Chapter 15

Sequence-Based Prediction of RNA-Binding Residues in Proteins

Rasna R. Walia, Yasser EL-Manzalawy, Vasant G. Honavar, and Drena Dobbs

Abstract

Identifying individual residues in the interfaces of protein–RNA complexes is important for understanding the molecular determinants of protein–RNA recognition and has many potential applications. Recent technical advances have led to several high-throughput experimental methods for identifying partners in protein–RNA complexes, but determining RNA-binding residues in proteins is still expensive and time-consuming. This chapter focuses on available computational methods for identifying which amino acids in an RNA-binding protein participate directly in contacting RNA. Step-by-step protocols for using three different web-based servers to predict RNA-binding residues are described. In addition, currently available web servers and software tools for predicting RNA-binding sites, as well as databases that contain valuable information about known protein–RNA complexes, RNA-binding motifs in proteins, and protein-binding recognition sites in RNA are provided. We emphasize sequence-based methods that can reliably identify interfacial residues without the requirement for structural information regarding either the RNA-binding protein or its RNA partner.

Key words Protein–RNA interfaces, Binding site prediction, Machine learning, RNA-binding proteins (RBPs), Ribonucleoprotein particles (RNPs), Homology-based prediction, RNABindRPlus, SNBRFinder, PS-PRIP, FastRNABindR

1 Introduction

RNA-binding proteins (RBPs) are key regulators of cellular and developmental processes [1], playing pivotal roles in the posttranscriptional splicing and localization of mRNAs [2–5], mediating the activities of noncoding RNAs (ncRNAs) [6, 7] and even "moonlighting" as metabolic enzymes [8, 9] and promoting phase transitions to generate stress granules inside cells [10]. Defects in RBPs and ribonucleoprotein particles (RNPs) have been linked to immunological disorders [11], cancer [12, 13], and neurodegenerative diseases in humans [5, 14]. Still, even though the human genome encodes more than 1500 different RNA-binding proteins [15,

Yaoqi Zhou et al. (eds.), *Prediction of Protein Secondary Structure*, Methods in Molecular Biology, vol. 1484,
DOI 10.1007/978-1-4939-6406-2_15, © Springer Science+Business Media New York 2017

16]—at least as many RBPs as DNA-binding transcription factors [17]—our understanding of the cellular roles of RBPs, how they recognize their targets, and how they are regulated has lagged far behind our understanding of transcription factors. Recent exciting developments have begun to close this gap, providing proteome-wide catalogs and databases of RNA-binding proteins, "RNA interactomes" or "RBPomes" [18–21], an impressive compendium of RNA recognition sites [22], detailed views of the architecture and dynamics of important RNP complexes and RNA viruses, e.g., refs. [23, 24], and substantial progress in engineering RBPs with customized functions and high specificity for desired RNA targets [25, 26].

RNA-binding proteins are often modular, and many well-characterized RBPs contain one or more conserved RNA-binding domains or motifs [1, 27]. The RNA recognition motif (RRM), for example, is one of the most abundant structural motifs in vertebrate proteins, and is found in ~2 % of all human proteins [25]. Other abundant RNA-binding domains and motifs include the KH, dsRBD, DEAD-Box, PUF, SAM, and ZnF domains [1, 27], all which have conserved structures and can be recognized in the primary sequences of proteins (*see* Subheading 3.1, **step 6** below). However, only ~50 % of the mRNA-binding proteins identified by "interactome capture" in HeLa cells contain a characterized RNA-binding domain [19]. Also, many RBPs bind RNA through intrinsically disordered regions (IDRs), which are thought to promote formation of extended interaction interfaces and contribute to the generation of higher order assemblies and the formation of RNA granules [28, 29]. Finally, a survey of available structures for protein–RNA complexes revealed that the majority of amino acids in the protein–RNA interface are not part of a characterized RNA-binding motif [30] and the presence of an RNA-binding signature does not conclusively identify the specific amino acids involved in RNA recognition and binding.

The most definitive way to identify RNA-binding residues (i.e., residues that directly contact RNA) (*see* **Note 1**) is to extract them from a high-resolution experimental structure of a protein–RNA complex. Three-dimensional structures are available for only a small fraction of the known protein–RNA complexes [31]. As of December 16, 2015, the number of solved structures in the Protein Data Bank (PDB) for protein–RNA complexes is only 1661 out of 114,402 total structures, and ~40 % of the RNA-containing structures in the PDB correspond to ribosomes. Protein–RNA complexes can be very difficult to crystallize and many are too large for structure determination using NMR spectroscopy [32, 33]. Fortunately, recent advances in NMR [34], cryo-electron microscopy [35], and small-angle X-ray scattering (SAXS) [36] offer tremendous promise for providing structural details for RNPs that have been recalcitrant to experimental structure determination. At present, in the absence of a 3D structure, several types of experiments can be used to identify RNA-binding residues that are required for function (e.g., site-specific mutagenesis) or residues that are either required for high affinity binding or are located in close proximity to RNA in protein–RNA complexes, either in vivo or in vitro (e.g., co-immunoprecipitation assays, cross-linking mass spectrometry, yeast 3-hybrid assays, footprinting, and electrophoretic shift assays (reviewed in refs. [1, 27, 38]).

The development of high-throughput CHIP and RNASeq-based methods that employ a combination of in vivo cross-linking and immunoprecipitation (e.g., RIP-Chip, HITS-CLIP, PAR-CLIP, iCLIP, and CRAC) has made it possible to identify RNAs bound by specific proteins on a genome-wide scale (reviewed in refs. [1, 39, 40]). Along with these advances, several powerful integrated biochemical/bioinformatics approaches can identify both the target RNAs and the specific ribonucleotides recognized by the RNA-binding proteins [41–43]. In contrast, at present, there are no truly high-throughput experimental approaches for identifying interfacial residues in the protein component of a protein–RNA complex, although CLAMP [44] and other cross-linking and combined cross-linking mass spectrometry methods can identify interfacial residues in both the protein and RNA [37, 45, 46]. Despite all of these impressive advances, the cost and effort involved in the experimental determination of protein–RNA complex structures and/or identifying specific RNA-binding residues in proteins, has created a need for reliable computational approaches that can predict the most likely RNA-binding residues in proteins.

Computational approaches to predicting protein–RNA interfaces have been the topic of several recent reviews and benchmark comparisons [31, 47–50]. These approaches can be broadly classified into sequence- and structure-based methods [31, 47]. Sequence-based methods use sequence-derived features (such as amino acid identity or physicochemical properties) of a target residue and its sequence neighbors to make predictions. Structure-based methods use structure-derived features (such as solvent-accessible surface area or secondary structure) of a target residue and its sequence or structural neighbors to make predictions. Both sequence-based and structure-based approaches could, in theory, take advantage of recognizable RNA-binding motifs in RBPs and protein-binding motifs in their RNA targets. But, although hundreds of RNA-binding domains, motifs and signatures are annotated in the **InterPro** resource [51], at present there is no comprehensive database focused specifically on RNA-binding motifs in proteins (*see* **Note 2**). For protein-binding motifs in RNA, there is a valuable compendium of "RNA-binding motifs" (i.e., RNA motifs recognized by RBPs) [22] and excellent databases of RNA sequence motifs and binding specificities [41, 43], which provide experimentally determined recognition sites in RNA for a large number of RBPs. Also, one of the protocols provided here, **PS-PRIP** (*see* Subheading 3.3) employs a dataset of interfacial sequence motifs from RBPs and their targets to predict RNA-binding residues *and* protein-binding residues in the RNA component of specific protein–RNA complexes [52].

Recent benchmark comparisons of software and servers for predicting RNA-binding residues in proteins [31, 47] have demonstrated that the performance of methods that require only sequence information is often superior to that of methods that require structural information. One reason for this is that the best sequence-based methods encode sequences using PSSMs (Position-Specific Scoring Matrices) (*see* **Note 3**), which capture powerful evolutionary information from large multiple alignments of homologous sequences. In considering potential RNA-binding residues in a specific protein of interest, however, the user is strongly encouraged to take advantage of any available structural information, especially in evaluating the validity of predictions. For example, in most cases, RNA-binding residues are located on the solvent-exposed surface of the protein. Any predicted RNA-binding residues that are buried in the three-dimensional structure of a protein should be viewed with suspicion, although buried interfacial residues in "unbound" protein structures can become exposed due to conformational changes in the protein that occur upon RNA binding [28, 53–55].

Another way in which structural information can be exploited to accurately identify potential RNA-binding residues is illustrated in the so-called "homology-based" approaches. Homology-based approaches take advantage of the observation that RNA-binding residues are often conserved across homologous proteins [56, 57]. Thus, if a "bound" structure is available for a close sequence homolog of the query protein, the RNA-binding residues of the query protein can be inferred, based on their alignment with the known RNA-binding residues in the homologous sequence. When applicable, homology-based approaches provide the most reliable computational predictions of RNA-binding sites, but they have an important limitation: if no homologs with experimentally determined bound structures are available for the query protein, no predictions can be generated. This limitation can be overcome by combining a homology-based method, with a machine learning-based method, which can return predictions for every residue in any protein. This is the strategy employed by **RNABindRPlus** (*see* Subheading 3.2), which combines a PSSM-based Support Vector Machine (SVM) with a homology-based method to generate highly reliable predictions [57], and by **SNBRFinder** (*see* Subheading 3.3), which combines an SVM classifier that uses sequence profiles, residue conservation scores, physicochemical properties and interface propensities, with a homology-based method that uses profile hidden Markov models (HMMs) to search for the homologs [58].

The major goal of the chapter is to provide a step-by-step protocol for predicting RNA-binding residues in proteins, with a focus on machine learning and homology-based methods. In keeping with the theme of this volume, the methods outlined here are sequence-based; they do not require structural information regarding the protein of interest. We also provide a brief guide to accessing and utilizing state-of-the-art computational methods, web servers and databases that provide information about interfaces in protein–RNA complexes and/or predictions of RNA-binding residues in proteins. For additional information, the reader is referred to two excellent reviews: a recent review by Si et al. [50], which includes a comprehensive table of available sequence, structure and docking based methods; and a review by Tuszynska et al. [59], which focuses on structural docking-based approaches which are not considered here.

In this chapter, we focus on currently available web-based computational tools for interface prediction, i.e., predicting which specific amino acid residues in an RNA-binding protein are involved in recognition of and binding to RNA. A few tools are also capable of predicting the converse, i.e., which ribonucleotides in the bound RNA directly contact the protein of interest (e.g., [52, 60, 61)]. Software and servers for partner prediction, i.e., predicting which RNA(s) bind to a specific protein of interest (or *vice versa*) in a protein–RNA complex or a protein–RNA interaction network, are not described here, but have been reviewed elsewhere [62–65]. Tools for predicting whether or not a query protein is likely to bind RNA are also available (e.g., Tartaglia [39, 66, 67)]. but are not considered here.

The protocol involves two major steps (illustrated in Fig. 1):

Step 1: Determine whether experimental data regarding RNA-binding residues in the query RNA-binding protein (or putative RNA-binding protein) are already available. This step is described in Subheading 3.1, which outlines strategies for exploiting available online databases and servers (provided in Table 1 below) that provide structural data regarding protein–RNA complexes, or focus on RNA-binding proteins, RNA-binding motifs, or protein–RNA interactions.

Step 2: If known RNA-binding residues cannot be identified using available resources, or if the user wishes to identify additional potential interfacial residues, use one (or, preferably, all three) of the following web-based tools for predicting RNA-binding residues in protein–RNA complexes:

- **RNABindRPlus** (*see* Subheading 3.2)—a hybrid machine learning/sequence homology-based approach developed by our group [57] which requires only sequence information for the protein(s) of interest. The accuracy of this and similar sequence-based methods from other groups is generally greater than that obtained using structure-based methods.

