Journal of
Computer Science & Systems Biology

Muppirala et al., J Comput Sci Syst Biol 2013, 6:4
DOI: 10.4172/jcsb.1000115

Review Article Open Access

Computational Tools for Investigating RNA-Protein Interaction Partners

Usha K Muppirala*, Benjamin A Lewis and Drena Dobbs

Bioinformatics and Computational Biology Program, Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, USA

Abstract

RNA-protein interactions are important in a wide variety of cellular and developmental processes. Recently, high-throughput experiments have begun to provide valuable information about RNA partners and binding sites for many RNA-binding proteins (RBPs), but these experiments are expensive and time consuming. Thus, computational methods for predicting RNA-Protein interactions (RPIs) can be valuable tools for identifying potential interaction partners of a given protein or RNA, and for identifying likely interfacial residues in RNA-protein complexes. This review focuses on the "partner prediction" problem and summarizes available computational methods, web servers and databases that are devoted to it. New computational tools for addressing the related "interface prediction" problem are also discussed. Together, these computational methods for investigating RNA-protein interactions provide the basis for new strategies for integrating RNA-protein interactions into existing genetic and developmental regulatory networks, an important goal of future research.

Keywords: RNA-protein interaction; RNA-binding protein; Partner prediction; Interface prediction; RNA-protein database

Abbreviations: RBPs: RNA-Binding Proteins; RPIs: RNA-Protein Interactions

Introduction

In the post-transcriptional regulation of gene expression, RNA-binding proteins (RBPs) interact with target mRNAs and non-coding RNAs (ncRNAs) to regulate a variety of cellular processes, including RNA splicing, RNA transport and stability, and translation [1-3]. RNA-protein interactions (RPIs) also play important roles in human health and diseases [4], as well as in viral replication [5], and pathogen resistance in plants [6]. Even though the human genome contains more than 400 known or predicted RBPs [7,8], the structures of RNA-protein complexes and the roles of RPIs in post-transcriptional regulatory networks [1,9], are much less well characterized than the DNA-protein complexes involved in transcriptional regulation. For example, on July 18, 2013, the Protein Data Bank (PDB) [10] contained only 1,593 structures of RNA-protein complexes, compared with more than 2,800 structures of DNA-protein complexes. Recently, however, new experimental approaches have been used to interrogate RNA-protein complexes and interaction networks. For example, high-throughput *in vivo* and *in vitro* experiments have been used to identify cellular RNA molecules that bind a protein of interest [11-13]. Global proteomic approaches have been applied to identify the entire mRNA-bound proteome [14].

The available structures of RNA-protein complexes in the PDB, databases of protein and RNA motifs, and a growing knowledge base regarding RNA and protein interactions in the literature have been exploited to develop computational methods for addressing several questions about RNA-protein interactions:

- Does this protein bind RNA?
- Which RNA molecules are bound by this protein?
- Which RNA sequence or structural motifs are recognized by this protein?
- Which amino acid residues are directly involved in binding RNA?

In this review, we focus on existing computational methods and web servers for predicting RNA-protein interaction partners. We also discuss recently developed "partner-aware" approaches for predicting RNA-protein interfaces, which use information about both the protein and RNA molecules to identify binding regions in either one or both sequences. Finally, available curated databases of RNA-protein interactions are briefly reviewed.

RNA-Protein Partner Prediction Methods and Web Servers

Table 1 summarizes the characteristics of computational methods available for predicting the interaction probability of a given RNA-protein pair. A general description of the machine learning methods and performance metrics discussed below is provided in Supplementary Text S1.

To the best of our knowledge, the first method for computationally predicting mRNA-protein interactions was proposed by Pancaldi and Bähler [15]. Their study took advantage of a dataset of 5,166 mRNA-RBP interactions detected using RNA immunopurification experiments performed in *S. cerevisiae* [16]. Two machine learning methods, Support Vector Machines (SVMs) and Random Forest (RF) classifiers (Supplementary Text S1), were used to predict the likelihood of interaction between an RBP and its target mRNAs. Input for the classifiers included more than 100 characteristic gene and protein features, but no motifs or experimentally measured binding specificities were used. Feature classes included gene ontology terms, predicted secondary structures, mRNA properties and genetic interactions. Overall, the RF classifier performed slightly better than SVM. In 2-fold cross validation experiments, an average prediction accuracy of 69% was obtained, with average sensitivity of 70% and specificity of 69%. When the authors tried to predict the mRNA targets

***Corresponding author:** Usha K Muppirala, Bioinformatics and Computational Biology Program, Department of Genetics, Development and Cell Biology, Iowa State University, Ames, Iowa, USA, E-mail: usha@iastate.edu

