou Y. Mol. BioSyst. 9 2417. 2013
10. Yan J, Friedrich S, Kurgan L. Brief Bioinform. May 1.2015 doi: 10.1093/bib/bbv023.
11. Muppirala UK, Lewis BA, Dobbs D. J. Comput. Sci. Syst. Biol. 2013; 6:182.
12. Cirillo D D, Livi CM, Agostini F, Tartaglia GG. Mol. BioSyst. 2014; 10:1632. [PubMed: 24756571]
13. Muppirala UK, Mann CM, Dobbs D. Methods Mol. Biol. In press. 2015
14. Choi S, Han K. BMC Bioinformatics. 2011; 12:S7. [PubMed: 22373313]
15. Choi S, Han K. Comp. Biol. Med. 2013; 43(11):1687.
16. Li X, Kazan H, Lipshitz HD, Morris Q. Wiley Interdiscip Rev RNA. 2014; 5:111. [PubMed: 24217996]
17. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, Honavar V, Dobbs D. Nucleic Acids Res. 2011; 39:D277. [PubMed: 21071426]
18. Dominguez C, Boelens R, Bonvin AMJJ. J. Am. Chem. Soc. 2003; 125:1731. [PubMed: 12580598]
19. Baldi P, Brunak S, Chauvin Y, Andersen CAF, Nielsen H. Bioinformatics. 2000; 16:412. [PubMed: 10871264]
20. Bellucci M, Agostini F, Masin M, Tartaglia GG. Nat. Methods. 2011; 8:444. [PubMed: 21623348]
21. Walia RR, Xue LC, Wilkins K, El-Manzalawy Y, Dobbs D, Honavar V. PLoS ONE. 2014; 9(5):e97725. [PubMed: 24846307]

22. Petrov AI, Zirbel CL, Leontis NB. RNA. 2013; 19:1327. [PubMed: 23970545]
23. Chojnowski G, Walen T, Bujnicki JM. Nucleic Acids Res. 2014; 42:D123. [PubMed: 24220091]



PDB ID	PChn	AA#	AA	Patom	RChn	NT#	NT	Ratom	Distance
1I94	S	77	THR	O	A	936	A	N6	3.35
1I94	S	78	ARG	CB	A	936	A	N6	4.75
1I94	S	79	THR	C	A	934	U	C4'	4.39

Fig. 1.

Generation of the protein-RNA motif lookup table. **A)** A sample subset of the protein and RNA interfacial motifs used to scan target sequences. **B)** The protein and RNA sequences of each protein-RNA pair in the training dataset are scanned with the interfacial motifs. For the purpose of illustration, only a small portion of the example sequences and a subset of the interfacial motifs (indicated in boxes) are shown. **C)** Interacting residues within a distance threshold of 5Å are identified. Only a subset of interactions identified in this example is shown. **D)** Only protein and RNA motif pairs that contain at least one such interaction between them are added to the protein-RNA motif lookup table. Of the eighteen possible protein-RNA motif pairs illustrated in this example, only four satisfy this criterion and are added to the lookup table.

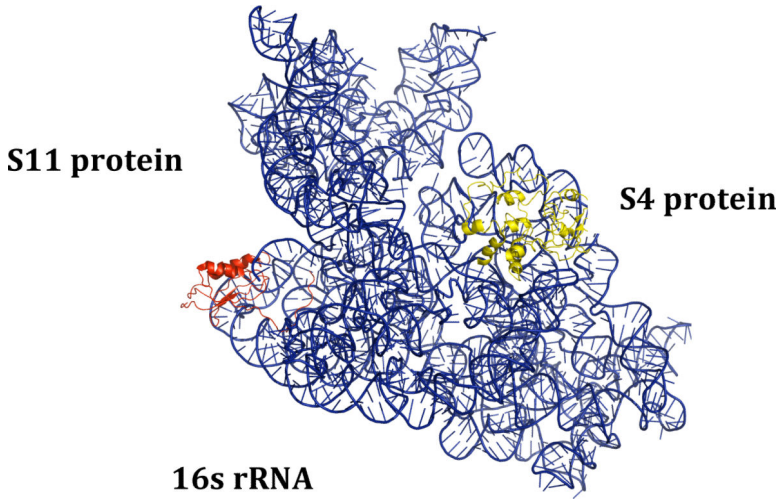


Fig. 2.

