Sequence-based prediction of RNA-protein interactions

by

Rasna Rani Walia

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of
DOCTOR OF PHILOSOPHY

Major: Bioinformatics and Computational Biology

Program of Study Committee:
Vasant Honavar, Co-major Professor
Drena Dobbs, Co-major Professor
Susan Carpenter
Robert Jernigan
Edward Yu

Iowa State University
Ames, Iowa
2014

DEDICATION

To my parents. Your love, support, and encouragement have kept me going throughout my life.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGEMENTS

I would like to take this opportunity to express my thanks to those who helped me with various aspects of conducting research and the writing of this thesis. First and foremost, to my advisor, Dr. Vasant Honavar, for his guidance, patience and support throughout this research and the writing of this thesis. To my co-major advisor, Dr. Drena Dobbs, whose selfless time and care were sometimes all that kept me going. Her insights and words of encouragement have often inspired me and renewed my hopes for completing my graduate education. I would also like to thank my committee members for their efforts and contributions to this work: Dr. Susan Carpenter, Dr. Robert Jernigan, and Dr. Edward Yu.

Many people have played pivotal roles in my life during the past few years and I have drawn inspiration from each and every one of them at some point or the other. To Divya Mistry, the one person who has been a constant in my life during my graduate education, I say "Asante sana"! Thank you so much for your endless support and encouragement. A heartfelt thanks to Pooja Arora, Scott Boyken, Deepak Dibya, Yasser El-Manzalawy, Rafael Jordan, Ben Lewis, Harris Lin, Richa Madan, Carla Mann, Divita Mathur, Usha Muppirala, Olga Nikolova, Deepak Reyon, Rahul Roy, Kannan Sankar, Fadi Towfic, Chijioke Umunnakwe, Li Xue, and Pete Zaback, with whom I have had various discussions regarding research, work, and life. A special mention goes out to Trish Stauble, the BCB Program Coordinator, for all her efforts to make the life of her "children" manageable while they are in Graduate School. I would like to acknowledge sources of funding throughout my time at Iowa State: NIH, NSF, and the James Cornette Fellowship.

Last but not the least, I would like to take this opportunity to thank my parents, Dr. Dharminder Walia and Mrs. Roopmala Walia, for always believing in me and teaching me the value of hard work. My sisters, Ranjeeta Walia, Roshni Walia, Simran Kohli, and Namrata Kohli, have been the silver linings in my cloudy days. To Anthony Betz, you came into my life like a breath of fresh air and have provided endless love, support and encouragement, especially during the final stages of my research. Thank you!

ABSTRACT

The interaction of RNAs with proteins is fundamental for executing many of the key roles they play in living systems, including translation, post-transcriptional regulation of gene expression, RNA splicing, and viral replication. Recently, new roles for RNA-protein interactions have emerged, following the discovery that the human genome is pervasively transcribed and produces thousands of non-coding RNAs (ncRNAs). Although the functions of many ncRNAs are not yet known, one emerging theme is that long non-coding RNAs (lncRNAs) often drive the formation of ribonucleoprotein (RNP) complexes, which in turn influence the regulation of gene expression. Although the human genome is predicted to encode almost as many different RNA-binding proteins as DNA-binding transcription factors, our current understanding of the cellular roles of RNA-binding proteins, how they recognize their targets, and how they are regulated, lags far behind our understanding of transcription factors.

To improve our comprehension of RNA-protein recognition and the regulation of RNA-protein interaction networks within cells, this dissertation has four related goals: (i) performing a rigorous and systematic evaluation of sequence- and structure-based methods for predicting RNA-binding residues in proteins; (ii) developing improved method for predicting interfacial residues in RNA-binding proteins, using only sequence information; (iii) generating a comprehensive collection of protein-RNA interaction motifs (PRIMs); and (iv) developing improved methods for RNA-protein interaction partner prediction.

First, we present a systematic evaluation of state-of-the-art machine learning methods for predicting RNA-binding residues in proteins, using three carefully curated benchmark datasets and a rich set of data representations. We show that sequence-based methods trained using position-specific scoring matrices (PSSMs) perform better than structure-based methods, which use more complex features extracted from the 3D structures of proteins. Second, we present RNABindRPlus, a new method for predicting RNA-binding residues in proteins, using only sequence information. The predictor combines output from an optimized Support Vector Machine (SVM) classifier with the output from a novel homology-based method (HomPRIP). We show that RNABindRPlus performs better than all currently available methods for predicting interfacial residues in proteins. Third, we extract more than 30,000 unique RNA-protein interfacial motifs (RPIMs), consisting of contiguous residues from both the RNA and protein chains of characterized RNA-protein complexes. Lastly, we demonstrate the utility of RPIMs in predicting RNA-protein interaction partners. We employ them in an innovative and significantly improved method for partner prediction and show that it has both a high true positive rate and a much lower false positive rate than other available methods. Taken together, the results presented here provide important new insights into the determinants of RNA-protein recognition, in addition to valuable new software tools for interrogating and predicting RNA-protein complexes and interaction networks.