Received July 22, 2013; **Accepted** August 14, 2013; **Published** August 21, 2013

Citation: Muppirala UK, Lewis BA, Dobbs D (2013) Computational Tools for Investigating RNA-Protein Interaction Partners. J Comput Sci Syst Biol 6: 182-187. doi:10.4172/jcsb.1000115

Method	Dataset	Features	Description
Pancaldi and Bähler [15]	5,166 mRNA-protein interacting pairs from immunopurification experiments	Predicted protein secondary structure, localization, protein physical properties, gene physical properties, UTR properties, genetic interactions	Protein and RNA sequences encoded using > 100 features are used to train SVM and RF classifiers
Bellucci et al. (catRAPID) [17]	7,409 interacting pairs from 858 RNA-protein complexes from PDB	Physicochemical properties including secondary structure propensities, hydrogen-bonding propensities, and van der Waals interaction propensities	Propensities are calculated for each amino acid and ribonucleotide to generate an interaction profile (http://service.tartaglialab.com/page/catrapid_group)
Muppirala et al. (RPISeq) [22]	2,241 interacting pairs from 943 RNA-protein complexes from PRIDB (RPI2241)	Sequence composition of proteins, represented as conjoint triads, and RNAs, represented as tetrads	Protein and RNA sequences encoded sequence-composition-based features are used to train SVM and RF classifiers (http://pridb.gdcb.iastate.edu/RPISeq)
Wang et al. [26]	RPI 2241 generated by Muppirala et al. & 367 interacting pairs from NPInter	Sequence composition of protein and RNA	Input to NB and ENB classifiers is a combination of protein triads and RNA triad features similar to those used in RPISeq

Table 1: Computational Methods for Predicting RNA-Protein Interaction Partners.

Database	URL	Description
BioGRID [43]	http://thebiogrid.org/	Manually curated protein and genetic interactions for major model organisms
IntAct [44]	http://www.ebi.ac.uk/intact/	Manually curated molecular interactions, including comprehensive data about their source experiments
NDB [42]	http://ndbserver.rutgers.edu/	Nucleic acid and DNA/RNA-protein complex structures, including derived data for nucleic acids
NPInter [18]	http://www.panrna.org/NPInter/index.php	Functional interactions of ncRNAs and protein-related biomolecules, classified into categories based on interaction type
PDB [10]	http://www.rcsb.org/pdb/home/home.do	Experimentally determined three-dimensional structures
PRD [45]	http://pri.hgc.jp/	RPIs from 22 species, focusing on gene-level information
PRIDB [24]	http://pridb.gdcb.iastate.edu/	Interface information from RNA-protein complex structures in browsable and machine-readable format
RBPDB [8]	http://rbpdb.ccbr.utoronto.ca/	Experimental data on binding preferences and specificities of RBPs
RPIntDB	http://pridb.gdcb.iastate.edu/RPISeq/	RPIs from databases and high-throughput experiments in literature

Table 2: Databases of RNA-Protein Interactions and Interfaces.

of individual RBPs that were not included in the training set, the performance of the classifiers was highly variable across the RBPs. On average, the classifiers performed with an accuracy of only 50%. Using the pre-rRNA processing factor Nop15p as an example, the authors demonstrated that their method performs better when the training set includes at least some of the known mRNA targets for a given RBP. The authors acknowledge that the main limitation of this method is that it requires many features of both the RNA and protein under consideration. Although some of these features are easy to compute, some of them may not be available for other RNA-protein pairs of interest, and they are not trivial to obtain experimentally. Hence, the method may have limited applicability.

Also in 2011, the catRAPID method for predicting long non-coding RNA (lncRNA) partners of RBPs was published [17]. This study used a dataset consisting of 858 RNA-protein complexes extracted from the PDB [10]. Values for several physicochemical properties, including secondary structure propensities, hydrogen bonding propensities and van der Waals interaction propensities, were combined to calculate an interaction profile for each lncRNA and protein, which was then used to calculate interaction propensities for every potential lncRNA-protein pair. The interaction propensity of a RNA-protein pair in the training dataset was reported using the discriminative power (DP), which ranges between 0 and 1, with higher confidence interactions having higher DP values. The reported discriminative power on a non-redundant training set was 78%. The performance of catRAPID was also evaluated on independent test sets composed of positive interactions from the NPInter database of ncRNA-protein interactions [18], for which 89% prediction accuracy was reported [17]. However, when tested on 12,000 randomly generated RNA associations with proteins extracted from a non-Nucleic Acid-binding dataset [19], ~ 30% of these were predicted to interact with RNA [17]. In a recent study [20], the authors used catRAPID to investigate ribonucleoprotein interactions linked to neurodegenerative diseases. An advantage of the catRAPID algorithm is that it is the only published method that simultaneously predicts the binding sites in both RNA and protein sequences [21]. The catRAPID web server is available at http://service.tartaglialab.com/page/catrapid_group.