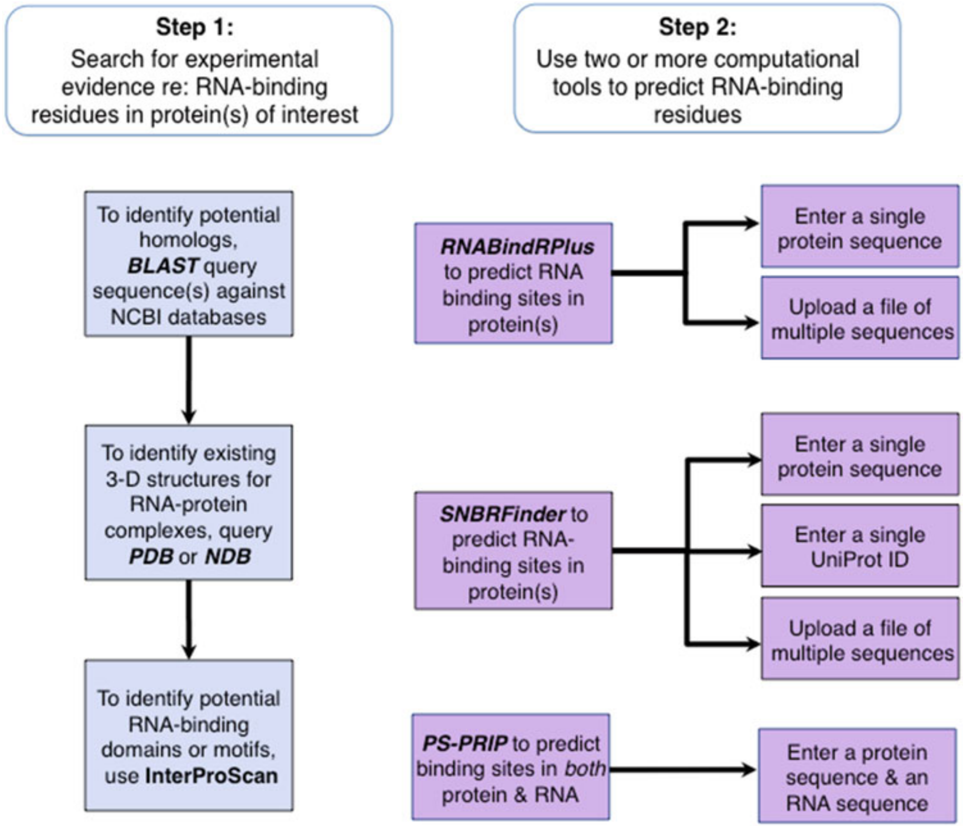

Step 1:
Search for experimental evidence re: RNA-binding residues in protein(s) of interest
Step 2:
Use two or more computational tools to predict RNA-binding residues
To identify potential homologs, BLAST query sequence(s) against NCBI databases
To identify existing 3-D structures for RNA-protein complexes, query PDB or NDB
To identify potential RNA-binding domains or motifs, use InterProScan
RNABindRPlus to predict RNA binding sites in protein(s)
Enter a single protein sequence
Upload a file of multiple sequences
SNBRFinder to predict RNA-binding sites in protein(s)
Enter a single protein sequence
Enter a single UniProt ID
Upload a file of multiple sequences
PS-PRIP to predict binding sites in both protein & RNA
Enter a protein sequence & an RNA sequence

Fig. 1 Flowchart for identifying potential RNA-binding residues in proteins

- **SNBRFinder** (*see* Subheading 3.3)—a method developed by Yang et al. [58], which can predict either RNA- or DNA-binding residues in proteins by combining a machine learning method with a template (homology)-based method. The key differences between SNBRFinder and RNABindRPlus are: (a) inputs to the SVM classifier in SNBRFinder include sequence profiles and other sequence descriptors such as residue conservation scores, physicochemical properties, and interface propensities, whereas the only inputs to the SVM for RNABindRPlus are sequence PSSMs; (b) SNBRFinder uses profile hidden Markov models to find remote homologs for the query protein, whereas RNABindRPlus uses BLAST searches.
- **PS-PRIP** (*see* Subheading 3.4)—a new motif-based method developed by our group [52], which can predict interfacial residues in both the protein and the RNA components of a protein–RNA complex and can provide "partner-specific" predictions.

Table 1
Databases of protein–RNA complexes and resources for analyzing interfaces and motifs in protein–RNA complexes

Database	Description	Reference
Databases of structures of RNA–protein complexes		
PDB (Protein Data Bank)	www.pdb.org This is a database of 3D macromolecular structures—protein–protein, protein–DNA, protein–RNA, and protein–ligand structures solved using X-ray crystallography, cryo-EM, NMR, and others	[68]
NDB (Nucleic Acid Database)	http://ndbserver.rutgers.edu/ This is a database of three-dimensional structural information for nucleic acids	[69]
PDBSum	https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html A pictorial database of PDB structures that provides access to interfacial residues in known structures	[70, 71]
Resources for analyzing interfaces and RNA-binding motifs in RNA		
BIPA (Biological Interaction Database for Protein–Nucleic Acid)	http://mordred.bioc.cam.ac.uk/bipa BIPA provides a list of protein–RNA (and protein–DNA) complexes from the PDB and displays RNA binding residues within the linear primary sequence of a chosen protein, or within a multiple sequence alignment of related RNA binding proteins (BIPA has not been updated since 2009 and is not fully functional at present)	[72]
InterPro & InterProScan	http://www.ebi.ac.uk/interpro/ InterPro classifies protein sequences into families using information from ten different databases; InterProScan identifies functional and/or conserved domains, motifs, and other important sites in protein sequences	[51, 73]
NPIDB (Nucleic Acid-Protein Interaction Database)	http://npidb.belozersky.msu.ru/ A database for extracting biologically meaningful characteristics of protein–RNA and protein–DNA complexes	[74]
DBBP (DataBase of Binding Pairs in protein–nucleic acid interactions)	http://bclab.inha.ac.kr/dbbp A database that provides structural data for hydrogen bonding interactions between proteins and nucleic acids	[75]
PRIDB (Protein RNA interface database)	http://pridb.gdcb.iastate.edu A database of protein–RNA complexes from the PDB, with tools for identifying and visualizing interfacial residues in both the protein and RNA sequences and structures. (PRIDB has not been updated since 2013 and is under remediation)	[76]

(continued)

Table 1 (continued)

Database	Description	Reference
RsiteDB	http://bioinfo3d.cs.tau.ac.il/RsiteDB/ This database stores information about the protein binding pockets that interact with single-stranded RNA nucleotide bases	[77]
ProNIT	http://www.abren.net/pronit/ A database of thermodynamic interaction data (binding constants, free energy change, and so on) between proteins and nucleic acids	[78]
RNA CoSSMos	http://cossmos.slu.edu/ A tool that provides information on secondary structural motifs such as bulges and hairpin loops of 3D protein–nucleic acid structures	[79]
RNA 3D Hub	http://rna.bgsu.edu/rna3dhub/ A suite of tools including the RNA Structure Atlas and RNA 3D Motif Atlas. These provide information about RNA 3D motifs	[80]
RNA Bricks	http://iimcb.genesilico.pl/rnabricks A database that provides information about recurrent RNA 3D motifs and their interactions, extracted from experimentally determined structures of RNA and RNA-protein complexes	[81]
Databases of recognition sites/protein-binding motifs in RNA		
CISBP-RNA	http://cisbp-rna.ccbr.utoronto.ca/ A database of inferred sequence binding preferences of RNA-binding proteins	[22]
RBPDB	http://rbpdb.ccbr.utoronto.ca/ A database of manually curated RNA-binding sites collected from literature	[41]

We encourage users to submit their proteins of interest to all three web servers described in this protocol because the underlying algorithms and datasets used for training and evaluating performance are different in each case, and the methods have different strengths and weaknesses. Even though all three methods have been shown to provide highly reliable predictions on benchmark datasets, it is not possible to guarantee an accurate prediction for any specific RNA-binding protein with any of these methods.

2 Materials

2.1 Databases of Experimentally Validated Protein–RNA Complexes and Resources for Analyzing Interfaces

Before using computational methods to *predict* RNA-binding residues, the user should first search for existing experimental data regarding interfacial residues in the specific RNA-binding protein(s) of interest, both in published literature and in relevant specialized databases. The "gold standard" for identifying RNA-binding residues in proteins is analysis of a high resolution three-dimensional structure of the protein bound to its cognate RNA, i.e., a "bound" structure of the complex containing the protein bound to RNA. The Protein Data Bank (PDB) [68] and the Nucleic Acid Database (NDB) [69] are two comprehensive databases of experimentally determined structures, from which residue and atomic-level information regarding the interfaces in macromolecular complexes can be extracted. Table 1 provides URLs for these two primary databases, followed by an alphabetical listing of several databases that contain valuable information about protein–RNA complexes and their interfacial residues, either derived from structures in the PDB/NDB or from other types of experiments. A suggested strategy for utilizing selected resources from this list is provided in Subheading 3.1 below.

2.2 Servers and Software for Predicting Interfaces in Protein–RNA Complexes

There are more than 20 published approaches for predicting RNA-binding residues in proteins (for a recent compilation, see [50]), and a few methods are capable of predicting interfacial residues in both the protein and the RNA components of a protein–RNA complex (e.g., [52, 82]). Subheadings 3.2–3.5 below focus on three methods (RNABindRPlus, SNBRFinder, PS-PRIP) that are freely available on web-based servers and have been shown to perform well on benchmark datasets. Table 2 lists these and several additional methods. Please note that not all of these are currently available as web-based servers.

2.3 The RNABindRPlus Server

RNABindRPlus [57] is a purely sequence-based method for predicting RNA-binding residues in putative RNA-binding proteins. It uses logistic regression to combine predictions from HomPRIP, a sequence homology-based method, with predictions from SVMOpt, an optimized Support Vector Machine (SVM) classifier. The SVM classifier utilizes sequence-based PSSMs as features. HomPRIP makes highly accurate predictions of RNA-binding residues when homologs (with solved structures) of the query protein can be found, but a major drawback is that no predictions are returned when no such homologs can be found. Additionally, HomPRIP cannot return predictions for parts of the query protein sequence that are not aligned with its homologs. This limitation of HomPRIP is overcome by combining it with a machine learning-based method, SVMOpt, which returns predictions for every residue in any protein sequence.

Table 2
Servers and software for predicting RNA-binding sites in proteins

Method	Description	Reference
BindN	http://bioinfo.ggc.org/bindn/ An SVM classifier that uses hydrophobicity, side chain pKa, molecular mass, and PSSMs for predicting RNA-binding residues; it can also predict DNA-binding residues	[83]
BindN+	http://bioinfo.gcc.org/bindn An updated version of BindN, that uses an SVM classifier based on PSSMs and several other descriptors of evolutionary information; it can also predict DNA-binding residues	[84]
catRAPID signature	http://s.tartaglialab.com/grant_submission/signature Predicts both RNA-binding and protein-binding residues in RNPs based on physicochemical features instead of sequence similarity searches	[82]
DR_bind1	http://drbind.limlab.ibms.sinica.edu.tw/ Predicts RNA-binding residues in proteins using information derived from 3D structure	
DRNA	http://sparks-lab.org/yueyang/DFIRE/dRNA-DB-service.php Predicts RNA-binding proteins and RNA-binding sites based on similarity to known structures	[85]
KYG	http://cib.cf.ocha.ac.jp/KYG Uses a set of scores based on the RNA-binding propensity of individual and pairs of surface residues of the protein, used alone or in combination with position-specific multiple sequence profiles	[86]
Meta predictor	http://iimcb.genesilico.pl/meta2/ A predictor that combines the output of PiRaNhA, PPRInt, and BindN+ to make predictions of RNA-binding residues using a weighted mean. (Not available as of March 2014)	[31]
NAPS	http://prediction.bioengr.uci.edu A modified C4.5 decision tree algorithm that uses amino acid identity, residue charge, and PSSMs to predict residues involved in DNA- or RNA-binding. (Not available as of March 2014)	[87]
OPRA	Uses path energy scores calculated using interface propensity scores weighted by the accessible surface area of a residue to predict RNA-binding sites. Available from the authors upon request	[88]
PPRInt	http://www.imtech.res.in/raghava/pprint/ An SVM classifier trained on PSSM profiles to predict RNA-binding residues	[89]
PS-PRIP	http://pridb.gdcb.iastate.edu/PSPRIP/ A partner-specific method for predicting RNA-binding residues in proteins and protein-binding residues in RNAs using sequence motifs extracted from interfacial regions in RNA-protein complexes	[52]
PRBR	http://www.cbi.seu.edu.cn/PRBR/ An enriched random forest classifier trained on predicted secondary structure, a combination of PSSMs with physic-chemical properties, a polarity-charge correlation, and a hydrophobicity correlation	[90]

(continued)

Table 2
(continued)

Method	Description	Reference
PRIP	Uses an SVM classifier and a combination of PSSM profiles, solvent accessible surface area, betweenness centrality, and retention coefficient as input features. Not accessible via a web server, but results can be obtained via correspondence with the author	[91]
RBScore	http://ahsoka.u-strasbg.fr/rbscore/ Utilizes a score derived from physicochemical and evolutionary features, integrating a residue neighboring network approach; it predicts both DNA- and RNA-binding residues in proteins	[92]
RISP	http://grc.seu.edu.cn/RISP An SVM-based method that uses evolutionary information in terms of PSSMs (Not available as of March 2014)	[93]
RNABindR	http://bindr.gdcb.iastate.edu/RNABindR/ A Naïve Bayes classifier that uses the amino acid sequence identity to predict RNA-binding residues in proteins (no longer maintained)	[94]
RNABindR v2.0	http://ailab1.ist.psu.edu/RNABindR/ An SVM classifier that uses sequence PSSMs to predict RNA-binding residues in proteins	[47]
RNABindRPlus	http://ailab1.ist.psu.edu/RNABindRPlus/ A predictor that combines an optimized SVM classifier with a sequence homology-based method to predict RNA-binding residues in proteins	[57]
RNApin	http://www.imtech.res.in/raghava/rnapin/ An SVM classifier that predicts protein-interacting nucleotides (PINs) in RNA	[61]
SNBRFinder	http://ibi.hzau.edu.cn/SNBRFinder/ A sequence-based hybrid predictor that combines a feature-based predictor and a template-based predictor to predict nucleic-acid binding residues in proteins	[95]
SPOT-Seq-RNA	http://sparks-lab.org/yueyang/server/SPOT-Seq-RNA/ A template-based technique for predicting RBPs, RNA-binding residues and complex structures	[95]

RNABindRPlus was trained on the RB198 dataset, and tested on two different datasets, RB44 and RB111. On a subset of proteins for which homologs with experimentally determined interfaces could be reliably identified, HomPRIP outperformed all other methods, achieving an MCC of 0.63 on RB44 and 0.83 on RB111. RNABindRPlus was able to predict RNA-binding residues of all proteins in both test sets, achieving an MCC of 0.55 on RB44 and 0.37 on RB111, and outperforming all other methods, including structure-based methods (e.g., KYG [86]).