CHAPTER 1. OVERVIEW

The interaction of proteins with ribonucleic acid (RNA) is vital to many processes that impact life. For example, RNA is used as the genetic material of a wide variety of highly infectious and sometimes deadly RNA viruses (viruses that exclusively utilize RNA as their genetic material) and retroviruses (viruses whose genetic material must be "reverse transcribed" into DNA at some stage of their lifecycle). Worldwide, 34.2 million people carry the human immunodeficiency virus-1 (HIV-1), the retrovirus that causes AIDS, with 2.1 million new infections in 2013[1]. Influenza, an RNA virus, affects 3-5 million people and is responsible for 250,000-500,000 yearly deaths worldwide[2]. Although the life cycles of these viruses are relatively well characterized, there is still no effective vaccine against HIV-1. Specific binding interactions between RNA genomes of these viruses and the proteins that help replicate their genomes play pivotal roles in viral replication and infectivity. If we could specifically target and block key sites in the viral RNA genome required for viral particle production, we would be able to prevent the infectious diseases associated with such viruses.

More generally, RNA molecules play important roles in all phases of protein production in living cells [Hutvagner and Simard (2008)]. They encode the genetic message (messenger RNA) from the DNA to the ribosome, help catalyze the addition of amino acids to a growing peptide chain, and regulate gene expression through various post-transcriptional gene regulation processes [Licatalosi and Darnell (2010); Schwanhausser et al. (2011); Xue and Barna (2012); Keene (2010); Fu and Ares Jr (2014)]. Ribonucleoprotein (RNP) complexes play essential roles in gene expression through the interplay of messenger RNAs (mRNAs), non-coding RNAs (ncRNAs), and RNA-binding proteins (RBPs). RNPs within cells form extensive regulatory

[1] Report by the United Nations Programme on AIDS (UNAIDS)

[2] The World Health Organization's (WHO) statistics

networks interconnecting the transcriptome and proteome [Faoro and Ataide (2014)]. The interaction of RNAs with proteins is fundamental for executing many of the key roles both molecules play in living systems.

A recent discovery is that the human genome is pervasively transcribed and produces many thousands of ncRNAs [Qu and Adelson (2012); Rinn and Chang (2012); Geisler and Coller (2013)]. It is believed that biological complexity generally correlates with the proportion of the genome that is non-protein coding; while only 2% of the mammalian genome encodes mRNA, the vast majority is transcribed as “long” or “short” ncRNAs [Fatica and Bozzoni (2014); Liu et al. (2013); Prasanth and Spector (2007)]. The Encyclopedia of DNA Elements (ENCODE) project was charged with identifying all functional elements in the human genome sequence [Consortium (2012)]. Many novel non-protein-coding transcripts have been identified, with many of these overlapping protein-coding loci. Others are located in regions of the genome previously thought to be transcriptionally silent [Birney et al. (2007)]. These findings have motivated a surge in experimental interrogations of RNA regulatory networks. With the advent of high-throughput methods such as RIP-ChIP [Keene et al. (2006)] and HITS-CLIP [Darnell (2010)], it is now possible to experimentally determine pairwise associations between RNAs and proteins. These kinds of data are beginning to make it possible to construct the first RNA-protein interaction networks [Ascano et al. (2011); Imig et al. (2012)]. At present, however, most of the thousands of potential RNA-protein interactions remain under-determined and uncharacterized [Glisovic et al. (2008)]. This calls for effective and reliable computational tools to predict such interactions, help generate testable hypotheses, and prioritize expensive experimental investigations.

Many RNA-protein interactions involve recogntion of specific sequence and structural features of the RNA by proteins; others are non-specific [Auweter et al. (2006); Lunde et al. (2007)]. Understanding the sequence and structural determinants of RNA-protein interactions is important both for understanding their roles in biological networks and for developing therapeutic applications.

1.1 Dissertation Organization

The dissertation is divided into the following sections:

Chapter 1 provides an overview of the overall goal and specific aims of this dissertation research, including a concise literature review, and a description of the organization of the dissertation.

Chapter 2 is a published manuscript entitled "Protein-RNA Interface Residue Prediction using Machine Learning: An Assessment of the State of the Art" (*Walia, R., Caragea, C., Lewis, B., Towfic, F., Terribilini, M., El-Manzalawy, Y., Dobbs, D., and Honavar, V. BMC Bioinformatics (2012), 13(1):89*). In this study, state-of-the-art machine learning methods for predicting RNA-binding residues in proteins were systematically evaluated using three carefully curated non-redundant benchmark datasets. Three different data representations (amino acid sequence identity, position specific scoring matrices (PSSMs), and smoothed PSSMs) using either sequence- or structure-based windows were implemented. Vasant Honavar, Drena Dobbs, and I conceived the study and contributed to experimental design and writing. I carried out the implementation, experiments, and analysis of data, and prepared the initial draft of the manuscript. I also implemented the webserver. Michael Terribilini and Benjamin Lewis prepared the datasets and performed preliminary experiments. Cornelia Caragea, Fadi Towfic, and Yasser El-Manzalawy assisted with experiments and analysis of data.

Chapter 3 is a published manuscripted entitled "RNABindRPlus: A Predictor that Combines Machine Learning and Sequence Homology-Based Methods to Improve the Reliability of Predicted RNA-Binding Residues in Proteins" (*Walia, R., Xue, L., Wilkins, K., El-Manzalawy, Y. Dobbs, D. and Honavar, V. PLoS ONE (2014), 9(5): e97725*). The paper described a novel method for predicting RNA-binding residues in proteins using sequence information only. I conceived and designed the experiments with guidance from Vasant Honavar and Drena Dobbs. I performed the bulk of the experiments and analysis of data, with assistance from Li Xue, Katherine Wilkins, and Yasser El-Manzalawy. I prepared the initial draft of the manuscript. Drena Dobbs and Vasant Honavar edited the manuscript.