A purely sequence-based approach to predict RPIs, RPISeq, was proposed by our group, also in 2011 [22]. RPISeq is a family of machine learning classifiers (RF and SVM) designed to predict the probability of interaction between a given protein and RNA. In this method, RNA sequences are encoded as normalized frequencies of RNA tetrads, and protein sequences are encoded using a conjoint triad feature (CTF) method originally proposed by Shen et al. [23]. In essence, RPISeq exploits the amino acid composition of protein sequences and ribonucleotide composition of RNA sequences to predict the probability that a given pair (one protein and one RNA) will interact. On a non-redundant dataset of 2241 interacting pairs (RPI2241) created from known RNA-protein complexes in PRIDB [24], the RPISeq-RF classifier performed slightly better (average accuracy 89.6%), compared to the RPISeq-SVM classifier (average accuracy 87.1%). On an independent test set composed of only positive examples generated

from NPInter, the RPISeq-RF classifier correctly predicted 80.2% of interactions, while RPISeq-SVM predicted 66.3% of interactions. RPISeq's performance on an independent negative dataset was not reported. RPISeq's performance, using sequence information alone, was comparable to that of Pancaldi and Bähler's [15] method, which uses extensive feature information. An independent experimental validation of RPISeq predictions was published in a recent study [25], in which RPISeq was used to predict that the linc-UBC1 RNA interacts with PRC2 (Polycomb Repressive Complex 2). This prediction was experimentally validated using RNA immunoprecipitation, which confirmed that linc-UBC1 physically interacts with two core protein components of the PRC2 complex, EZH2 and SUZ12. RPISeq is available as a web server at http://pridb.gdcb.iastate.edu/RPISeq.

Another sequence-based method, similar to RPISeq, was proposed by Wang et al. [26]. This study also used the RPI2241 dataset [22], as one of the training datasets, a variation of the conjoint triad feature representation as protein descriptors and frequencies of RNA triads as RNA descriptors. The feature vector also included all combinations of protein and RNA descriptors. Only those features that were enriched in the training dataset were used as input for Naïve Bayes (NB) and Extended Naïve Bayes (ENB) classifiers (Supplementary Text S1). In cross-validation experiments using the RPI2241 dataset, the ENB classifier had a slightly better accuracy than the NB classifier (74% vs. 73%). The classifiers were also evaluated on known interactions from an independent dataset extracted from NPInter, with a reported predictive power of 79% (using the ENB classifier trained on RPI2241). In another experiment, the authors used a dataset of 30 ncRNAs and 759 proteins to predict RNA-protein interactions in *C. elegans*. They used an ncRNA pull-down experiment to validate these predictions for one selected ncRNA, sbRNA CeN72. The experiments identified 51 proteins that interact with CeN72. However, the ENB classifier predicted a total of 207 CeN72 interacting proteins (Supplemental Table S5 in [26]); of these, only 10 were true positive predictions. Although the authors claim that their method outperforms other existing methods, no evidence was presented to support this claim. In fact, as summarized in Supplementary Table S2, the published results demonstrate that RPISeq-RF [22] outperforms the ENB classifier [26].

In summary, except for Pancaldi and Bähler's approach [15], all of the methods discussed above use sequence information as the primary input to make predictions. This is a distinct advantage when making predictions on proteins or RNAs for which little information is available, other than the sequence. Also, every method except that of Pancaldi and Bähler [15], uses training data partly derived from three-dimensional structures of complexes in the PDB. Because the number of experimentally determined structures of RNA-protein complexes is relatively small and the PDB does not yet encompass all possible types of RNA-protein interactions, one should use caution when interpreting these predictions. A weakness of all of these predictors is the use of a negative dataset generated from random pairings of RNAs and proteins (in which many false negative examples may be included). Using real negative examples based on experimental interaction data would be desirable and would increase confidence in the predictions.

In conclusion, researchers interested in predicting RPIs are advised to compare results of more than one method. At present, only two of the methods described above are available as web-based servers (Table 1).