2.4 The SNBRFinder Server

SNBRFinder is a sequence-based predictor that combines predictions from a Support Vector Machine (SVM) classifier, SNBRFinder^{F}, with predictions from a template-based classifier, SNBRFinder^{T}.

SNBRFinder^{F} utilizes a sliding window of the target residues and five neighboring residues on each side to represent the sequential environment. The features used as inputs to the classifier include the sequence profile, residue conservation scores, predicted structural features, physicochemical properties, interface propensity, sequential position, and two global features, sequence length and the global amino acid composition.

SNBRFinder^{T} is a template-based method, i.e., a method that utilizes sequence or structural alignments to retrieve homologs/templates of a query protein and then infer binding residue information for the query protein. SNBRFinder^{T} uses the HHblits program [96] to identify homologs of the query protein. HHblits represents both the query and database sequences using profile hidden Markov models (HMM), and then compares the two to identify homologs of the query protein. For each query and homolog pair, a probability score is output for evaluating the similarity between the aligned HMMs. The higher the score is, the better the alignment is and vice versa. Specifically, a residue in the query protein is predicted to be RNA-binding with a probability score of 1 if it is matched with a binding residue in the homolog, otherwise the residue is predicted to be non RNA-binding with a probability score of 0.

On the RB44 [31] dataset, SNBRFinder had an MCC of 0.48, whereas RNABindRPlus had an MCC of 0.49. In terms of AUC values, SNBRFinder and RNABindRPlus achieved very similar results, with both getting 0.84.

2.5 The PS-PRIP Server

PS-PRIP [52] is a motif-based method that predicts interfacial residues for both the RNA and protein components of protein–RNA complexes in a partner-specific manner (*see* **Note 4**). PS-PRIP requires as input the sequences of both the RNA-binding protein and its putative bound RNA(s). Although no structural information is required, PS-PRIP exploits the co-occurrence of specific pairs of short protein and RNA sequence motifs (5 amino acids long and 5 ribonucleotides long) from a database of motifs extracted from interfaces in known protein–RNA complexes from the PDB. On an independent dataset of 327 RNA-protein pairs, PS-PRIP obtained a sensitivity of 0.64, precision of 0.80, and MCC of 0.59 compared to RNABindRPlus with values of 0.88, 0.76, and 0.71, respectively. In addition to providing predicted RNA-binding residues in proteins, PS-PRIP makes predictions of protein-binding residues in RNAs, although with much lower accuracy. Other methods designed to predict protein-binding residues in RNA have been published recently (e.g., [61, 82]).

3 Methods

3.1 Searching Existing Literature and Databases for Relevant Experimental Data

Currently, all computational methods for predicting RNA-binding residues in proteins return only *predicted* interfacial residues, even when the actual interfaces are known from experimental data. Thus, before using software to predict potential RNA-binding residues, the user should search published literature and existing databases for experimentally identified interactions involving the protein of interest (*see* **Note 5**). If the query protein is newly discovered or has no known function, the user should first search for potential homologs using a BLAST search. As outlined below, both the original query sequence and its homologs can be used to search databases of known protein–RNA interactions, such as those listed in Table 1.

1. If the query protein sequence corresponds to an "unknown" or novel protein, run the sequence through **NCBI's BLAST server**, available at http://blast.ncbi.nlm.nih.gov/Blast.cgi [97, 98] or use similar genomics resources elsewhere (e.g., http://www.ebi.ac.uk/Tools/sss/). BLAST (Basic Local Alignment Search Tool) finds highly similar sequences in the NCBI or ENSEMBL databases (*see* **Note 6**). A good starting point for most protein sequence searches is SMARTBLAST, available here: http://blast.ncbi.nlm.nih.gov/smartblast/ (*see* **Note 7**). If the query sequence itself is not available in one of the NCBI or ENSEMBL databases, potential homologs identified by BLAST can be used as the query for subsequent searches in the databases listed in **steps** 2–**6** below (*see* **Note 8**).
2. If the query protein has been previously identified and/or analyzed, a search using the **NCBI "Protein"** tool may quickly reveal previously annotated RNA-binding domains or motifs and links to experimentally determined structures. Enter the name of the protein (or name of a potential homolog, identified in **step 1**) into the box provided here: (http://www.ncbi.nlm.nih.gov/protein/). In the list of "Items" returned, click on the protein name from the appropriate organism to access the full GenBank protein entry. Then, examine information on the right side of the GenBank protein page; for example, if a high resolution structure is available, it will appear under the "Protein 3D Structure" header. Under the "Related Information" header, click on "Conserved Domains (Concise)" or "Conserved Domains (Full)" to access any annotated RNA-binding domains (or other conserved domains) identified in the protein sequence. The "Conserved Domains" results page also provides links to available three-dimensional structure(s) similar to that of the query protein, if available. Other links on this page can lead to additional information regarding potential RNA-binding domains in the protein of interest (*see* **Note 9**).

3. In every case, the user should search the **Protein Data Bank (PDB),** available at www.rcsb.org [68] for any available structures of protein–RNA complexes that contain the protein of interest. The PDB contains over 1600 three-dimensional structures of protein–RNA complexes determined using experiments such as X-ray crystallography, nuclear magnetic resonance (NMR) imaging, and cryo-electron microscopy. The PDB has a powerful search engine that allows the database to be queried in a variety of ways, e.g., by protein (or RNA) name, sequence, or GO terms. The PDB also provides excellent structure visualization tools as well as links to valuable third-party resources for visualizing and analyzing the structures of macromolecules (*see* **Note 10**).

4. Similarly, the **Nucleic Acid Database (NDB**), available at http://ndbserver.rutgers.edu [69], is another valuable resource that focuses on experimentally determined three-dimensional structures of nucleic acids, including both protein–RNA and protein–DNA complexes. The NDB contains only a subset of structures in the PDB, making it easier for the user to focus on structures that contain RNA. Also, the NDB provides convenient access to a wide variety of tools and software specifically designed for analyzing RNA sequences and structures (*see* **Note 11**).

5. If it is possible to identify a structure for the query protein–RNA complex (or a homologous complex) in one of the previous steps, the user can quickly obtain a graphical representation of the protein–RNA interface, using **PDBSum** [70, 71] available at: https://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html. Enter the 4-letter PDB code in the box provided and click "Find." At the top of the PDBSum entry page that appears, click on the "DNA/RNA" link to access a page listing all of the nucleic acid chains in the complex. Then click on "**NUCPLOT**" to visualize the ribonucleotides that are contacted by individual amino acids, as well as additional information (backbone *vs.* phosphate group contacts, hydrogen bonding, etc.). Another way to identify the RNA-binding amino acids is to click on the "Protein" link at the top of the page to reveal a diagrammatic representation of the protein sequence, in which Residue Contacts to DNA/RNA are labeled. Tools for visualizing, analyzing and manipulating the structure are provided by both the PDB and NDB (*see* **Notes 10** and **11**). *See* Table 1 for additional tools that provide detailed information about the interfacial residues (e.g., NPIDB [74], DBBP [75]).

6. If no structure for the query protein–RNA complex can be identified, the user can **search for known RNA-binding domains or motifs in the protein sequence.** Typically, only a few of the amino acids in well-characterized RNA-binding

domains or motifs (e.g., the RNA Recognition Motif (RMM), which is ~90 amino acids) are actually "interfacial residues" involved in contacting RNA (*see* **Note 1**). But, if the query protein does contain such a conserved domain or motif, homologous structures are likely available and can indicate which amino acids are directly involved in recognizing and binding RNA. The EMBL-EBI's **InterPro** [51] is a valuable comprehensive resource that includes more than 10 databases of protein structural and functional motifs, and an integrated tool, **InterProScan** [73], which can be used to identify all known motifs, including RNA-binding motifs, in a protein of interest. Access InterPro here: http://www.ebi.ac.uk/interpro/ and enter the query protein sequence in the text box. Within a few minutes, a "Results" page will appear, providing a graphical summary of all domains, motifs and signatures identified, with links to additional information about each.

7. For many RNA-binding proteins, recognition motifs (i.e., the specific RNA sequences bound by the RBP) are now known [1, 22, 99]. Several valuable databases and tools are available if the user wishes to identify known or potential recognition sites in the RNA component of a specific protein–RNA complex. Databases of experimentally defined RNA sequence motifs that are bound by RBPs include: CISBP-RNA [22], RBPDB, [41], and RBPMotif [43]. Databases of RNA structural motifs, e.g., BRICKS [81] and the RNA 3D Motif Atlas [80], are also available, but these have not yet been systematically annotated regarding their protein-binding activities. Also, a valuable tool for mapping binding sites for RBPs within the genomes of several model organisms is RBPMap [100], which is available at: http://rbpmap.technion.ac.il.

3.2 Using RNABindRPlus to Predict RNA-Binding Residues in Proteins

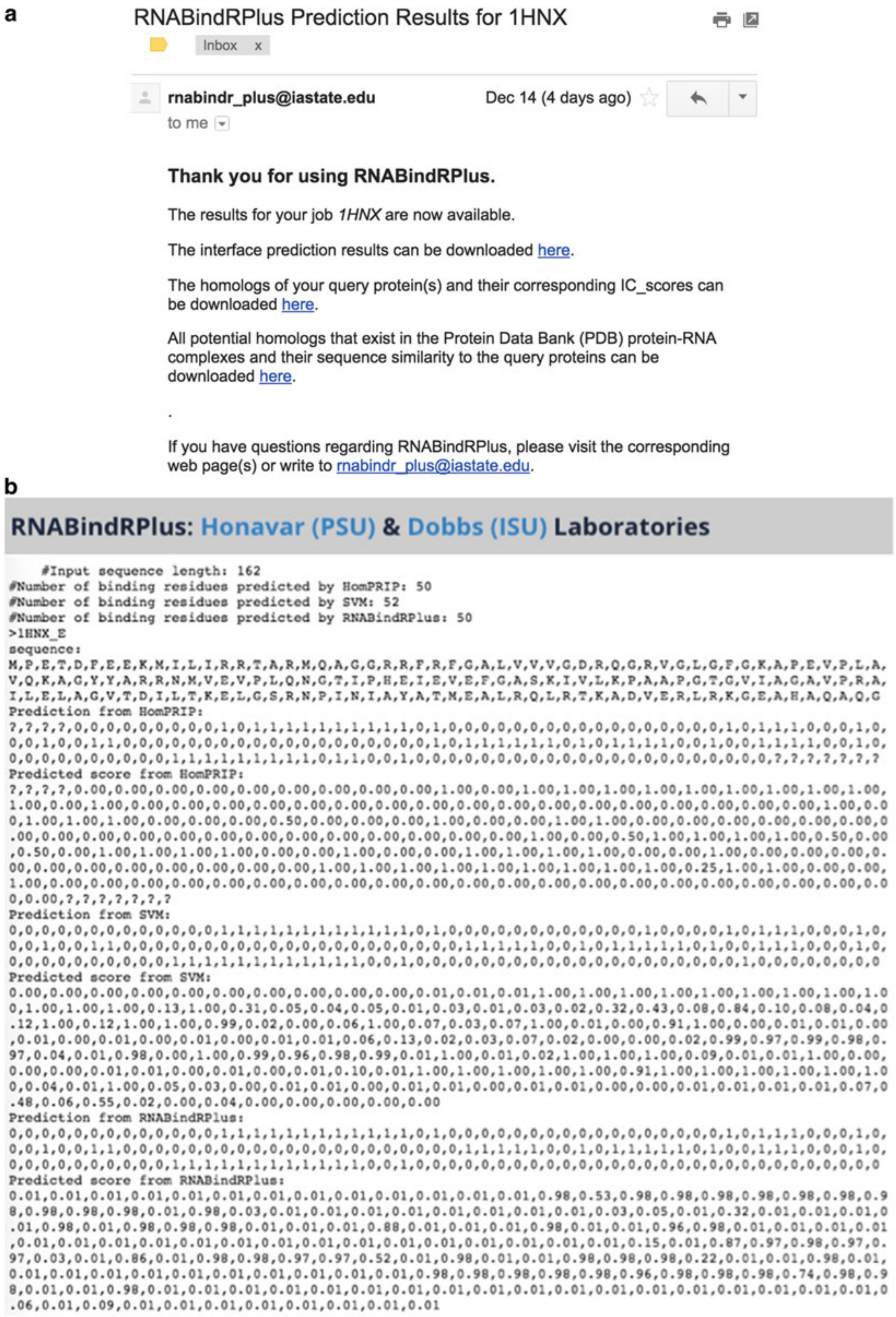
The **RNABindRPlus** method implements a combination of a machine learning method (**SVMOpt**) and a sequence homology-based method (**HomPRIP**) to predict RNA-binding residues in proteins [57] (*see* Subheading 2.3). Given a single protein sequence (or a file of multiple protein sequences), RNABindRPlus can predict which amino acid residues are mostly likely to bind RNA. Run times can be slow when large numbers of protein sequences are submitted in a single job (*see* **Note 12**). A faster version of the server is under development (*see* **Note 13**).