Chapter 4 is a manuscript to be submitted to Nucleic Acids Research entitled "Discovering

Interaction Motifs in the Interfaces of RNA-protein Complexes". The paper describes a new and systematic way of extracting and representing motifs at the interfaces of RNA-protein complexes. I conceived the study, carried out the experiments and analysis, and wrote the initial draft of the manuscript. Drena Dobbs and Vasant Honavar provided guidance throughout the project, contributed to experimental design, and edited the manuscript.

Chapter 5 is a manuscript to be submitted to Bioinformatics, entitled "Predicting RNA-Protein Interaction Partners using RNA-Protein Interaction Motifs". The paper describes a new sequence-based method to predict RNA-protein interaction partners. I conceived the study, created the test datasets, carried out the experiments and analysis, and wrote the initial draft of the manuscript. Drena Dobbs and Vasant Honavar contributed to experimental design, in addition to supervising the work and editing the manuscript.

Chapter 6 contains a summary of the contributions of this dissertation, and describes some directions for future work.

1.2 Experimental Methods to Identify RNA-Protein Interactions

The most definitive way of identifying RNA-protein interactions at the molecular level is to solve the three dimensional structures of RNA-protein complexes, using X-ray crystallography [Ke and Doudna (2004)] or Nuclear Magnetic Resonance (NMR) spectroscopy [Wu et al. (2005)]. Both methods provide structures at atomic resolution, and allow researchers to accurately identify RNA-binding residues in proteins as well as protein-binding residues in RNAs. However, many RNA-protein complexes are difficult to crystallize due to the difficulty in obtaining well-ordered crystals for X-ray diffraction analysis [Ke and Doudna (2004)], and NMR spectroscopy is limited to RNA-protein complexes with molecular weight lower than 40 kDa [Dominguez et al. (2011)]. Unlike X-ray crystallography, NMR spectroscopy can produce different models of RNA-protein complexes, which can be used to obtain some insight into their dynamics. Small-angle X-ray scattering (SAXS) is emerging as a complementary method to X-ray crystallography and NMR for studying structures of RNA-protein complexes at lower resolution. The method is especially useful for gaining insight into the flexible and disordered region of RNA-protein complexes [Rambo and Tainer (2010); Tuukkanen and Svergun (2014)].

In the absence of three dimensional structures of RNA-protein complexes, there are other experimental methods for identifying RNA-protein interactions [Glisovic et al. (2008); McHugh et al. (2014)]; both in terms of identifying interface residues in RNAs and proteins, and in providing partner information, i.e., is a particular protein likely to interact with a given RNA and vice-versa.

SELEX (systematic evolution of ligands by exponential enrichment) [Tuerk and Gold (1990)] and RNAcompete [Ray et al. (2009)] are both *in vitro* methods for elucidating the RNA-binding preferences of RBPs. A major disadvantage of these methods is that they don't take into account the cellular environment in which the RNA-protein interactions normally take place.

RIP (RNA Immunoprecipitation)-Chip [Keene et al. (2006)] and RIP-Seq are two variants of RIP-based methods to identify RNA targets of an RBP *in vivo*. In both cases, the RNP is immunoprecipitated from cell extracts using an antibody against the protein of interest. In RIP-Chip, the RNAs bound to the RBP are identified using mircroarray techniques after extraction and purification from the protein. In RIP-Seq, the RNAs are identified using high-throughput sequencing. The main disadvantages of RIP-Chip and RIP-Seq are that they don't allow identification of the actual binding sites in the RNAs to nucleotide resolution and tend to have a high signal-to-noise ratio as well as low stringency.

CLIP [Ule et al. (2005)] (Crosslinking immunoprecipitation)-based methods are very similar to RIP-based methods, except for an additional crosslinking step which attempts to overcome limitations of RIP-based methods. There are four main variants of CLIP-based methods: (i) HITS-CLIP [Darnell (2010)](also known as CLIP-Seq) uses high-throughput sequencing of RNA isolated by crosslinking immunoprecipitation. It utlizes in-vivo UV crosslinking technologies and next-generation sequencing; (ii) PAR-CLIP [Hafner et al. (2010a)] (photoactivatable ribonucleoside enhanced crosslinking and immunoprecipitation) uses the incorporation of photoreactive ribonucleoside analogues (e.g. 4-thiouridine and 6-thioguanosine) into RNA transcripts synthesized in living cells. This has the potential to increase the efficiency of crosslinking; (iii) iCLIP [Konig et al. (2010)] permits the identification of the actual crosslink site between the protein and the RNA with a resolution of one or a few nucleotides; and (iv) CRAC [Bohnsack et al. (2012)] (crosslinking and high-throughput cDNA analysis) relies on reverse transcription errors (substitutions and deletions) at crosslink sites to map contact sites. The biggest disadvantage of CLIP-based methods is the likely bias towards identifying sites that can be most efficiently cross-linked, which may not represent the entire RNA-binding landscape for any single RBP. However, these methods can reveal the sites of direct contact between RBPs and RNAs.

HiTS-RAP [Tome et al. (2014)] (high-throughput sequencing-RNA affinity profiling) is a recently developed method to measure interaction affinities quantitatively between RBPs and their associated RNAs. RIPiT-Seq [Singh et al. (2014)] (RNA:protein immunoprecipitation in tandem high-throughput sequencing) is another recently developed method that is suitable for studying multi-protein RNA complexes and for revealing targets and binding sites of compositionally dynamic RNPs.