E. coli 16S ribosomal RNA (blue) interaction with S4 ribosomal protein S4 (yellow) and ribosomal protein S11 (red). (PDB ID: 4GAS)

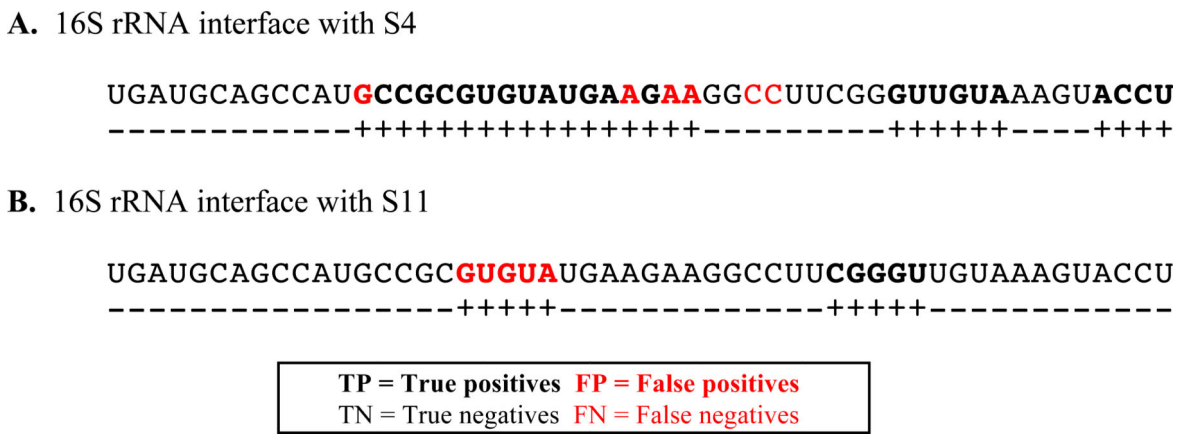


Fig. 3.

Partner-specific interface prediction in the *E, coli* 16S ribosomal RNA (PDB ID: 4GAS). Different protein-binding residues are predicted for the same RNA sequences between nt 386-437 of 16S RNA when the segment is paired with two different protein partners. **A.** Ribosomal protein S4 protein. **B.** Ribosomal protein S11 protein. '+' indicates "positive/binding" and '-' indicates "negative/non-binding" predictions.

Table 1

RNA-binding residue prediction performance using 5-fold cross-validation on a non-redundant dataset of 1,130 protein-RNA pairs

Protein motif length	RNA motif length	Specificity	Sensitivity	MCC
5	4	0.90	0.65	0.58
5	5	0.92	0.61	0.58
5	6	0.94	0.54	0.54

Table 2

Protein-binding residue prediction performance using 5-fold cross validation on a non-redundant dataset of 1,310 protein-RNA pairs

Protein Motif length	RNA motif length	Specificity	Sensitivity	MCC
5	4	0.35	0.89	0.07
5	5	0.69	0.75	0.13
5	6	0.91	0.55	0.21

Table 3

Prediction performance on an independent test set of 327 protein-RNA pairs using protein and RNA motifs of length 5.

Prediction	Specificity	Sensitivity	MCC
RNA-binding amino acids in proteins	0.92	0.64	0.59
Protein-binding nucleotides in RNA	0.67	0.79	0.13

Table 4

Performance in predicting RNA binding residues in protein-RNA complexes containing RNAs of different lengths

Prediction	Specificity	Sensitivity	MCC
RNAs > 250 nts (83 total)	0.88	0.75	0.62
RNAs 50-100 nts (79 total)	0.99	0.03	0.03

Table 6

Performance comparison of PS-PRIP and RNABindRPlus in predicting of RNA binding residues.

Method	Specificity	Sensitivity	Precision	MCC
PS-PRIP	0.92	0.64	0.80	0.59
RNABindRPlus	0.85	0.88	0.76	0.71
RNABindRPlus (SVM-only)	0.74	0.90	0.65	0.61