1. Access the **RNABindRPlus** web server at: http://ailab1.ist.psu.edu/RNABindRPlus/.
2. **To predict RNA-binding residues in a single putative RNA-binding protein**: Enter the protein sequence in FASTA format (*see* **Note 14**) in the text box provided on the homepage.

3. **To predict RNA-binding residues for multiple putative RNA-binding proteins**: In this case, the user has two options: (a) Enter the protein sequences in FASTA format in the text box provided; or (b) upload a FASTA formatted file of protein sequences by clicking the "Choose file" button on the homepage.
4. Provide an email address where results should be sent. Computing the results requires approximately 10 min per protein sequence submitted to RNABindRPlus (*see* **Notes 12** and **13**).
5. The user has the option of excluding highly similar proteins from the homolog list, at the desired sequence identity level by selecting the check box at the bottom of the submission page. To obtain the most reliable predictions, leave this option blank (*see* **Note 15**).
6. Once all submission fields have been filled, click on the "Submit" button. The user will receive an email confirming that the job is currently running. RNABindRPlus results will be returned to the user by email.
7. Figure 2 shows results returned by RNABindRPlus for the S5 protein from the 30S ribosomal subunit of *T. thermophilus*, which corresponds to protein chain E, in PDB structure 1HNX). Figure 2a shows the *Results Summary* email, which contains several links that can be clicked to display selected portions of the results. Figure 2b (*Interface Prediction Results)* displays predictions from three different methods: HomPRIP (homology-based method), SVMOpt (optimized SVM) and RNABindRPlus (which combines predictions from HomPRIP and SVMOpt). The first section of output for each method (e.g., Prediction from HomPRIP), is a list of the predictions for each residue, where "1" corresponds to predicted interfacial residues (i.e., RNA-binding) and "0" corresponds to predicted non-interfacial residues. The second section of output (e.g., "Predicted score from HomPRIP") gives the probability score for each residue (where a probability of ≥0.5 means the residue is an interface residue, otherwise it is a non-interface residue). Figure 2c (*Homologs of the query protein*) displays a list of homologous proteins identified by HomPRIP, the homology-based component of RNABindRPlus, along with their corresponding interface conservation scores (IC_scores) (*see* **Note 16**). These are the homologous proteins used for inferring RNA-binding residues in the query protein using HomPRIP. Figure 2d (*All potential homologs in the PDB*) shows only a portion of the output providing information about all potential homologs found in the PDB for the query protein.

3.3 Using SNBRFinder to Predict RNA-Binding Residues in Proteins

SNBRFinder is a sequence-based hybrid predictor that combines predictions from a Support Vector Machine method, SNBRFinder^F, with predictions from a template-based method, SNBRFinder^T [58] (*see* Subheading 2.4). The inputs to the SVM method include



a

RNABindRPlus Prediction Results for 1HNX

rnabindr_plus@iastate.edu Dec 14 (4 days ago)

to me

Thank you for using RNABindRPlus.

The results for your job *1HNX* are now available.

The interface prediction results can be downloaded here.

The homologs of your query protein(s) and their corresponding IC_scores can be downloaded here.

All potential homologs that exist in the Protein Data Bank (PDB) protein-RNA complexes and their sequence similarity to the query proteins can be downloaded here.

.

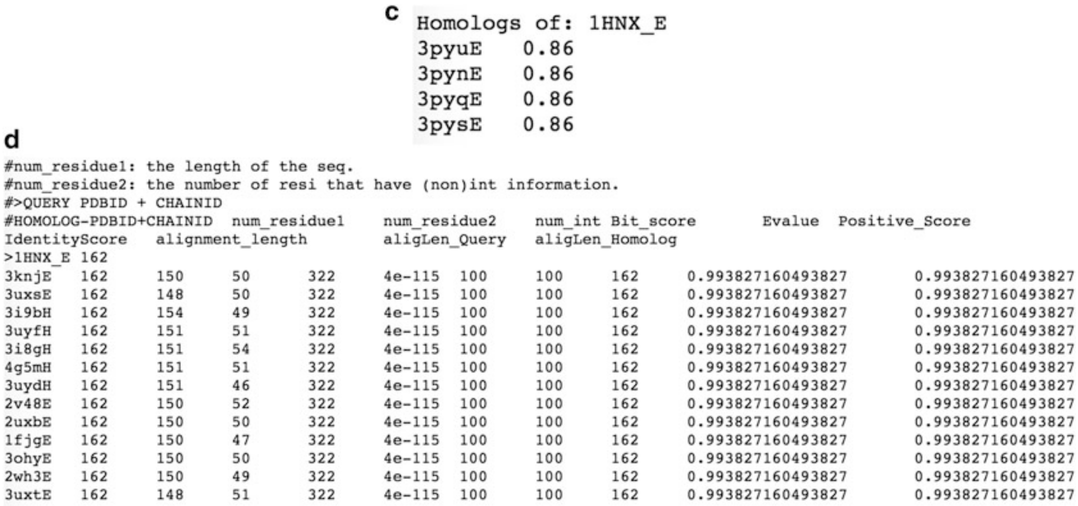
If you have questions regarding RNABindRPlus, please visit the corresponding web page(s) or write to rnabindr_plus@iastate.edu.

b

RNABindRPlus: Honavar (PSU) & Dobbs (ISU) Laboratories

```
    #Input sequence length: 162
#Number of binding residues predicted by HomPRIP: 50
#Number of binding residues predicted by SVM: 52
#Number of binding residues predicted by RNABindRPlus: 50
>1HNX_E
sequence:
M,P,E,T,D,F,E,E,K,M,I,L,I,R,R,T,A,R,M,Q,A,G,G,R,R,F,R,F,G,A,L,V,V,V,G,D,R,Q,G,R,V,G,L,G,F,G,K,A,P,E,V,P,L,A,
V,Q,K,A,G,Y,Y,A,R,R,N,M,V,E,V,P,L,Q,N,G,T,I,P,H,E,I,E,V,E,F,G,A,S,K,I,V,L,K,P,A,A,P,G,T,G,V,I,A,G,A,V,P,R,A,
I,L,E,L,A,G,V,T,D,I,L,T,K,E,L,G,S,R,N,P,I,N,I,A,Y,A,T,M,E,A,L,R,Q,L,R,T,K,A,D,V,E,R,L,R,K,G,E,A,H,A,Q,A,Q,G
Prediction from HomPRIP:
?,?,?,?,0,0,0,0,0,0,0,0,0,1,0,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,1,0,0,0,1,0,
0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,1,1,1,1,0,1,0,1,1,1,1,0,0,1,0,0,1,1,1,1,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,0,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,?,?,?,?,?,?,?
Predicted score from HomPRIP:
?,?,?,?,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,1.00,0.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,
1.00,0.00,1.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,1.00,0.0
0,1.00,1.00,1.00,0.00,0.00,0.00,0.50,0.00,0.00,0.00,1.00,0.00,0.00,1.00,1.00,0.00,0.00,0.00,0.00,0.00,0.00,0
.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,1.00,0.00,0.50,1.00,1.00,1.00,1.00,0.50,0.00
,0.50,0.00,1.00,1.00,1.00,1.00,0.00,0.00,1.00,0.00,0.00,1.00,1.00,1.00,1.00,0.00,0.00,1.00,0.00,0.00,0.00,0.
00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,0.25,1.00,1.00,0.00,0.00,
1.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.0
0,0.00,?,?,?,?,?,?,?
Prediction from SVM:
0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,1,0,1,1,1,0,0,0,1,0,
0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,1,0,1,1,1,1,1,0,1,0,0,1,1,1,0,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0,0
Predicted score from SVM:
0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.00,0.01,0.01,0.01,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.00,1.0
0,1.00,1.00,1.00,0.13,1.00,0.31,0.05,0.04,0.05,0.01,0.03,0.01,0.03,0.02,0.32,0.43,0.08,0.84,0.10,0.08,0.04,0
.12,1.00,0.12,1.00,1.00,0.99,0.02,0.00,0.06,1.00,0.07,0.03,0.07,1.00,0.01,0.00,0.91,1.00,0.00,0.01,0.01,0.00
,0.01,0.00,0.01,0.00,0.01,0.00,0.01,0.01,0.06,0.13,0.02,0.03,0.07,0.02,0.00,0.00,0.02,0.99,0.97,0.99,0.98,0.
97,0.04,0.01,0.98,0.00,1.00,0.99,0.96,0.98,0.99,0.01,1.00,0.01,0.02,1.00,1.00,1.00,0.09,0.01,0.01,1.00,0.00,
0.00,0.00,0.01,0.01,0.00,0.01,0.00,0.01,0.10,0.01,1.00,1.00,1.00,1.00,1.00,0.91,1.00,1.00,1.00,1.00,1.00,1.0
0,0.04,0.01,1.00,0.05,0.03,0.00,0.01,0.01,0.00,0.01,0.01,0.00,0.01,0.01,0.00,0.00,0.01,0.01,0.01,0.01,0.07,0
.48,0.06,0.55,0.02,0.00,0.04,0.00,0.00,0.00,0.00,0.00
Prediction from RNABindRPlus:
0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,1,0,0,0,1,0,
0,0,1,0,0,1,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,0,0,1,0,1,1,1,1,1,0,1,0,0,1,1,1,0,0,0,1,0,
0,0,0,0,0,0,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0,0
Predicted score from RNABindRPlus:
0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.98,0.53,0.98,0.98,0.98,0.98,0.98,0.98,0.9
8,0.98,0.98,0.98,0.01,0.98,0.03,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.03,0.05,0.01,0.32,0.01,0.01,0.01,0
.01,0.98,0.01,0.98,0.98,0.98,0.01,0.01,0.01,0.88,0.01,0.01,0.01,0.98,0.01,0.01,0.96,0.98,0.01,0.01,0.01,0.01
,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.15,0.01,0.87,0.97,0.98,0.97,0.
97,0.03,0.01,0.86,0.01,0.98,0.98,0.97,0.97,0.52,0.01,0.98,0.01,0.01,0.98,0.98,0.98,0.22,0.01,0.01,0.98,0.01,
0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.98,0.98,0.98,0.98,0.98,0.96,0.98,0.98,0.98,0.74,0.98,0.9
8,0.01,0.01,0.98,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0
.06,0.01,0.09,0.01,0.01,0.01,0.01,0.01,0.01,0.01,0.01
```

Fig. 2 (**a**) RNABindRPlus results notification email obtained for the *T. thermophilus* S5 protein. (**b**) RNABindRPlus prediction results for the *T. thermophilus* S5 protein. Results are also returned for the two individual components of RNABindRPlus, HomPRIP and SVMOpt. For each method, under the header "Prediction from," the predicted RNA-binding residues are represented by a string of 1's and 0's, where "1" and "0" correspond to predicted RNA-binding and non-RNA binding residues, respectively. See text for additional details.



```
Homologs of: 1HNX_E
3pyuE    0.86
3pynE    0.86
3pyqE    0.86
3pysE    0.86
```