Mass spectrometry [Aebersold and Mann (2003)] has been used to identify protein-protein interactions and map interface residues. It is now being adapted to identify the proteins that interact with RNA [Ascano et al. (2013); Butter et al. (2009); Klass et al. (2013); Schmidt et al. (2012); Scheibe et al. (2012)] and, in a few cases, has been used to identify interfacial residues [Kramer et al. (2014)]. Other molecular and biochemical methods for identifying and analyzing RNA-protein interactions include yeast 3-hybrid assays [Hook et al. (2005)], RNA affinity columns [Walker et al. (2008)], gel shift assays [Gagnon and Maxwell (2011)], and co-immunoprecipitation assays [Conrad (2008); Jedamzik and Eckmann (2009)].

The above mentioned methods provide ways to experimentally determine either the partners in an RNA-protein interaction, the specific binding sites on either the RBPs or RNAs in RNPs, or both. The methods all have their advantages and pitfalls (see [Riley and Steitz (2013)] for a good review). In addition to the PDB [Berman et al. (2000)] and NDB [Coimbatore Narayanan et al. (2014)], which store three-dimensional structural information about proteins, interacting macromolecules, and nucleic acids, databases such as doRiNA [Anders et al. (2011)], CLIPZ [Khorshid et al. (2010)], NPInter [Wu et al. (2006); Yuan et al. (2014)], and RBPDB [Cook et al. (2011)] collect and store information about RNA-protein interaction data determined experimentally. However, despite recent advances, there is a huge difference in numbers of known RNAs and RNPs and numbers that have been structurally characterized or shown to interact using high-throughput technologies. Computational methods provide viable and cost-effective approaches for predicting RNA-protein partners or RNA- and protein-binding residues in RNPs, and reducing the experimental search space for researchers.

1.3 Computational Prediction of RNA-Protein Interfaces

RNA-protein interactions are controlled by structural properties such as the shapes and charges of the interacting macromolecules, as well as high order structures [Iwakiri et al. (2012)]. Important features of RNA-protein interfaces include both sequence-based characteristics (i.e., derived only from the amino acid sequence of the proteins and/or ribonucleotides in RNA) and structure-based characteristics (derived from the secondary or tertiary structures of the proteins and/or RNA).

Several studies have shown that positively charged residues (more specifically, Arg and Lys) and polar residues (e.g. His) are preferred at RNA-binding sites in proteins [Bahadur et al. (2008); Ellis et al. (2007); Jones et al. (2001); Gupta and Gribskov (2011); Terribilini et al. (2006b); Treger and Westhof (2001)]. It is no surprise that positively charged residues are preferred as interfacial residues in RBPs; this gives them the ability to interact with the negatively-charged phosphate backbone of RNA. Polar residues such as Asp, Gln, and Ser are well suited to form hydrogen bonds with RNA. Hydrophobic residues are generally disfavored in RNA-binding interfaces, as are negatively charged and aliphatic residues [Gupta and Gribskov (2011)]. However, stacking interactions of ribonucleotide bases with aromatic rings of His and Tyr often occur in RNA-protein interfaces [Gupta and Gribskov (2011)].

Interfacial residues in RNA-binding proteins tend to be protruding compared to non-interfacial residues [Towfic et al. (2010)]. The surface shape of the protein and the secondary structure of the RNA molecule are both important in determining binding specificity. For example, “dented” protein surfaces are more likely to interact with unpaired nucleotides than paired ones [Iwakiri et al. (2012)] and at such interfaces, hydrogen bonds are prominent between the protein backbone and RNA bases.

The RNA-binding interfaces in RBPs are clusters of positively charged residues scattered on the protein surface, in contrast to DNA-binding interfaces, which tend to form continuous surface patches [Shazman and Mandel-Gutfreund (2008); Shazman et al. (2011)]. Also, within the primary sequence of RNA-binding proteins, interfacial residues tend to occur as short stretches of contiguous residues; these are often brought into close proximity in the folded protein, forming the clusters of postively charged residues on the surface [Terribilini et al. (2006b)].

The packing of RNA-protein interfaces is looser than that of protein-protein and DNA-protein complexes [Jones et al. (2001); Huang et al. (2013)]. This has been attributed to the more flexible and complex tertiary structures that RNA chains can form. The secondary structure states of both amino acid residues and ribonucleotides are important in RNA-protein interactions. In RNAs, unpaired bases are over-represented in protein binding regions. Unpaired bases are more accessible than paired bases for interactions with amino acids, and this is a major difference between RNA-protein and DNA-protein interactions [Gupta and Gribskov (2011)]. The idea that RNA-protein interfaces are loosely packed as compared to protein-protein and DNA-protein interfaces is supported by the over-representation of both unpaired bases and irregularly structured amino acids in RNA-binding regions [Gupta and Gribskov (2011)].

Computational methods for predicting RNA-binding residues in proteins (reviewed in [Perez-Cano and Fernandez-Recio (2010a); Puton et al. (2012); Walia et al. (2012)]) typically employ machine learning algorithms (e.g. Naïve Bayes, Random Forest, Support Vector Machine) that utilize several of the features mentioned above to discriminate between interfacial and non-interfacial regions. The methods can be broadly divided into: (i) sequence-based methods that encode information about the target residue and its neighboring sequence residues using features derived from the sequence of the protein, e.g., amino acid identity (which indirectly exploits information about the interface propensities of different amino acid residues [Terribilini et al. (2006b, 2007)]) or, position-specific scoring matrices (PSSMs), derived from multiple sequence alignments of homologous sequences [Kumar et al. (2008); Wang et al. (2008); Walia et al. (2012)]; (ii) structure-based methods that encode information about the target residue and its geometrical neighbors using features derived from the 3D-structure of the protein e.g. surface shape [Towfic et al. (2010)]; and (iii) hybrid methods, which combine sequence and structural features, e.g. combining PSSMs with solvent accessible surface area [Maetschke and Yuan (2009)].