Web Servers for Partner Prediction

The catRAPID server (http://service.tartaglialab.com/page/catrapid_group) developed by Bellucci et al. [17] provides an estimate of the interaction propensities of given RNA and protein sequences. The output is displayed as a heat-map of interaction scores, with x and y axes representing the RNA and protein sequences, respectively. The overall interaction score and the corresponding discriminative power (predictive measure for binding) are also reported. This server provides another module called catRAPID strength that predicts the "strength" of a RNA-protein pair, by comparing its interaction propensity with the interaction propensities of a reference set of 100 proteins and 100 RNAs.

The RPISeq web server (http://pridb.gdcb.iastate.edu/RPISeq) implements the RPISeq method developed by Muppirala et al. [22]. RPISeq takes as input a pair of RNA and protein sequences, and outputs the interaction probability computed by SVM and RF classifiers trained using the RPI2241 dataset. It also accepts batch submission of multiple proteins or RNAs. Currently, users can input a maximum of 100 sequences. This limitation can be overcome by using a stand-alone version of the program, which is freely available from the authors.

RNA-Protein Interface Prediction Methods

So far, we have discussed computational methods for predicting the likelihood that a given RNA-protein pair will interact. Understanding how individual RNAs and proteins specifically recognize each other is an important aspect of this problem, and requires characterization of interfacial contacts at the residue and atomic level. As a step toward deciphering the rules that govern recognition specificity in RNA-protein interfaces, many computational methods (both sequence-based and structure-based) have been developed for predicting RNA-binding residues in proteins. Three recent reviews have summarized and compared these methods [21,27,28], which we will not reconsider here. With one exception, all published methods for predicting RNA-binding residues in a protein of interest do not take into account the specific RNA partner with which it interacts (i.e. they are or "partner-agnostic" or "non-partner specific" methods. Here, we will focus instead on methods that are "partner-aware" or "partner-specific." For protein-protein complexes, the partner-specific approach has been shown to provide improved interface predictions over non-partner specific methods in several studies (e.g. [29,30]).

The first partner-specific RNA-binding residue prediction method was proposed by the Han et al. [31,32]. In this work, both protein and RNA features were used as input to an SVM classifier to predict RNA-binding residues. Length and amino acid composition of the protein, along with features such as solvent accessible surface area and interaction propensity of an amino acid triplet were used to encode the input protein. The input RNA was encoded as a 4 element vector representing the sum of the normalized position of each ribonucleotide in the RNA sequence. In 5-fold cross-validation experiments on a dataset of 3,149 RNA-protein interacting pairs, prediction accuracy was 84%, with a correlation coefficient (CC) 0.41. On an independent dataset comprising 267 RPIs, accuracy was 90%, with CC of 0.24 [32]. Comparison with non-partner specific methods on the same datasets showed that the performance of the partner-specific approach was superior in terms of CC, and comparable in terms of overall accuracy. It seems likely that using more descriptive features to encode the sequence of the RNA partner could provide improved performance.

A second partner-specific prediction method for identifying binding sites in both the protein and RNA partners of an interacting pair is catRAPID [17]. As discussed above, catRAPID predicts

interaction partners based on the interaction propensities of individual residues [17]. In several cases, catRAPID binding site predictions correlate well with experimental results [20,33], but the performance of this method has not been evaluated systematically on benchmark datasets. Therefore, it is difficult to comment on the relative accuracy of this method in predicting interfacial residues in either RNA or protein sequences.

Sequence and Structural Motifs in RNA-Protein Interfaces

Structural analyses of RNA-protein complexes and sequence data from high-throughput RNA-protein interaction experiments have led to a rapid expansion in the collections of structural and sequence motifs associated with interfaces in RNA-protein complexes. Databases of protein motifs (e.g. ProSite [34]) and RNA motifs (e.g. FR3D [35]) are valuable resources for investigating recognition principles in RNA-protein interactions. In addition to their utility for identifying binding sites in novel proteins and RNAs, motifs can provide insight into the biological functions of protein or RNA families. Well-characterized RNA-binding motifs in proteins include the RNA recognition motif (RRM), the K-homology (KH) domain, the Pumilio/FBF (PUF) domain, and the double-stranded RNA-binding domain (dsRBD) (recently reviewed in [36]). The number of characterized RNA

Structural motifs are smaller, but include several well-studied examples, such as pseudo knots, tetra-loops and kink turns [37]. RNA sequence motifs that serve as recognition sites for RBPs have been identified using *in vitro* selection methods such as SELEX [38] and RNAcompete [39]. High-throughput approaches for capturing *in vivo* RNA-protein complexes by Tap-tagging and immunoprecipitation [16], or UV cross linking and immunoprecipitation of RNA-protein complexes combined with microarray or RNA-Seq analysis [11,12] have resulted in a dramatic increase in our understanding of recognition motifs in cellular RNAs. Experimental data from such studies have been analyzed to determine sequence and structural features of recognition motifs for RBPs using methods such as RNAcontext [40]. These data are now available in resources such as the RBPDB database [8], and in RBPMotif [41], a web server for identifying sequence and structure preferences of RBPs.