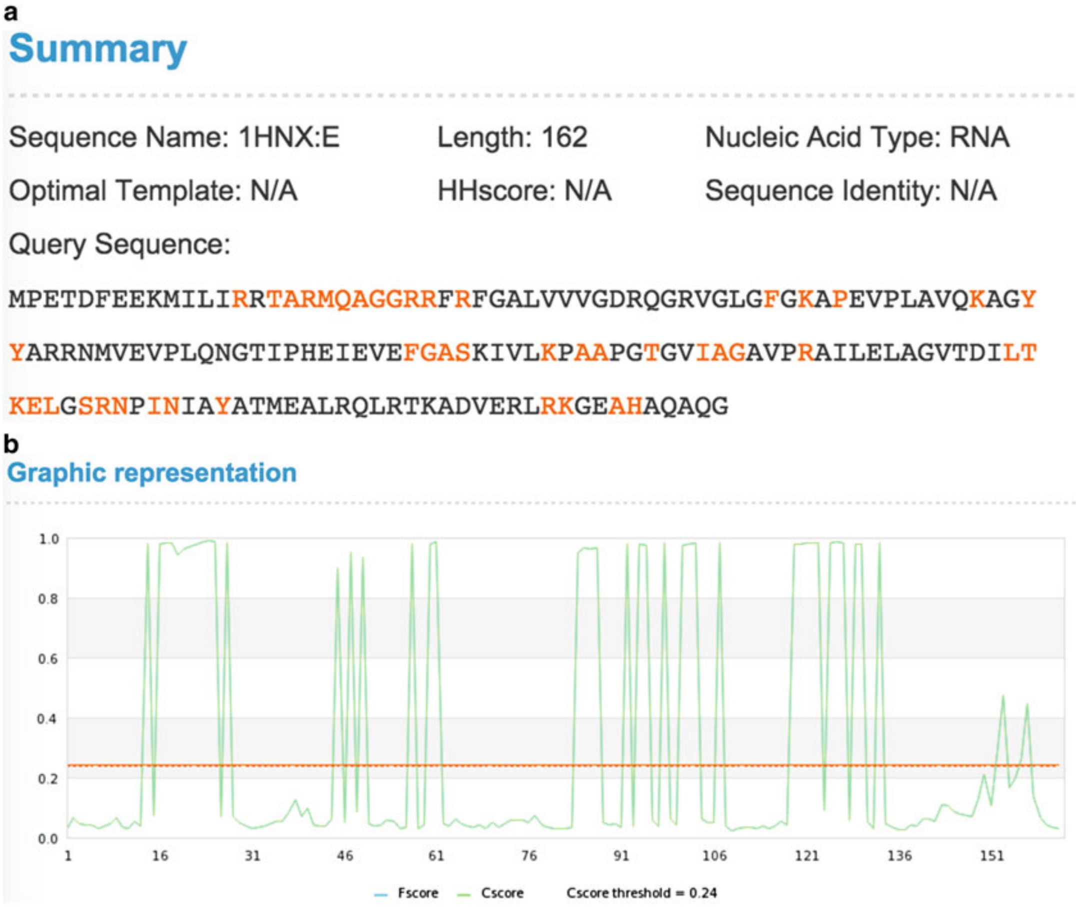
```
#num_residue1: the length of the seq.
#num_residue2: the number of resi that have (non)int information.
#>QUERY PDBID + CHAINID
#HOMOLOG-PDBID+CHAINID  num_residue1      num_residue2      num_int Bit_score       Evalue  Positive_Score
IdentityScore    alignment_length         aligLen_Query     aligLen_Homolog
>1HNX_E 162
3knjE    162      150      50       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3uxsE    162      148      50       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3i9bH    162      154      49       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3uyfH    162      151      51       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3i8gH    162      151      54       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
4g5mH    162      151      51       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3uydH    162      151      46       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
2v48E    162      150      52       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
2uxbE    162      150      50       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
1fjgE    162      150      47       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3ohyE    162      150      50       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
2wh3E    162      150      49       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
3uxtE    162      148      51       322      4e-115   100      100      162      0.993827160493827        0.993827160493827
```

Fig. 2 (continued) (**c**) List of homologs and IC scores obtained by RNABindRPlus. These are the homologs used by HomPRIP for making the homology-based predictions. (**d**) List of all potential homologs with structures in the PDB for *T. thermophilus* S5 protein identified by RNABindRPlus. num_residue1 (e.g., 162) denotes the number of amino acids in the query protein; num_residue2 shows the number of amino acids (e.g., 150) in the homolog of the query protein (e.g., 3KNJ, chain E); num_int is the number of binding residues (e.g., 50) in the homolog of the query protein; Bit_score (e.g., 322) gives an indication of the quality of the alignment between the query protein and its homolog—the higher the score, the better the alignment; Evalue is the number of hits expected by chance when searching the database of homologous proteins—the lower the Evalue, the more significant a match to a database sequence is; Positive_Score gives an indication of how many amino acids in the query protein were at least similar to the amino acid sequences found in the database; IdentityScore gives an indication of how many exact matches the query protein had with amino acid sequences in the database; alignment_length is an indication of the number of residues in the query protein aligned with homologs from the database; aligLen_Query is the alignment_length divided by the length of the query protein; aligLen_Homolog is the alignment_length divided by the length of the homolog of the query protein

sequence profiles and other sequence descriptors, such as residue conservation scores, physicochemical properties, and interface propensities. SNBRFinder^T uses profile hidden Markov models to find remote homologs of the query protein sequence, but the basic methodology used for building the classifier is similar to that used in RNABindRPlus.

1. Access the SNBRFinder web server at http://ibi.hzau.edu.cn/SNBRFinder/index.php.
2. Use the radio buttons provided to choose one of three different options for submitting a protein sequence: (a) enter the amino acid sequence in FASTA format; (b) upload a protein sequence file by clicking on "Browse File"; or (c) input UniProt IDs for retrieval (*see* **Note 17**).
3. The user has the option of filtering out proteins homologous to the query protein sequence by specifying a sequence identity threshold. By default, the method excludes homologous templates that share ≥30% sequence identity. To obtain the most reliable predictions, leave this option blank (*see* **Note 18**).

4. Because SNBRFinder can predict either RNA- or DNA-binding residues in proteins, the user should select the binding nucleic acid type (RNA) from a drop-down list. By default, the selection is "DNA."
5. Before clicking on the "submit" button, the user can optionally enter an email address. After the job is submitted, a webpage showing the job id and indicating that the job is running should appear. This page also includes the URL where prediction results will be posted, after they become available. If an email address was provided, the URL will also be included in the email. Typically, results are returned to users after about 15 min.
6. Figure 3 shows results returned by SNBRFinder for the S5 protein from the 30S ribosomal subunit of *T. thermophilus*, which corresponds to protein chain E, in PDB structure 1HNX. Figure 3a shows a summary of the results, in which the query sequence is


a

Summary

Sequence Name: 1HNX:E Length: 162 Nucleic Acid Type: RNA

Optimal Template: N/A HHscore: N/A Sequence Identity: N/A

Query Sequence:

MPETDFEEKMILIRRTARMQAGGRRFRFGALVVVGDRQGRVGLGFGKAPEVPLAVQKAGY

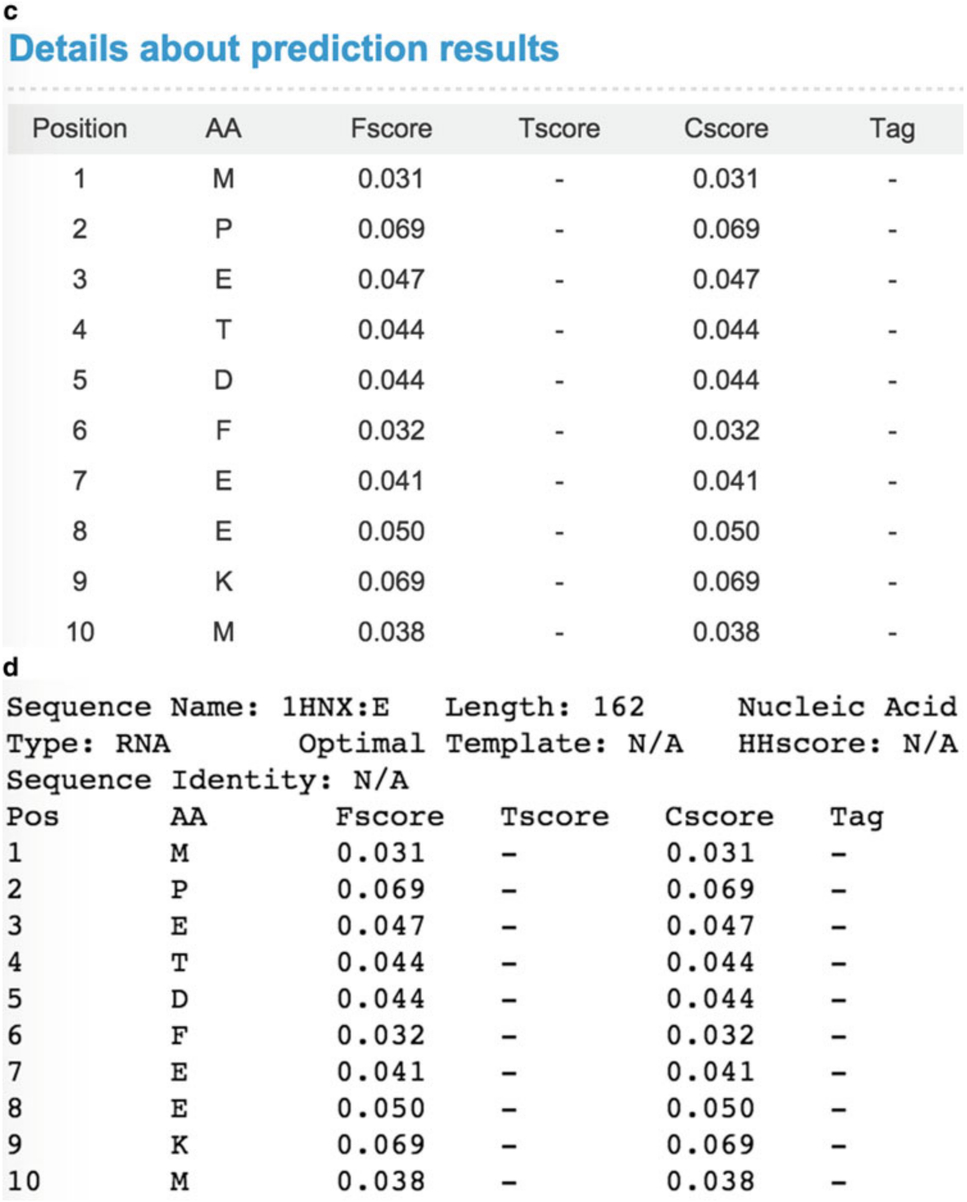
YARRNMVEVPLQNGTIPHEIEVEFGASKIVLKPAAPGTGVIAGAVPRAILELAGVTDILT

KELGSRNPINIAYATMEALRQLRTKADVERLRKGEAHAQAQG

b

Graphic representation

Fig. 3 (**a**) SNBRFinder prediction results summary for the *T. thermophilus* S5 protein. Predicted RNA-binding residues are shown in red. (**b**) Graphical representation of SNBRFinder predictions for the *T. thermophilus* S5 protein. Fscore is the prediction score returned by the feature-based component, SNBRFinder^{F}, and Cscore is the prediction score returned by the combination of the feature-based component and homology/template-based component, SNBRFinder^{T}, of SNBRFinder.



Details about prediction results

Position	AA	Fscore	Tscore	Cscore	Tag
1	M	0.031	-	0.031	-
2	P	0.069	-	0.069	-
3	E	0.047	-	0.047	-
4	T	0.044	-	0.044	-
5	D	0.044	-	0.044	-
6	F	0.032	-	0.032	-
7	E	0.041	-	0.041	-
8	E	0.050	-	0.050	-
9	K	0.069	-	0.069	-
10	M	0.038	-	0.038	-

```
Sequence Name: 1HNX:E   Length: 162     Nucleic Acid
Type: RNA        Optimal Template: N/A   HHscore: N/A
Sequence Identity: N/A
Pos        AA         Fscore    Tscore    Cscore    Tag
1          M          0.031     -         0.031     -
2          P          0.069     -         0.069     -
3          E          0.047     -         0.047     -
4          T          0.044     -         0.044     -
5          D          0.044     -         0.044     -
6          F          0.032     -         0.032     -
7          E          0.041     -         0.041     -
8          E          0.050     -         0.050     -
9          K          0.069     -         0.069     -
10         M          0.038     -         0.038     -
```