In comparative studies, sequence-based methods that utilize evolutionary information in the form of position specific scoring matrices (PSSMs) have been shown to have superior performance to other sequence-based methods [Perez-Cano and Fernandez-Recio (2010a); Walia et al. (2012)]. Recently, several publications have appeared in which these methods have been used and their predictions have been independently validated by experimental biologists [Das et al. (2013); Qu et al. (2014)]. In constrast, predictors that rely on structure-derived features have limited applicability, because the number of solved RNA-protein structures lags far behind the number of sequences available for the same. Also, it is not clear how conformational changes in proteins upon binding RNA affect the performance of structure-based methods that are trained on features derived from datasets of proteins bound to RNA, instead of ideally looking at the apo-form of the protein [Puton et al. (2012)]. Perhaps surprisingly, we found that sequence-based methods actually perform better than structure-based methods for predicting RNA-binding residues [Walia et al. (2012) and Chapter 2 in this thesis]. To explore a possible explanation for this, we have evaluated the extent to which conformational changes occur in proteins upon RNA-binding [Appendix C].

Thus, sequence-based methods for making predictions of RNA-binding residues in proteins are of interest both because sequences are much more readily available than structures, and because sequence-based methods currently have competitive (if not superior) performance over structure-based methods. The latter finding suggests that RNA-binding amino acid sequence motifs are important for the specificity in RNA-protein interactions, and that these sequence motifs, as well as sequence motifs that include both amino acids from the protein *and* ribonucleotides in the RNA, could potentially be exploited for predicting partners in RNA-protein complexes or interaction networks. We explore this possibility in Chapter 5.

In contrast to the growing body of literature on predicting RNA-binding residues in proteins, only four publications have addressed the problem of predicting protein-binding residues in RNA [Bellucci et al. (2011); Choi and Han (2013); Gupta (2011); Muppirala (2013b)]. Gupta et al. [Gupta (2011)] developed an information theoretic model to predict protein-binding residues in RNAs. The method achieved an accuracy of about 60%. They also concluded that RNA structure is more important in distinguishing protein-binding sites from non-binding sites than RNA sequence. Muppirala et al. [Muppirala (2013b)] developed a method for predicting protein-binding sites on RNAs that utlizes RNA sequence motifs of length 5 or greater. The method doesn't achieve a very high correlation coefficient, but shows the utility of sequence features in making predictions of interacting residues on the RNA. Further research is needed to improve the prediction of protein-binding sites in RNAs.

1.4 RNA-Binding Domains and Motifs in Proteins

Some RNA-binding domains (RBDs) recognize their target sites mainly by their shape and geometry and others are sequence-specific but are sensitive to secondary structure context [Li et al. (2014)]. Most, but not all, RBPs contain RNA-binding domains (RBDs) that are needed for recognition of their RNA targets and for sequence specificity [Hogan et al. (2008)]. Some common RBDs include RNA recognition motifs (RRMs), K-homology (KH) domains, zinc finger domains, and double-stranded RNA-binding domains (dsRBDs) (see [Lunde et al. (2007)] for a good review). The presence of characterized RNA-binding domains in RBPs is most often indicative of a certain type of fold. For example, in RRM domains, the typical secondary structural topology is $\beta\alpha\beta\beta\alpha\beta$,where β represents a beta sheet and α an alpha-helical structure. At the primary sequence level, these domains are composed of stretches of consecutive amino acids, only a few of which directly interact with the RNA.

Several online databases, such as PROSITE [Sigrist et al. (2010, 2013)] and Pfam [Finn et al. (2014)], describe protein families, domains, and functional sites. PROSITE offers tools for protein sequence analysis and motif detection. Pfam provides a variety of information on different protein families, including multiple sequence alignments, and protein domain architectures.

1.5 Protein-Binding Domains and Motifs in RNA

Until recently, protein-binding domains in RNAs have received much less attention than RNA-binding domains in proteins. The high-througput methods (cited above) have changed that, and led to experimentally validated consensus sequences for target sites of approximately 50 different RNA-binding proteins, most of which bind mRNA. A recently developed method, RNAcompete [Ray et al. (2009)], can be used to determine the binding preferences of individual RBPs for a large number of RNA sequences. Consensus motifs generated from these data can be used to scan mRNA transcripts to identify potential RNA-binding sites using methods such as RNAcontext [Kazan et al. (2010)]. RBPmotif and CISBP-RNA Database [Ray et al. (2013)] catalogue binding sites of RNA-binding proteins derived from RNAcompete experiments. RBPmap [Paz et al. (2014)] is a new method that has been developed to accurately predict and map the binding sites of RBPs.

Also recently, impressive progress has been made in identifying and characterizing RNA structural motifs common to sets of related RNAs, e.g., kink turns, tetraloops, and pseudoknots. The RNA 3D Motif Atlas [Petrov et al. (2013)] and RNA Bricks [Chojnowski et al. (2014)] are databases that provide detailed information about RNA 3D motifs and their interactions, primarily with other RNA sequences.