RNA-Protein Interaction Databases

At present, there is no single comprehensive database of RNA-protein interactions. Widely used databases that contain RNA-protein complexes, and/or interactions as part of a broader collection include structure databases, such as the PDB [10] and NDB [42], as well as interaction databases, such as BioGRID [43] and IntAct [44]. The Protein Data Bank (PDB) [10] is a comprehensive database of experimentally determined three-dimensional structures of macromolecules, including both proteins and nucleic acids. The Nucleic Acid Database (NDB) [42] contains experimental 3D structural information for nucleic acids, and includes both DNA-protein and RNA-protein complexes. BioGRID [43] is a curated database of protein interactions and genetic interactions from more than 45 model organisms. The IntAct database [44] primarily contains protein-protein interactions, although it also includes some protein-small molecule, protein-nucleic acid and protein-gene locus interactions. In the remainder of this section, several databases that focus on RNA-protein interactions are discussed. Table 2 provides URLs for these.

The first three databases discussed below, PRD [45], NPInter [18] and RPIntDB (http://pridb.gdcb.iastate.edu/RPISeq/)are collections of RNA-protein interaction partners. They focus on binary interactions between proteins and RNAs and do not provide residue or atomic level information about interfaces. Most interactions in these databases are extracted from results of low-throughput, or more recently, high-throughput experiments in published literature.

In contrast, PRIDB (http://pridb.gdcb.iastate.edu) [24] is a collection of interfaces in RNA-protein complexes, derived from experimentally determined structures deposited in the PDB. Databases similar to PRIDB, but not focused exclusively on RNA-protein complexes, include ProNIT (http://www.abren.net/pronit/) [46], which contains experimentally determined thermodynamic interaction data for protein-nucleic acid interactions; BIPA (http://mordred.bioc.cam.ac.uk/bipa)[47], the Biological Interaction Database for Protein-Nucleic Acid; and NPIDB (http://npidb.belozersky.msu.ru) [48], which also includes structural information for both DNA-protein and RNA-protein complexes, as well as several online tools for analysis.

The final database included in this section, RDPDB [7,8], is a recently expanded collection of RNA-binding proteins and their experimentally determined target RNAs. This database provides information about both RNA-protein interaction partners and their interfaces, with a focus on the RNA recognition preferences of individual RPBs.

PRD

PRD (http://pri.hgc.jp/) [45] is the most comprehensive database of RNA-protein interactions currently available. It contains more than 10,000 documented physical interactions between RNA and proteins. It includes interactions from BioGRID [43], IntAct [44], and the PDBj [49]. The PRD interaction data model is based on the HUPO POSI-MI model, and the database can be searched using 11 different fields (e.g. Gene ID, experiment, biological function), or using text keywords. Each interaction record contains information about both the protein and RNA involved, the experimental method used to detect the interaction, and references. Biological functions and information regarding binding sites are also provided, when available. Search results can be exported in PSI-MI XML files.

NPInter

NPInter (http://www.panrna.org/NPInter/index.php) [18] was the first database developed to collect experimentally determined functional interactions between ncRNAs and protein-related biomolecules (PRMs), i.e. proteins, mRNAs or genomic DNAs. Interactions involving tRNAs and rRNAs are not included. In 2006, NPInter contained 700 interactions from six model organisms. NPInter version 2.0, available in 2013, now contains more than 200,000 interactions from 18 different organisms. It classifies the interactions into eight categories: 'ncRNA binds protein', 'ncRNA regulates mRNA expression', 'ncRNA indirectly regulates a gene activity', 'ncRNA expression is regulated by protein', 'ncRNA affects protein activity', 'ncRNA activity is affected by protein', 'genetic interaction between ncRNA gene and protein gene' and 'other linkages'. Users can search NPInter by molecule type (ncRNA, miRNA, protein) by ID (NONCODE, miRBase, UniProt, PubMed), or using text queries. NPInter provides a BLAST option to query protein, ncRNA and miRNA sequences. Multiple download options are also provided.