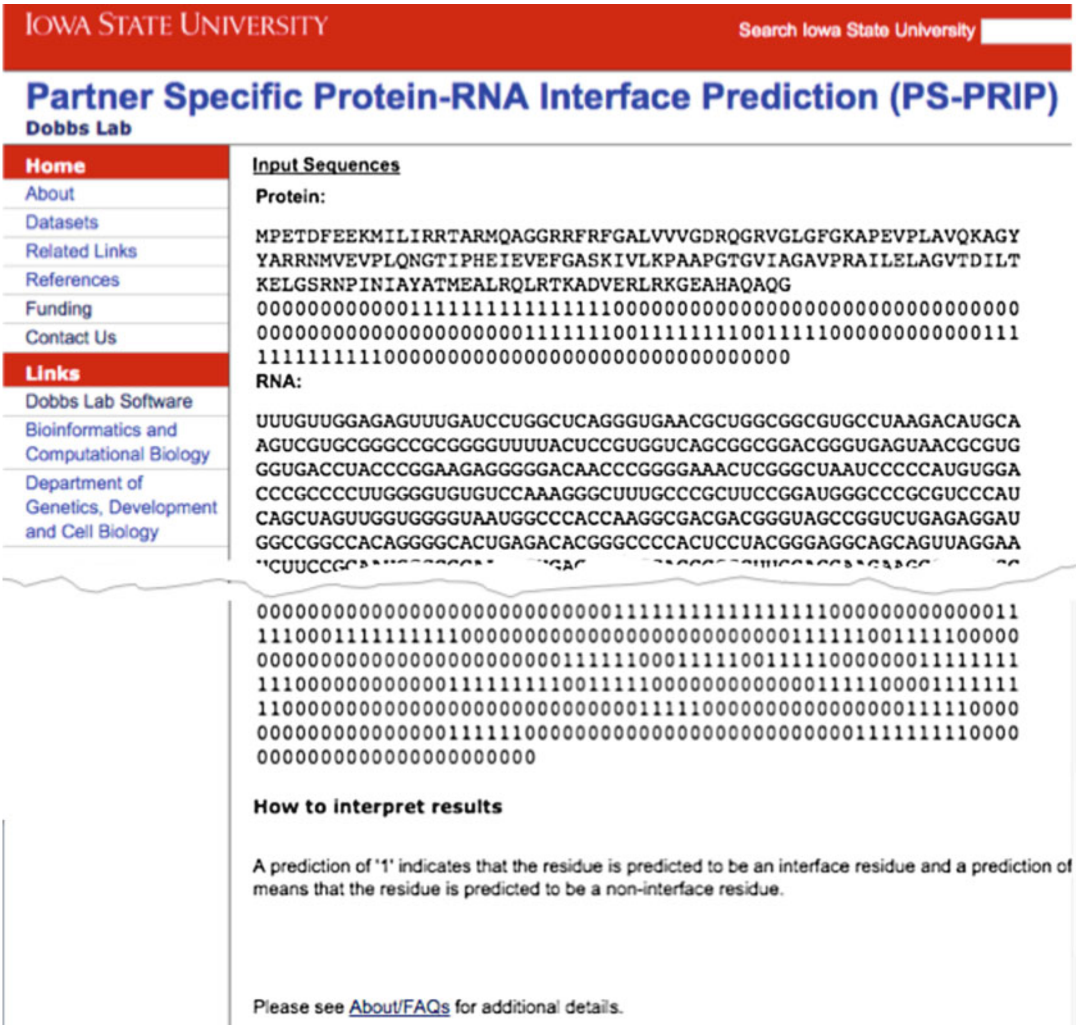
Fig. 3 (continued) (**c**) Table showing SNBRFinder a sample of the detailed results for the *T. thermophilus* S5 protein. See text for additional details. (**d**) Downloadable results from SNBRFinder. Only a portion of the returned results is shown

displayed with predicted interfacial residues highlighted in red text; the query sequence name, length, nucleic acid type, as well as the PDB ID of the optimal template used for making the prediction, the HHscore, if any (*see* **Note 19**), and the % sequence identity (between the query and the optimal template) are also provided. For this example, SNBRFinder was not able to find an optimal template, so HHscore and sequence identity have a value of N/A. Figure 3b shows a graphical representation of the results, which displays a plot of the Fscore and Cscore for each residue, and the Cscore threshold above which a residue is considered an interfacial residue (*see* **Note 20**). Because no optimal template was found for 1HNX chain E, the Fscore is equivalent to the Cscore. Figure 3c shows a detailed results table, which lists each amino acid residue, along with its associated Fscore, Tscore (if any), and Cscore, as well as the "tag" for each amino acid ("+" for interfacial residue, "-" for non-interfacial residue). Figure 3d shows a portion of the results in plain text format, which can be obtained by clicking the "Download the result" link in the top right corner of the "Result" page.

3.4 Using PS-PRIP to Predict Both RNA-Binding and Protein-Binding Residues in RNPs

PS-PRIP (Partner-Specific protein–RNA Interface Prediction) is a sequence motif-based method that can simultaneously predict interfacial residues for both the RNA and protein components of protein–RNA complexes [52] (*see* Subheading 2.5). PS-PRIP is a partner-specific method (*see* **Note 4**), which means that, given the sequences of a protein and several potential interacting RNAs, it can identify which amino acid residues contact each RNA binding partner. In other words, if the protein binds to different RNAs using distinct (or overlapping) interfaces, PS-PRIP can distinguish between these RNA-binding sites. PS-PRIP requires *both* the protein sequence and its partner RNA sequence as input. If the user does not have any potential RNA sequence(s) for testing, methods such as RPI-Seq or catRAPID can be used to infer potential partner RNAs for a specific protein (reviewed in refs. [62–65]). In addition to the sequences of the protein and its RNA-binding partners, PS-PRIP utilizes a dataset of interfacial motifs extracted from solved protein–RNA complexes in the PDB [68]. For predicting RNA-binding residues in proteins, the use of such interfacial motifs by PS-PRIP appears to provide improved precision over RNABindRPlus and other sequence-based interface prediction servers [52]. At present, the RNA-binding residues predicted by PS-PRIP are much more reliable than the protein-binding residues predicted in the bound RNA component.

1. Access the PS-PRIP server at http://pridb.gdcb.iastate.edu/PSPRIP/index.html.
2. Enter a protein sequence and the sequence for an RNA known or expected to be its binding partner in plain text format (protein sequence only and RNA sequence only, without any header information) into the text boxes provided on the homepage (*see* **Note 21**). Then click the "Submit" button.
3. Figure 4 shows results returned by PS-PRIP for the S5 protein from the 30S ribosomal subunit of *T. thermophilus*, which corresponds to protein chain E, in PDB structure 1HNX. In this case, the 16S rRNA corresponding to RNA chain A in the 1HNX structure was provided as input to PS-PRIP, in order to obtain a


IOWA STATE UNIVERSITY
Search Iowa State University
Partner Specific Protein-RNA Interface Prediction (PS-PRIP)
Dobbs Lab
Home
About
Datasets
Related Links
References
Funding
Contact Us
Links
Dobbs Lab Software
Bioinformatics and Computational Biology
Department of Genetics, Development and Cell Biology

Input Sequences

Protein:

```
MPETDFEEKMILIRRTARMQAGGRRFRFGALVVVGDRQGRVGLGFGKAPEVPLAVQKAGY
YARRNMVEVPLQNGTIPHEIEVEFGASKIVLKPAAPGTGVIAGAVPRAILELAGVTDILT
KELGSRNPINIAYATMEALRQLRTKADVERLRKGEAHAQAQG
000000000000111111111111111100000000000000000000000000000000
000000000000000000000111111100111111110011111000000000000111
111111111100000000000000000000000000000000
```

RNA:

```
UUUGUUGGAGAGUUUGAUCCUGGCUCAGGGUGAACGCUGGCGGCGUGCCUAAGACAUGCA
AGUCGUGCGGGCCGCGGGGUUUUACUCCGUGGUCAGCGGCGGACGGGUGAGUAACGCGUG
GGUGACCUACCCGGAAGAGGGGGACAACCCGGGGAAACUCGGGCUAAUCCCCCAUGUGGA
CCCGCCCCUUGGGGUGUGUCCAAAGGGCUUUGCCCGCUUCCGGAUGGGCCCGCGUCCCAU
CAGCUAGUUGGUGGGGUAAUGGCCCACCAAGGCGACGACGGGUAGCCGGUCUGAGAGGAU
GGCCGGCCACAGGGGCACUGAGACACGGGCCCCACUCCUACGGGAGGCAGCAGUUAGGAA
CUUCCG
```

```
000000000000000000000000000011111111111111111000000000000011
111000111111111100000000000000000000000000111111001111100000
000000000000000000000000111111000111110011111000000011111111
111000000000000111111111001111100000000000001111100001111111
110000000000000000000000000000111110000000000000000111110000
000000000000000111111000000000000000000000000001111111110000
00000000000000000000000
```

How to interpret results

A prediction of '1' indicates that the residue is predicted to be an interface residue and a prediction of means that the residue is predicted to be a non-interface residue.

Please see About/FAQs for additional details.

Fig. 4 PS-PRIP prediction results for the *T. thermophilus* S5 protein bound to 16S rRNA. Sequences shown correspond to protein chain E and RNA chain A in the PDB structure 1HNX. Under each sequence, the predicted interfacial residues are represented by a string of 1's and 0's, where "1" and "0" correspond to predicted binding and non-binding residues, respectively

"partner-specific" prediction. On the results page, the S5 protein sequence and 16S rRNA sequences are displayed. In the lines below each sequence, the interfacial residues are indicated by a string of 1's and 0's, where "1" and "0" correspond to predicted interfacial and non-interfacial residues, respectively.

3.5 Actual RNA-Binding Residues Compared with Predictions Using Three Different Methods

Figure 5 shows a comparison of the predicted RNA-binding residues in the *T. thermophilus* S5 ribosomal protein, for which a 3D structure is available in the PDB (1HNX; protein chain E, RNA chain A). The top line shows the amino acid sequence of the S5 protein, with red letters denoting the actual RNA-binding residues (58 out of 162 total residues), defined on the basis of a 5 Å

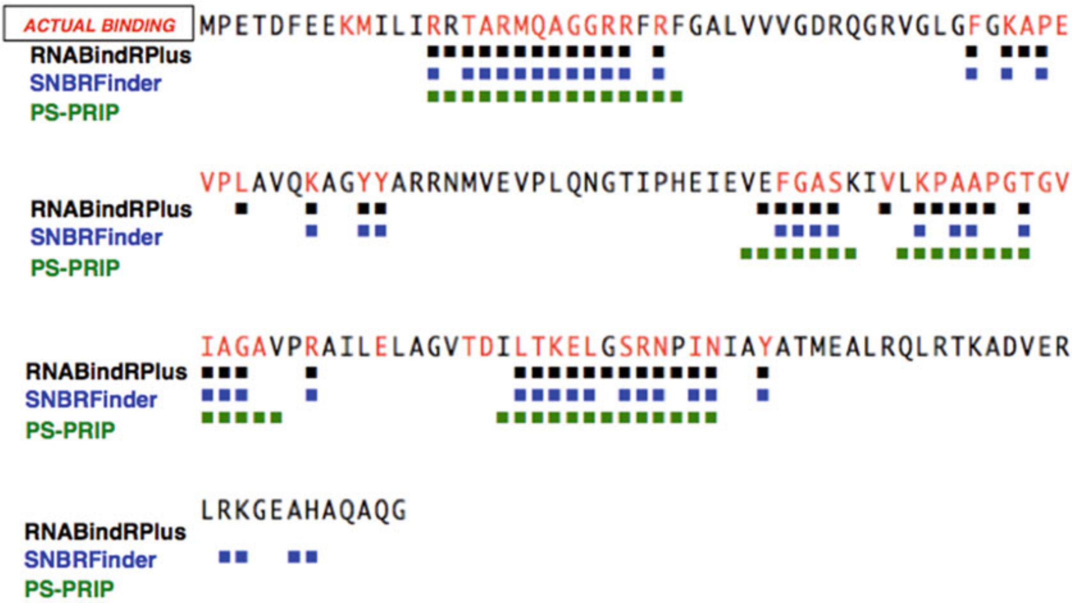

ACTUAL BINDING
MPETDFEEKMILIRRTARMQAGGRRFRFGALVVVGDRQGRVGLGFGKAPE
RNABindRPlus
SNBRFinder
PS-PRIP
VPLAVQKAGYYARRNMVEVPLQNGTIPHEIEVEFGASKIVLKPAAPGTGV
RNABindRPlus
SNBRFinder
PS-PRIP
IAGAVPRAILELAGVTDILTKELGSRNPINIAYATMEALRQLRTKADVER
RNABindRPlus
SNBRFinder
PS-PRIP
LRKGEAHAQAQG
RNABindRPlus
SNBRFinder
PS-PRIP

Fig. 5 Actual vs. predicted RNA-binding residues in the *T. thermophilus* S5 ribosomal protein sequence. *Top line:* Actual RNA-binding residues are shown in *red*, non-binding residues are *black. Lower lines:* Predictions obtained using RNABindRPlus, SNBRFinder and PS-PRIP. Colored boxes indicate predicted RNA-binding residues. Sequence corresponds to: PDB 1HNX; protein chain E

distance cutoff (*see* **Note 1**). RNA-binding residues predicted by RNABindRPlus, SNBRFinder and PS-PRIP are shown below. In this example, all three methods were able to identify the majority of the 58 RNAbinding residues: RNABindRPlus (46/58) SNBRFinder (41/58), PS-PRIP (33/58). A small number of false positive predictions were returned by RNABindRPlus (4), SNBRFinder (4), and a larger number by PS-PRIP (12).