1.6 Computational Methods for Predicting RNA-Protein Interaction Partners and Networks

With the recent advent of high-throughput *in vitro* and *in vivo* methods for identifying RNA-protein interactions, it has been possible to characterize the RNA partners for a rapidly expanding set of RNA-binding proteins [Konig et al. (2010, 2012); Hafner et al. (2010a); Granneman et al. (2009); Leung et al. (2011)] and to identify the protein partners of several RNAs [Faoro and Ataide (2014)]. These experimental methods have different advantages and disadvantages [Riley and Steitz (2013)]. Computational methods that can predict RNA-protein partners are thus viable and cost-effective approaches for reducing the experimental search space for basic researchers and for identifying potential targets for clinical intervention in genetic or infectious diseases.

To date, only a few computational methods have been developed for predicting RNA-protein interaction partners (reviewed in detail in [Muppirala (2013a); Cirillo et al. (2014)]).

Two recent methods developed after these reviews were published include: (i) lncPro [Lu et al. (2013)] and (ii) Oli as well as its various versions [Livi and Blanzieri (2014)]. lncPro uses matrix multiplication to encode feature vectors derived from the RNAs and proteins to predict association between long non-coding RNAs (lncRNAs) and proteins. Oli and its various derivatives are protein-specific predictors of mRNA binding, which use features derived from the RNA sequence (e.g., binding motifs and predicted secondary structures). Direct comparisons of methods for RNA-protein partner prediction on large datasets have been difficult to conduct, because most methods: (i) do not have webservers; or (ii) have webservers that cannot deal with large datasets/batch submissions; or (iii) do not have standalone versions of the methods to run on large datasets.

Most of the machine learning approaches developed for predicting RNA-protein interaction partners are supervised learning methods [Bellucci et al. (2011); Muppirala et al. (2011)]. Thus these methods suffer from the lack of negative data (i.e., validated examples of RBPs and the RNAs with which they do *not* interact) and consequently, tend to have high false positive rates. Muppirala et al. [Muppirala et al. (2011)] generated negative data for training their methods by randomly pairing RBPs with RNAs and ensuring that the resulting pairs did not appear in the positive test set. However, they did not include any negative examples in the test set. Belluci et al. [Bellucci et al. (2011)] and Lu et al. [Lu et al. (2013)] define interacting and non-interacting pairs by considering whether the atoms from the protein and RNA are within a certain distance from one another (The former use 7Å and the latter use 5Å). Methods end up testing their performance on positive data extracted from NPInter [Wu et al. (2006); Yuan et al. (2014)], which contains experimentally validated functional interactions between lncRNAs and other biomolecules. Since no negative data is included in the test set, it is not possible to reliably determine the false positive rates of the different methods. While it should be possible to improve these methods by extracting negative examples from raw RNA-Seq data obtained in high-throughput PAR-Clip or similar experiments, such an approach has not yet appeared in the literature. In future directions (Chapter 6), we propose using negative datasets extracted from high-throughput experiments. This improved dataset of experimentally validated negative examples should be valuable not only for extending the experiments presented in this dissertation, but also to the larger community of researchers investigating RNA-protein complexes and interaction networks.

1.7 Research Aims

The long term goal of this research is to improve our understanding of the molecular mechanisms of RNA-protein recognition and the regulation of RNA-protein interactions in cells. The primary goal of this work is to develop improved algorithms and tools for analyzing and predicting RNA-protein interactions.This work addresses both the “interface prediction” problem and the “partner prediction” problem through the following specific aims:

1. **To perform a rigorous and systematic evaluation of sequence- and structure-based methods for predicting RNA-binding residues in proteins**, by directly implementing and evaluating various methods using a common set of carefully curated benchmark datasets. This will provide insight into: (a) specific sequence and structural features that best differentiate interfacial residues from non-interfacial residues; (b) the comparative performance of sequence- versus structure-based methods.

2. **To develop improved methods for predicting interfacial residues in RNA-binding proteins, using sequence information alone**. This will allow us to take advantage of the wealth of RNA-binding protein sequences, without having to rely on the relatively sparse information available on the structures of RBPs.

3. **To generate a comprehensive collection of novel RNA-protein interaction motifs** (RPIMs) that contain elements from *both* the RNA and protein partners of RNA-protein complexes.

4. **To develop improved methods for predicting interaction partners in RNA-protein complexes and interaction networks** by exploiting the interaction motifs identified in Aim 3.

CHAPTER 2. PROTEIN-RNA INTERFACE RESIDUE PREDICTION USING MACHINE LEARNING: AN ASSESSMENT OF THE STATE OF THE ART

Paper originally appeared in BMC Bioinformatics, Vol 13, 2012[1]

Rasna R. Walia, Cornelia Caragea, Benjamin Lewis, Fadi Towfic, Michael Terribilini, Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar

Abstract

Background: RNA molecules play diverse functional and structural roles in cells. They function as messengers for transferring genetic information from DNA to proteins, as the primary genetic material in many viruses, as catalysts (ribozymes) important for protein synthesis and RNA processing, and as essential and ubiquitous regulators of gene expression in living organisms. Many of these functions depend on precisely orchestrated interactions between RNA molecules and specific proteins in cells. Understanding the molecular mechanisms by which proteins recognize and bind RNA is essential for comprehending the functional implications of these interactions, but the recognition 'code' that mediates interactions between proteins and RNA is not yet understood. Success in deciphering this code would dramatically impact the development of new therapeutic strategies for intervening in devastating diseases such as AIDS and cancer. Because of the high cost of experimental determination of protein-RNA interfaces, there is an increasing reliance on statistical machine learning methods for training predictors of RNA-binding residues in proteins. However, because of differences in the choice of datasets, performance measures, and data representations used, it has been difficult to obtain an accurate assessment of the current state of the art in protein-RNA interface prediction.