RPIntDB

The RNA-Protein Interaction Data Base (RPIntDB), (http://pridb.

gdcb.iastate.edu/RPISeq/) was developed as a component of the RPISeq server [22]. The database includes experimentally validated RNA-protein interactions from several sources. It includes 11,815 proteins and 2,408 RNAs extracted from known RNA-protein complexes in PRIDB (as of March 2011), 242 ncRNAs and 282 proteins from ncRNA-protein interactions in the NPInter database [18], and 13,243 RPIs from high-throughput experiments published in literature [16]. Users can query RPIntDB to determine whether there is experimental evidence that a specific protein of interest is involved in an RPI. In the current version of RPIntDB, the service runs a BLAST search against the database and returns protein sequences that fall within a user-specified e-value threshold, along with their experimentally validated interacting RNA partners. The corresponding source(s) of the interaction are displayed in the output results.

PRIDB

The Protein-RNA Interface Database (PRIDB) http://pridb.gdcb.iastate.edu [24] is a comprehensive database of RNA-protein interfaces extracted from RNA-protein complexes in the PDB. It contains 16,350 proteins and 3,398 RNAs from 1,484 RNA-protein complexes (as of July 1 2013). PRIDB displays interfacial residues on protein and RNA sequences. It also displays known RNA-binding domains or motifs from ProSite [34] and RNA structural motifs from FR3D [35]. Atomic-level contact details for interfaces in the RNA-protein complexes can be visualized using an integrated JMol applet or downloaded in a machine-readable format. PRIDB also provides several reduced-redundancy benchmark datasets of RNA-binding protein chains.

RBPDB

The RNA-Binding Protein Database (RBPDB) (http://rbpdb.ccbr.utoronto.ca) [7,8] is a highly valuable compendium of experimentally determined RNA-binding specificities for RBPs from human, mouse, *D. melanogaster* and *C. elegans*. RBPDB contains target site preferences for more than 200 RBPs, extracted from almost 1,500 RNA-binding experiments. RBPDB catalogs data from 14 types of RNA-binding experiments and includes binding site sequence logos for more than 70 RBPs. The database can be searched by RBD, experiment type, species and gene name.

Future Directions

The emergence of high-throughput experimental approaches for interrogating RNA-protein interactions is generating a vast amount of new data, which will undoubtedly lead to improved computational methods for analyzing and predicting RNA-protein interfaces and interaction partners. Despite recent advances in both experimental and computational methodology, identifying the interaction partner(s) for a specific protein or RNA sequence is still an immensely challenging task. For example, even though the compendium of RNA-binding proteins and their targets published by the Hughes and Morris laboratories includes RBP recognition sites for more than 200 different RBPs [7], this impressive number corresponds to less than half of the known RBPs encoded in the human genome [8]. An analysis of the mRNA-bound proteome of a human kidney cell line identified ~ 800 bound proteins [14], nearly one third of which were not previously annotated as RNA-binding. With such large numbers of RPBs, each of which binds multiple mRNA and/or ncRNA targets, another difficult task will be to identify which combinations of RBPs determine specific post-transcriptional fates of individual mRNAs and ncRNAs. Progress in this direction was demonstrated in a quantitative proteomic analysis in *S. cerevisiae*, which identified sets of RBPs that bind simultaneously to common RNA targets [50]. Computational tools for constructing and interrogating RNA-protein interaction networks and for integrating RPIs into existing gene and protein interaction networks will be needed.

Obtaining high-resolution experimental structures of RNA-protein complexes is notoriously difficult and time consuming [51,52]. Thus, improved methods for computational modeling will be important for gaining insight into molecular details of interfaces in recalcitrant RNA-protein complexes. Algorithms for RNA-protein docking (not discussed in this review), although still somewhat naïve relative to those for small molecule and protein docking, are already benefitting from the increased availability of RNA-containing complex structures [53-55]. Finally, another important future direction in research on RNA-protein interactions is the rational design of RNA-protein interfaces. Engineered DNA binding proteins, such as ZFNs and TALENS, have become enormously powerful tools for genome engineering, and are poised to enter clinical settings [56-58]. Likewise, RNA-binding proteins engineered to recognize specific RNA sequences [36] could become valuable tools for manipulating post-transcriptional regulatory networks in the research laboratory, and potentially, important therapeutic agents for treating genetic and infectious diseases.

Acknowledgements and Funding

We thank Rasna Walia, Xue Li and Pete Zaback for suggestions and critical comments on the manuscript. This work was partially supported by funding from National Institutes of Health (GM066387 to D.D.).