In this particular example, "better than average" results were obtained because the S5 protein is a highly conserved component of the 30S ribosomal subunit. For the S5 protein, the RNA-binding residues predicted by PS-PRIP are less reliable than those predicted by RNABindRPlus and SNBRFinder. But, because the sequence of the bound RNA is also available, PS-PRIP also returns predictions for *protein*-binding residues in the 16S rRNA, which the other two servers cannot do. This example illustrates that although the overall performance of PS-PRIP was superior in terms of *precision* when tested on a benchmark dataset [52], both RNABindRPlus and SNBRFinder may perform better on certain proteins. Given the purpose of this chapter, the important point is that all three servers predict similar patches of RNA-binding residues, providing the user with a remarkably accurate prediction of the RNA-binding residues in the S5 protein, without using any structural information in order to make these predictions.

In closing, we again encourage users to submit query protein(s) of interest to at least two or three different servers from the list in Table 2, and to evaluate predictions in the context of the 3D structure, if available. All prediction results should be interpreted with caution: the computational tools are intended to help users identify the most probable RNA-binding residues in proteins, i.e., to generate hypotheses that can limit the number of experiments needed to determine RNA-binding residues using biochemical or biophysical approaches.

4 Notes

1. RNA-binding residues in proteins or other **"interfacial residues"** in the interface formed when a protein binds RNA (or DNA or another protein) are typically defined in one of two ways: (a) using a contact distance threshold, e.g., an interfacial residue is any amino acid with a heavy atom within *n* Å of a heavy atom in the bound RNA (where *n* typically ranges from 3.5 to 8 Å); (b) residues whose accessible surface area is reduced by >1 Å^2 upon complex formation [101]. It is very important to take into account how interfacial residues are defined when comparing the performance of various computational methods for predicting RNA-binding residues in proteins [47].
2. Two databases that once provided comprehensive information about interfaces in protein–RNA complexes in the PDB are no longer up-to-date: **PRIDB** [76] and **BIPA** [72]. Efforts to update PRIDB are underway. Two resources that are currently maintained and provide detailed information about interfaces in RBPs include: **NPIDB** [74] and **DBBP** [75].
3. A **position-specific scoring matrix (PSSM)** is a type of weighted scoring matrix derived from a set of aligned sequences that are considered to be homologous or functionally related [102]. PSSMs can be very sensitive because they capture important evolutionary information by exploiting the large number of protein sequences currently available.
4. A **partner-specific prediction method** takes into account the potential interacting partner(s) in predicting interfacial residues. For example, if a protein binds two distinct RNAs, RNA-1 and RNA-2, a partner-specific method will return one set of amino acids that specifically interact with only RNA-1, and a second set of amino acids that specifically interact with only RNA-2. Note that the two sets of RNA-binding residues may overlap.
5. At present, none of the available servers for predicting RNA-binding residues in proteins provide the user with existing information regarding experimentally determined RNA-binding residues (i.e., the servers always return *predicted* RNA-

binding residues, which may not be the same as the actual interfacial residues determined by experiment). Thus, as a first step, the user should always search published literature (via search engines such as **NCBI/PubMed** (http://www.ncbi.nlm.nih.gov/) or **Google Scholar** (http://scholar.google.com) and relevant databases (*see* Subheading 3.1) for existing experimental data regarding the specific RNA-binding protein(s) of interest. In addition to the resources described in Subheading 3.1 and Table 1, many new databases and servers that provide extensive information regarding protein–RNA complexes, RNA-binding proteins and their recognition sites, and in vivo protein–RNA interaction networks are becoming available. OMICtools (http://omictools.com) provides an extensive and up-to-date directory of these resources [103].

6. Users unfamiliar with **BLAST** should first read BLAST documentation and/or tutorials. A beginner's guide is available here: ftp://ftp.ncbi.nlm.nih.gov/pub/factsheets/HowTo_BLASTGuide.pdf.
7. **SmartBLAST** is a new version of BLAST that is faster than BLASTp and offers a user-friendly graphical view. For additional information, see: http://ncbiinsights.ncbi.nlm.nih.gov/2015/07/29/smartblast/.
8. **Tip:** Because proteins from humans are usually much better annotated than those from other organisms, valuable clues regarding potential RNA-binding domains or motifs in a protein can be obtained by visiting the NCBI GenBank Protein entry for the human homolog of a query sequence, if available.
9. Under the **"Related Information"** header on the GenBank Protein entry page, the user can access several different types of information, e.g., clicking on the **"Related Structures (Summary)"** link returns structurally related proteins found in NCBI's Molecular Modeling Database (MMDB), as well as an alignment of the query protein sequence with its potential homolog(s), and links for visualizing the 3D structures. Alternatively, the user can perform BLAST or Conserved Domain searches by clicking links under the **"Analyze this sequence"** header (located at the top of right-side panel), but it is usually more efficient to take advantage of precomputed information available under "Related Resources," e.g., "Blink" (for BLAST results, instead of "Run Blast"); or "CDD Search Results" (instead of "Identify Conserved Domains").
10. The **PDB Advanced Search** (http://www.rcsb.org/pdb/search/advSearch.do?search=new) is a powerful tool that allows the user to BLAST a sequence of interest against all structures in the database, to identify GO annotations, citations in publications, etc. In addition, the PDB offers several

built-in visualization tools (http://www.rcsb.org/pdb/secondary.do?p=v2/secondary/visualize.jsp—RCSBviewer) as well as links to additional resources and software for analyzing macromolecular structures (http://www.rcsb.org/pdb/static.do?p=general_information/web_links/index.html)

11. The **NDB** [69] focuses on structures that contain either RNA or DNA and provides links to many valuable RNA sequence and structure analysis tools (http://ndbserver.rutgers.edu/ndbmodule/services/index.html) as well as software for identifying RNA motifs and for predicting secondary and tertiary structures of RNA molecules (http://ndbserver.rutgers.edu/ndbmodule/services/softwares.html).
12. Currently, there is a wait of approximately 10 min per protein sequence submitted to RNABindRPlus. The rate-limiting step is generating the PSSMs using PSI-BLAST [98]. To obtain results more quickly, the user is encouraged to split large jobs into several smaller submissions (e.g., if the user would like to submit 100 proteins, she/he should submit 5 smaller jobs of 20 proteins each).
13. A faster version of this server, **FastRNABindR**, is under development. When it becomes available, a link to FastRNABindR will be provided on the RNABindRPlus website (http://ailab1.ist.psu.edu/RNABindRPlus/).
14. The user should submit the protein sequence in upper case letters to the RNABindRPlus web server. Note that this server predicts RNA-binding residues in proteins, so RNA nucleotides are not valid input.
15. The homology-based component of RNABindRPlus, **HomPRIP**, searches for homologs of the query protein. Excluding similar sequences (>30 % sequence identity) ensures that the homolog and the query protein are not the same. This is useful for stringently evaluating performance of RNABindRPlus in comparison with other methods, but is not the best strategy for a user interested in identifying potential RNA-binding residues. To obtain the best possible prediction of RNA-binding residues, the user should take full advantage of all available homologous sequences (i.e., should *not* eliminate any potential homologs).
16. The **IC_score** (interface conservation score) measures the correlation between the interface and non-interface residues of a query protein Q and its putative sequence homolog H when the two are aligned. It is a measure of how well the RNA-binding residues of Q are conserved (and subsequently, can be predicted from known interface residues of homologous proteins) in protein H. However, computing the IC_score requires knowledge of interface residues in both the query protein and

its homolog. Fortunately, for a query protein with unknown RNA-binding residues, the IC_score can be estimated using BLAST alignment statistics between Q and H [57].

17. SNBRFinder allows submission of at most five sequences each time, for any of the submission options. When entering multiple UniProt IDs, IDs should be separated by commas.
18. Like RNABindRPlus, SNBRFinder allows the user to specify which sequences to exclude when searching for homologous templates, using a sequence identity cutoff. Protein templates that are more similar to the query protein are likely to return better results than templates that are less similar. The sequence identity cutoff utilized depends on the user's objective (*see* **Note 15**). To obtain the best possible prediction of RNA-binding residues, the user should take full advantage of all available homologous sequences. In contrast, for a rigorous performance comparison with other methods, a lower sequence identity cutoff should be used (i.e., to evaluate the sensitivity and specificity of the methods).
19. **HHscore** is a score that indicates the similarity score between the query protein and its best homolog/template.
20. SNBRFinder calculates the probability score of each residue being an RNA-binding residue using the following formula:

$$\text{Cscore} = \begin{cases} \alpha\text{Fscore} + (1-\alpha)\text{Tscore if HHscore} \geq \text{cutoff} \\ \text{Fscore otherwise} \end{cases}$$

where Fscore is the output of SNBRFinder$^{\text{F}}$ (support vector machine component) and Tscore is the output of SNBRFinder$^{\text{T}}$ (template-based component), $\alpha = 0.6$ and cutoff = 85 %.

21. A current limitation of PS-PRIP is that it has a minimum length requirement for both the protein and RNA sequences: proteins must be ≥25 amino acids in length and RNAs must be ≥100 nucleotides in length.

Acknowledgments

This work was supported in part by NSF DBI0923827 to DD, by NIH GM066387 to VGH and DD, by a Presidential Initiative for Interdisciplinary Research (PIIR) award to DD from Iowa State University, and by the Edward Frymoyer Chair in Information Sciences and Technology held by VGH at Pennsylvania State University. RRW is currently supported by an appointment to the ARS-USDA Research Participation Program administered by the Oak Ridge Institute for Science and Education (ORISE) through an interagency agreement between the US Department of Energy

(DOE) and USDA. ORISE is managed by ORAU under DOE contract number DE-AC05-06OR23100. We thank Carla Mann and Usha Muppirala for valuable discussions.

References

1. Re A, Joshi T, Kulberkyte E et al (2014) RNA-protein interactions: an overview. Methods Mol Biol 1097:491–521
2. Lee Y, Rio DC (2015) Mechanisms and regulation of alternative pre-mRNA splicing. Annu Rev Biochem 84:291–323
3. Fu X-D, Ares M Jr (2014) Context-dependent control of alternative splicing by RNA-binding proteins. Nat Rev Genet 15(10):689–701
4. Singh G, Pratt G, Yeo GW et al (2015) The clothes make the mRNA: past and present trends in mRNP fashion. Annu Rev Biochem 84:325–354
5. Bryant CD, Yazdani N (2016) RNA binding proteins, neural development and the addictions. Genes Brain Behav 15(1):169–186.
6. Hogg JR, Collins K (2008) Structured non-coding RNAs and the RNP renaissance. Curr Opin Chem Biol 12(6):684–689
7. Cech TR, Steitz JA (2014) The noncoding RNA revolution-trashing old rules to forge new ones. Cell 157(1):77–94
8. Castello A, Hentze MW, Preiss T (2015) Metabolic enzymes enjoying new partnerships as RNA-binding proteins. Trends Endocrinol Metab 26(12):746–757
9. Beckmann BM, Horos R, Fischer B et al (2015) The RNA-binding proteomes from yeast to man harbour conserved enigmRBPs. Nat Commun 6:10127
10. Lin Y, Protter DS, Rosen MK et al (2015) Formation and maturation of phase-separated liquid droplets by RNA-binding proteins. Mol Cell 60(2):208–219
11. Kafasla P, Skliris A, Kontoyiannis DL (2014) Post-transcriptional coordination of immunological responses by RNA-binding proteins. Nat Immunol 15(6):492–502
12. Darnell RB (2010) RNA regulation in neurologic disease and cancer. Cancer Res Treat 42(3):125–129
13. Wurth L, Gebauer F (2015) RNA-binding proteins, multifaceted translational regulators in cancer. Biochim Biophys Acta 1849(7):881–886
14. Pilaz LJ, Silver DL (2015) Post-transcriptional regulation in corticogenesis: how RNA-binding proteins help build the brain. Wiley Interdiscip Rev RNA 6(5):501–515
15. Gerstberger S, Hafner M, Tuschl T (2014) A census of human RNA-binding proteins. Nat Rev Genet 15(12):829–845
16. Neelamraju Y, Hashemikhabir S, Janga SC (2015) The human RBPome: from genes and proteins to human disease. J Proteomics 127(Pt A):61–70
17. Vaquerizas JM, Kummerfeld SK, Teichmann SA et al (2009) A census of human transcription factors: function, expression and evolution. Nat Rev Genet 10(4):252–263
18. Tsvetanova NG, Klass DM, Salzman J et al (2010) Proteome-wide search reveals unexpected RNA-binding proteins in *Saccharomyces cerevisiae*. PLoS One 5(9)
19. Castello A, Fischer B, Eichelbaum K et al (2012) Insights into RNA biology from an atlas of mammalian mRNA-binding proteins. Cell 149(6):1393–1406
20. Hashemikhabir S, Neelamraju Y, Janga SC (2015) Database of RNA binding protein expression and disease dynamics (READ DB). Database (Oxford) 2015:bav072
21. Tamburino AM, Ryder SP, Walhout AJ (2013) A compendium of *Caenorhabditis elegans* RNA binding proteins predicts extensive regulation at multiple levels. G3 (Bethesda) 3(2):297–304
22. Ray D, Kazan H, Cook KB et al (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499(7457):172–177
23. Jiang J, Chan H, Cash DD et al (2015) Structure of *Tetrahymena* telomerase reveals previously unknown subunits, functions, and interactions. Science 350(6260):aab4070. doi: 10.1126/science.aab4070
24. Zhang X, Ding K, Yu X et al (2015) In situ structures of the segmented genome and RNA polymerase complex inside a dsRNA virus. Nature 527(7579):531–534
25. Chen Y, Varani G (2013) Engineering RNA-binding proteins for biology. FEBS J 280(16):3734–3754
26. Wei H, Wang Z (2015) Engineering RNA-binding proteins with diverse activities. Wiley Interdiscip Rev RNA 6(6):597–613
27. Lunde BM, Moore C, Varani G (2007) RNA-binding proteins: modular design for efficient function. Nat Rev Mol Cell Biol 8(6):479–490