[1] Copyright retained by authors

Results: We provide a review of published approaches for predicting RNA-binding residues in proteins; and a systematic comparison and critical assessment of protein-RNA interface residue predictors trained using these approaches on three carefully curated non-redundant datasets. We directly compare two widely used machine learning algorithms (Naïve Bayes (NB) and Support Vector Machine (SVM)) using three different data representations in which features are encoded using either sequence or structure-based windows. Our results show that (i) Sequence-based classifiers that use a position-specific scoring matrix (PSSM)-based representation (PSSMSeq) outperform those that use an amino acid identity based representation (IDSeq) or a smoothed PSSM (SmoPSSMSeq); (ii) Structure-based classifiers that use smoothed PSSM representation (SmoPSSMStr) outperform those that use PSSM (PSSMStr) as well as sequence identity based representation (IDStr). PSSMSeq classifiers, when tested on an independent test set of 44 proteins, achieve performance that is comparable to that of three state-of-the-art structure-based predictors (including those that exploit geometric features), with respect to *Matthews Correlation Coefficient* (MCC) although the structure-based methods achieve substantially higher *Specificity* (albeit at the expense of *Sensitivity*) compared to sequence-based methods. We also find that the expected performance of the classifiers on a residue level can be markedly different from that on a protein level. Our experiments show that the classifiers trained on three different non-redundant protein-RNA interface datasets achieve comparable cross-validation performance. However, we find that the results are markedly affected by differences in the distance threshold used to define interface residues.

Conclusions: Our results demonstrate that protein-RNA interface residue predictors that use a PSSM-based encoding of sequence windows outperform classifiers that use other encodings of sequence windows. While structure-based methods that exploit geometric features can yield significant increases in the *Specificity* of protein-RNA interface residue predictions, such increase is offset by decreases in *Sensitivity.* Our results underscore the importance of comparing alternative methods using rigorous statistical procedures, multiple performance measures, and datasets that are constructed based on several alternative definitions of interface residues and redundancy cutoffs as well as including tests on independent test sets into the comparisons.

2.1 Background

RNA molecules play important roles in all phases of protein production and processing in the cell [Fabian et al. (2010); Hogan et al. (2008); Huntzinger and Izaurralde (2011); Licatalosi and Darnell (2010)]. They carry the genetic message from DNA to the ribosome, help catalyze the addition of amino acids to a growing peptide chain, and regulate gene expression through miRNA pathways. RNA molecules also serve as the genetic material of many viruses. Many of the key functions of RNA molecules are mediated through their interactions with proteins. These interactions involve sequence-specific recognition and recognition of structural features of the RNA by proteins, as well as non-specific interactions. Consequently, understanding the sequence and structural determinants of protein-RNA interactions is important both for understanding their fundamental roles in biological networks and for developing therapeutic applications.

The most definitive way to verify RNA-binding sites in proteins is to determine the structure of the relevant protein-RNA complex by X-ray crystallography or NMR spectroscopy. Unfortunately, protein-RNA complex structures have proven difficult to solve experimentally. Other methods for determining RNA-binding sites in proteins are costly and time consuming, usually requiring site-directed mutagenesis and low-throughput RNA-binding assays [Hellman and Fried (2007); Mills et al. (2007); Ule et al. (2005)]. Despite experimental challenges, the number of protein-RNA complexes in the PDB has grown rapidly in recent years (yet still lags far behind protein-DNA complexes). As of March 2012, there were 2,200 protein-DNA complexes in the Protein Data Bank [Berman et al. (2002)] and 1,186 protein-RNA complexes.

The difficulties associated with experimental determination of RNA-binding sites in proteins and their biological importance have motivated several computational approaches to these problems [Perez-Cano and Fernandez-Recio (2010a); Puton et al. (2012)]. Computational methods can rapidly identify the most likely RNA-binding sites, thus focusing experimental efforts to identify them. Ideally, such methods rely on readily available information about the RNA-binding protein, such as its amino acid sequence. Accurate prediction of protein-RNA interactions can contribute to the development of new molecular tools for modifying gene expression, novel therapies for infectious and genetic diseases, and a detailed understanding of molecular mechanisms involved in protein-RNA recognition. In addition to reducing the cost and effort of experimental investigations, computational methods for predicting RNA-binding sites in proteins may provide insights into the recognition code(s) for protein-RNA interactions. Several previous studies have analyzed protein-RNA complexes to define and catalog properties of RNA-binding sites [Ellis et al. (2007); Jeong et al. (2004b); Jeong and Miyano (2006b); Jones et al. (2001); Kim et al. (2003); Treger and Westhof (2001)]. These studies have identified specific interaction patterns between proteins and RNA and suggested sequence and structural features of interfaces that can be exploited in machine learning methods for analyzing and predicting interfacial residues in protein-RNA complexes.