References

1. Kishore S, Luber S, Zavolan M (2010) Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. Brief Funct Genomics 9: 391-404.
2. Licatalosi DD, Darnell RB (2010) RNA processing and its regulation: Global insights into biological networks. Nat Rev Genet 11: 75-87.
3. Singh R (2002) RNA-protein interactions that regulate pre-mRNA splicing. Gene Expr 10: 79-92.
4. Khalil AM, Rinn JL (2011) RNA-protein interactions in human health and disease. Semin Cell Dev Biol 22: 359-365.
5. Li Z, Nagy PD (2011) Diverse roles of host RNA binding proteins in RNA virus replication. RNA Biol 8: 305-315.
6. Zvereva AS, Pooggin MM (2012) Silencing and innate immunity in plant defense against viral and non-viral pathogens. Viruses 4: 2578-2597.
7. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, et al. (2013) A compendium of RNA-binding motifs for decoding gene regulation. Nature 499: 172-177.
8. Cook KB, Kazan H, Zuberi K, Morris Q, Hughes TR (2011) RBPDB: a database of RNA-binding specificities. Nucleic Acids Res 39: D301-308.
9. Mittal N, Roy N, Babu MM, Janga SC (2009) Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. Proc Natl Acad Sci U S A 106: 20300-20305.
10. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, et al. (2000) The Protein Data Bank. Nucleic Acids Res 28: 235-242.
11. Ankö ML, Neugebauer KM (2012) RNA-protein interactions in vivo: global gets specific. Trends Biochem Sci 37: 255-262.
12. König J, Zarnack K, Luscombe NM, Ule J (2012) Protein-RNA interactions: New genomic technologies and perspectives. Nat Rev Genet 13: 77-83.
13. Riley KJ, Steitz JA (2013) The "Observer Effect" in genome-wide surveys of protein-RNA interactions. Mol Cell 49: 601-604.
14. Baltz AG, Munschauer M, Schwanhäusser B, Vasile A, Murakawa Y, et al. (2012) The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts. Mol Cell 46: 674-690.
15. Pancaldi V, Bähler J (2011) In silico characterization and prediction of global

protein-mRNA interactions in yeast. Nucleic Acids Res 39: 5826-5836.

16. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO (2008) Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. PLoS Biol 6: e255.
17. Bellucci M, Agostini F, Masin M, Tartaglia GG (2011) Predicting protein associations with long noncoding RNAs. Nat Methods 8: 444-445.
18. Wu T, Wang J, Liu C, Zhang Y, Shi B, et al. (2006) NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. Nucleic Acids Res 34: D150-152.
19. Stawiski EW, Gregoret LM, Mandel-Gutfreund Y (2003) Annotating nucleic acid-binding function based on protein structure. J Mol Biol 326: 1065-1079.
20. Cirillo D, Agostini F, Klus P, Marchese D, Rodriguez S, et al. (2013) Neurodegenerative diseases: quantitative predictions of protein-RNA interactions. RNA 19: 129-140.
21. Cirillo D, Agostini F, Tartaglia GG (2013) Predictions of protein-RNA interactions. Wiley Interdiscip Rev Comput Mol Sci 3:161-175.
22. Muppirala UK, Honavar VG, Dobbs D (2011) Predicting RNA-protein interactions using only sequence information. BMC Bioinformatics 12: 489.
23. Shen J, Zhang J, Luo X, Zhu W, Yu K, et al. (2007) Predicting protein-protein interactions based only on sequences information. Proc Natl Acad Sci U S A 104: 4337-4341.
24. Lewis BA, Walia RR, Terribilini M, Ferguson J, Zheng C, et al. (2011) PRIDB: a Protein-RNA interface database. Nucleic Acids Res 39: D277-D282.
25. He W, Cai Q, Sun F, Zhong G, Wang P, et al. (2013) linc-UBC1 physically associates with polycomb repressive complex 2 (PRC2) and acts as a negative prognostic factor for lymph node metastasis and survival in bladder cancer. Biochim Biophys Acta 1832: 1528-1537.
26. Wang Y, Chen X, Liu ZP, Huang Q, Wang Y, et al. (2013) De novo prediction of RNA-protein interactions from sequence information. Mol Biosyst 9: 133-142.
27. Puton T, Kozlowski L, Tuszynska I, Rother K, Bujnicki JM (2012) Computational methods for prediction of protein-RNA interactions. J Struct Biol 179: 261-268.
28. Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, et al. (2012) Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 13: 89.
29. Ahmad S, Mizuguchi K (2011) Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. PLoS One 6: e29104.
30. Xue LC, Dobbs D, Honavar V (2011) HomPPI: A class of sequence homology based protein-protein interface prediction methods. BMC Bioinformatics 12: 244.
31. Shrestha R, Kim J, Han K (2008) Prediction of RNA-binding residues in proteins using the interaction propensities of amino acids and nucleotides. In: Advanced intelligent computing theories and applications, Springer-Verlag, Berlin, Heidelberg, Germany 114-121.
32. Choi S, Han K (2011) Prediction of RNA-binding amino acids from protein and RNA sequences. BMC Bioinformatics 12: S7.
33. Agostini F, Cirillo D, Bolognesi B, Tartaglia GG (2013) X-inactivation: Quantitative predictions of protein interactions in the Xist network. Nucleic Acids Res 41: e31.
34. Sigrist CJ, Cerutti L, de Castro E, Langendijk-Genevaux PS, Bulliard V, et al. (2010) PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 38: D161-166.
35. Sarver M, Zirbel CL, Stombaugh J, Mokdad A, Leontis NB (2008) FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. J Math Biol 56: 215-252.
36. Chen Y, Varani G (2013) Engineering RNA-binding proteins for biology. FEBS J 280: 3734-3754.
37. Fritsch V, Westhof E (2010) The architectural motifs of folded RNAs. The Chemical Biology of Nucleic Acids.
38. Tuerk C, Gold L (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science 249: 505-510.
39. Ray D, Kazan H, Chan ET, Peña Castillo L, Chaudhry S, et al. (2009) Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. Nat Biotechnol 27: 667-670.
40. Kazan H, Ray D, Chan ET, Hughes TR, Morris Q (2010) RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. PLoS Comput Biol 6: e1000832.
41. 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: W180-186.
42. Berman HM, Olson WK, Beveridge DL, Westbrook J, Gelbin A, et al. (1992) The nucleic acid database. A comprehensive relational database of three-dimensional structures of nucleic acids. Biophys J 63: 751-759.
43. Stark C, Breitkreutz BJ, Chatr-Aryamontri A, Boucher L, Oughtred R, et al. (2011) The BioGRID Interaction Database: 2011 update. Nucleic Acids Res 39: D698-704.
44. Kerrien S, Aranda B, Breuza L, Bridge A, Broackes-Carter F, et al. (2012) The IntAct molecular interaction database in 2012. Nucleic Acids Res 40: D841-846.
45. Fujimori S, Hino K, Saito A, Miyano S, Miyamoto-Sato E (2012) PRD: A protein-RNA interaction database. Bioinformation 8: 729-730.
46. Kumar MD, Bava KA, Gromiha MM, Prabakaran P, Kitajima K, et al. (2006) ProTherm and ProNIT: Thermodynamic databases for proteins and protein-nucleic acid interactions. Nucleic Acids Res 34: D204-D206.
47. Lee S, Blundell TL (2009) BIPA: A database for protein-nucleic acid interaction in 3D structures. Bioinformatics 25: 1559-1560.
48. Kirsanov DD, Zanegina ON, Aksianov EA, Spirin SA, Karyagina AS, et al. (2013) NPIDB: Nucleic acid-Protein interaction dataBase. Nucleic Acids Res 41: D517-D523.
49. Kinjo AR, Suzuki H, Yamashita R, Ikegawa Y, Kudou T, et al. (2012) Protein Data Bank Japan (PDBj): Maintaining a structural data archive and resource description framework format. Nucleic Acids Res 40: D453-D460.
50. Klass DM, Scheibe M, Butter F, Hogan GJ, Mann M, et al. (2013) Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in *Saccharomyces cerevisiae*. Genome Res 23: 1028-1038.
51. Ke A, Doudna JA (2004) Crystallization of RNA and RNA-protein complexes. Methods 34: 408-414.
52. Scott LG, Hennig M (2008) RNA structure determination by NMR. Methods Mol Biol 452: 29-61.
53. Tuszynska I, Bujnicki JM (2011) DARS-RNP and QUASI-RNP: new statistical potentials for protein-RNA docking. BMC Bioinformatics 12: 348.
54. Li CH, Cao LB, Su JG, Yang YX, Wang CX (2012) A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RNA docking decoys. Proteins 80: 14-24.
55. Huang Y, Liu S, Guo D, Li L, Xiao Y (2013) A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci Rep 3: 1887.
56. Joung JK, Sander JD (2013) TALENs: A widely applicable technology for targeted genome editing. Nat Rev Mol Cell Biol 14: 49-55.
57. Reyon D, Tsai SQ, Khayter C, Foden JA, Sander JD, et al. (2012) FLASH assembly of TALENs for high-throughput genome editing. Nat Biotechnol 30: 460-465.
58. Rahman SH, Maeder ML, Joung JK, Cathomen T (2011) Zinc-finger nucleases for somatic gene therapy: The next frontier. Hum Gene Ther 22: 925-933.