28. Varadi M, Zsolyomi F, Guharoy M et al (2015) Functional advantages of conserved intrinsic disorder in RNA-binding proteins. PLoS One 10(10):e0139731
29. Calabretta S, Richard S (2015) Emerging roles of disordered sequences in RNA-binding proteins. Trends Biochem Sci 40(11):662–672
30. Terribilini M, Lee JH, Yan C et al (2006) Prediction of RNA binding sites in proteins from amino acid sequence. RNA 12(8):1450–1462
31. Puton T, Kozlowski L, Tuszynska I et al (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol 179(3):261–268
32. Ke A, Doudna JA (2004) Crystallization of RNA and RNA-protein complexes. Methods 34(3):408–414
33. Wu H, Finger LD, Feigon J (2005) Structure determination of protein/RNA complexes by NMR. Methods Enzymol 394:525–545
34. Carlomagno T (2014) Present and future of NMR for RNA-protein complexes: a perspective of integrated structural biology. J Magn Reson 241:126–136
35. Binshtein E, Ohi MD (2015) Cryo-electron microscopy and the amazing race to atomic resolution. Biochemistry 54(20):3133–3141
36. Hennig J, Sattler M (2015) Deciphering the protein-RNA recognition code: combining large-scale quantitative methods with structural biology. Bioessays 37(8):899–908
37. Faoro C, Ataide SF (2014) Ribonomic approaches to study the RNA-binding proteome. FEBS Lett 588(20):3649–3664
38. McHugh CA, Russell P, Guttman M (2014) Methods for comprehensive experimental identification of RNA-protein interactions. Genome Biol 15(1):203
39. Campbell ZT, Wickens M (2015) Probing RNA-protein networks: biochemistry meets genomics. Trends Biochem Sci 40(3):157–164
40. Cook KB, Hughes TR, Morris QD (2015) High-throughput characterization of protein-RNA interactions. Brief Funct Genomics 14(1):74–89
41. Cook KB, Kazan H, Zuberi K et al (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39(Database issue):D301–D308
42. Li X, Kazan H, Lipshitz HD et al (2014) Finding the target sites of RNA-binding proteins. Wiley Interdiscip Rev RNA 5(1):111–130
43. Kazan H, Morris Q (2013) RBPmotif: a web server for the discovery of sequence and structure preferences of RNA-binding proteins. Nucleic Acids Res 41(Web Server issue):W180–W186
44. Banerjee H, Singh R (2008) A simple cross-linking method, CLAMP, to map the sites of RNA-contacting domains within a protein. Methods Mol Biol 488:181–190
45. Kramer K, Sachsenberg T, Beckmann BM et al (2014) Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. Nat Methods 11(10):1064–1070
46. Qamar S, Kramer K, Urlaub H (2015) Studying RNA-protein interactions of pre-mRNA complexes by mass spectrometry. Methods Enzymol 558:417–463
47. Walia RR, Caragea C, Lewis BA et al (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 13(1):89
48. Zhao H, Yang Y, Zhou Y (2013) Prediction of RNA binding proteins comes of age from low resolution to high resolution. Mol Biosyst 9(10):2417–2425
49. Nagarajan R, Gromiha MM (2014) Prediction of RNA binding residues: an extensive analysis based on structure and function to select the best predictor. PLoS One 9(3):e91140
50. Si J, Cui J, Cheng J et al (2015) Computational prediction of RNA-binding proteins and binding sites. Int J Mol Sci 16(11):26303–26317
51. Mitchell A, Chang HY, Daugherty L et al (2015) The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res 43(Database issue):D213–D221
52. Muppirala UK, Lewis BA, Mann CM et al (2016) A motif-based method for predicting interfacial residues in both the RNA and protein components of protein-RNA complexes. Pac Symp Biocomput 2016:445–455. doi:10.1142/9789814749411_0041
53. Williamson JR (2000) Induced fit in RNA-protein recognition. Nat Struct Biol 7(10):834–837
54. Ellis JJ, Jones S (2008) Evaluating conformational changes in protein structures binding RNA. Proteins 70(4):1518–1526
55. Sankar K, Walia R, Mann C et al (2014) An analysis of conformational changes upon RNA-protein binding. In: ACM BCB 2014 5th ACM conference on bioinformatics, computational biology, and health informatics, Washington, DC, 2013. ACM New York, NY, USA ©2014 pp 592–593 doi: 10.1145/2649387.2660790
56. Spriggs RV, Jones S (2009) RNA-binding residues in sequence space: conservation and interaction patterns. Comput Biol Chem 33(5):397–403
57. Walia RR, Xue LC, Wilkins K et al (2014) RNABindRPlus: a predictor that combines machine learning and sequence homology-based methods to improve the reliability of

predicted RNA-binding residues in proteins. PLoS One 9(5):e97725

58. Yang X, Wang J, Sun J et al (2015) SNBRFinder: a sequence-based hybrid algorithm for enhanced prediction of nucleic acid-binding residues. PLoS One 10(7):e0133260
59. Tuszynska I, Matelska D, Magnus M et al (2014) Computational modeling of protein-RNA complex structures. Methods 65(3):310–319
60. Gupta A, Gribskov M (2011) The role of RNA sequence and structure in RNA—protein interactions. J Mol Biol 409(4):574–587
61. Panwar B, Raghava GP (2015) Identification of protein-interacting nucleotides in a RNA sequence using composition profile of trinucleotides. Genomics 105(4):197–203
62. Mann C, Muppirala UK, Dobbs DL (2016) Computational prediction of RNA-protein interactions. Methods Mol Biol. In press
63. Muppirala UK, Lewis BA, Dobbs D (2013) Computational tools for investigating RNA-protein interaction partners. J Comput Sci Syst Biol 6:182–187
64. Cirillo D, Livi CM, Agostini F et al (2014) Discovery of protein-RNA networks. Mol Biosyst 10(7):1632–1642
65. Marchese D, Livi CM, Tartaglia GG (2016) A computational approach for the discovery of protein-RNA networks. Methods Mol Biol 1358:29–39
66. Zhao H, Yang Y, Janga SC et al (2014) Prediction and validation of the unexplored RNA-binding protein atlas of the human proteome. Proteins 82(4):640–647
67. Kumar M, Gromiha MM, Raghava GP (2011) SVM based prediction of RNA-binding proteins using binding residues and evolutionary information. J Mol Recognit 24(2):303–313
68. Berman HM, Westbrook J, Feng Z et al (2000) The protein data bank. Nucleic Acids Res 28(1):235–242
69. Coimbatore Narayanan B, Westbrook J, Ghosh S et al (2014) The nucleic acid database: new features and capabilities. Nucleic Acids Res 42(Database issue):D114–D122
70. de Beer TA, Berka K, Thornton JM et al (2014) PDBsum additions. Nucleic Acids Res 42(Database issue):D292–D296
71. Laskowski RA, Hutchinson EG, Michie AD et al (1997) PDBsum: a Web-based database of summaries and analyses of all PDB structures. Trends Biochem Sci 22(12):488–490
72. Lee S, Blundell TL (2009) BIPA: a database for protein-nucleic acid interaction in 3D structures. Bioinformatics 25(12):1559–1560
73. Jones P, Binns D, Chang HY et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30(9):1236–1240
74. Kirsanov DD, Zanegina ON, Aksianov EA et al (2013) NPIDB: nucleic acid—protein interaction database. Nucleic Acids Res 41(D1):D517–D523
75. Park B, Kim H, Han K (2014) DBBP: database of binding pairs in protein-nucleic acid interactions. BMC Bioinformatics 15(Suppl 15):S5
76. Lewis BA, Walia RR, Terribilini M et al (2011) PRIDB: a protein-RNA interface database. Nucleic Acids Res 39(Database issue):D277–D282
77. Shulman-Peleg A, Nussinov R, Wolfson HJ (2009) RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. Nucleic Acids Res 37(Suppl 1):D369–D373
78. Kumar MDS, Bava KA, Gromiha MM et al (2006) ProTherm and ProNIT: thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34(Database issue):D204–D206
79. Vanegas PL, Hudson GA, Davis AR et al (2012) RNA CoSSMos: characterization of secondary structure motifs—a searchable database of secondary structure motifs in RNA three-dimensional structures. Nucleic Acids Res 40(Database issue):D439–D444
80. Petrov AI, Zirbel CL, Leontis NB (2013) Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. RNA 19(10):1327–1340
81. Chojnowski G, Walen T, Bujnicki JM (2014) RNA Bricks—a database of RNA 3D motifs and their interactions. Nucleic Acids Res 42(Database issue):D123–D131
82. Livi CM, Klus P, Delli Ponti R et al (2015) catRAPID signature: identification of ribonucleoproteins and RNA-binding regions. Bioinformatics. Oct 31. pii: btv629. [Epub ahead of print]
83. Wang L, Brown SJ (2006) BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. Nucleic Acids Res 34(suppl 2):W243–W248
84. Wang L, Huang C, Yang MQ et al (2010) BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst Biol 4(Suppl 1):S3
85. Zhao H, Yang Y, Zhou Y (2011) Structure-based prediction of RNA-binding domains and RNA-binding sites and application to

structural genomics targets. Nucleic Acids Res 39(8):3017–3025

86. Kim OTP, Yura K, Go N (2006) Amino acid residue doublet propensity in the protein–RNA interface and its application to RNA interface prediction. Nucleic Acids Res 34(22):6450–6460
87. Carson MB, Langlois R, Lu H (2010) NAPS: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res 38(Web Server Issue):W431–W435
88. Pérez-Cano L, Fernández-Recio J (2010) Optimal protein-RNA area, OPRA: a propensity-based method to identify RNA-binding sites on proteins. Proteins 78(1):25–35
89. Kumar M, Gromiha MM, Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins 71(1):189–194
90. Ma X, Guo J, Wu J et al (2011) Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins 79(4):1230–1239
91. Maetschke SR, Yuan Z (2009) Exploiting structural and topological information to improve prediction of RNA-protein binding sites. BMC Bioinformatics 10(1):341
92. Miao Z, Westhof E (2015) Prediction of nucleic acid binding probability in proteins: a neighboring residue network based score. Nucleic Acids Res 43(11):5340–5351
93. Tong J, Jiang P, Lu Z-H (2008) RISP: a web-based server for prediction of RNA-binding sites in proteins. Comput Methods Programs Biomed 90(2):148–153
94. Terribilini M, Sander JD, Lee JH et al (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucleic Acids Res 35(Web Server issue):W578–W584
95. Yang Y, Zhao H, Wang J et al (2014) SPOT-Seq-RNA: predicting protein-RNA complex structure and RNA-binding function by fold recognition and binding affinity prediction. Methods Mol Biol 1137:119–130
96. Remmert M, Biegert A, Hauser A et al (2012) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat Methods 9(2):173–175
97. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. J Mol Biol 215(3):403–410
98. Altschul SF, Madden TL, Schaffer AA et al (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
99. Lambert N, Robertson A, Jangi M et al (2014) RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins. Mol Cell 54(5):887–900
100. Paz I, Kosti I, Ares M Jr et al (2014) RBPmap: a web server for mapping binding sites of RNA-binding proteins. Nucleic Acids Res 42(Web Server issue):W361–W367
101. Jones S, Daley DT, Luscombe NM et al (2001) Protein-RNA interactions: a structural analysis. Nucleic Acids Res 29(4):943–954
102. Stormo GD, Schneider TD, Gold L et al (1982) Use of the "Perceptron" algorithm to distinguish translational initiation sites in *E. coli*. Nucleic Acids Res 10(9):2997–3011
103. Henry VJ, Bandrowski AE, Pepin AS et al (2014) OMICtools: an informative directory for multi-omic data analysis. Database (Oxford). doi:10.1093/database/bau069