Over the past 5 years, a large number of methods for predicting RNA-binding residues in proteins have been published [Chen and Lim (2008); Cheng et al. (2008); Huang et al. (2010); Jeong et al. (2004a); Jeong and Miyano (2006a); Kim et al. (2006); Terribilini et al. (2006b); Wang and Brown (2006a); Kumar et al. (2008); Tong et al. (2008); Wang et al. (2008); Maetschke and Yuan (2009); Spriggs et al. (2009); Liu et al. (2010); Li et al. (2010); Zhang et al. (2010b); Towfic et al. (2010); Ma et al. (2011); Wang et al. (2011); Chen et al. (2011)]. In these studies, a variety of sequence, structural, and evolutionary features have been used as input to different machine learning methods such as Naïve Bayes (NB), Support Vector Machine (SVM), and Random Forest (RF) classifiers [Perez-Cano and Fernandez-Recio (2010a); Puton et al. (2012)]. Most of the methods train classifiers or predictors which accept a set of residues that are sequence or structure neighbors of the target residue as input, and produce, as output, a classification as to whether the target residue is an interface residue. Such methods can be broadly classified into: (i) Sequence-based predictors that encode the target residue and its sequence neighbors in terms of sequence-derived features, e.g. identities of the amino acids, dipeptide frequencies, position-specific scoring matrices (PSSMs) generated by aligning the sequence with its homologs, or physico-chemical properties of amino acids; (ii) Structure-based predictors that encode the target residue and its spatial neighbors in terms of structure-derived features, e.g. parameters that describe the local surface shape; and (iii) Methods that use both sequence- and structure-derived features. The predictors not only differ in terms of the specific choice of the sequence or structure-derived features used for encoding the input, but also in the methods used (if any) to select a subset of features from a larger set of candidate features, and the specific machine learning algorithms used to train the classifiers, e.g. NB, SVMs and RF classifiers.

Identifying the relative strengths and weaknesses of the various combinations of machine learning methods and data representations is a necessary prerequisite for developing improved methods. However, because of differences in the criteria used to define protein-RNA interfaces, choice of datasets, performance measures, and data representations used for training and assessing the performance of the resulting predictors, as well as the general lack of access to implementations or even complete detailed descriptions of the methods, it has been difficult for users to compare the results reported by different groups. Consequently, most existing comparisons of alternative approaches, including that of Perez-Cano and Fernandez-Recio (2010a), rely on extrapolations of results obtained using different datasets, experimental protocols, and performance metrics. Implementations of some methods are only accessible in the form of web servers. Recently, Puton et al. (2012) presented a review of computational methods for predicting protein-RNA interactions, in which they compare the performance of multiple web servers that implement different sequence and/or structure-based predictors on a dataset of 44 RNA-binding proteins (RB44). Such a comparison is valuable for users because it identifies servers that provide more reliable predictions. Use of such servers to compare different methods or data representations can be problematic, however, because it is often impossible to definitively exclude overlap between the training data used for developing the prediction server and the test data used for measuring the performance of the server. In cases where one has access to the code used to generate data representations and implement the machine learning methods, it is possible to use statistical cross-validation to obtain rigorous estimates of the comparative performance of the methods. Comparison of performance of alternative methods based on published studies is fraught with problems, because of differences in the details of the evaluation procedures. For example, some studies use sequence-based cross-validation [Mitchell (1997)] where, on each cross-validation run, the predictor is trained on a subset of the protein sequences and tested on the remaining sequences. Other studies use window-based cross-validation, where sequence windows extracted from the dataset are partitioned into training and test sets used in cross-validation runs. Still others report the performance of classifiers on independent (blind) test sets. Window-based cross-validation has been shown to yield overly optimistic assessments of predictor performance because it does not guarantee that the training and test sequences are disjoint [Caragea et al. (2007a)]. Even when sequence-based cross-validation is used, the performance estimates can be biased by the degree of sequence identity shared among proteins included in the dataset. The lower the percentage of sequence identity, i.e., redundancy, allowed within the datasets, the smaller the number of sequences in the datasets and the harder the prediction problem becomes. While some studies have used reduced redundancy datasets, others have reported performance on highly redundant datasets. Taken together, all of these factors have made it difficult for the scientific community to understand the relative strengths and weaknesses of the different methods and to obtain an accurate assessment of the current state of the art in protein-RNA interface prediction.

Against this background, this paper presents a direct comparison of sequence-based and structure-based classifiers for predicting protein-RNA interface residues, using several alternative data representations and trained and tested on three carefully curated benchmark datasets, using two widely used machine learning algorithms (NB and SVM). We also compare the performance of the best sequence-based classifier with other more complex structure-based classifiers on an independent test set. The goal of this work is to systematically survey some of the most commonly used methods for predicting RNA-binding residues in proteins and to recommend methodology to evaluate machine learning classifiers for the problem. The main emphasis is the evaluation procedure of the different classifiers, i.e., the similarity of protein chains within the datasets used, the way in which cross-validation is carried out (sequence- versus window-based), and the performance metrics reported. Our results suggest that the PSSM-based encoding using amino acid sequence features outperforms other sequence-based methods. In the case of simple structure-based predictors, the best performance is achieved using a smoothed PSSM representation. Interestingly, the performance of the different classifiers was generally invariant across the three non-redundant benchmark datasets (containing 106, 144, and 198 protein-RNA complexes) used in this study. Implementation of the best performing sequence-based predictor is available at http://einstein.cs.iastate.edu/RNABindR/. We also make the datasets available to the community to facilitate direct comparison of alternative machine learning approaches to protein-RNA interface prediction.

2.1.1 Sequence-based Methods

Early work targeted towards the prediction of interface residues in complexes of RNA and protein was carried out by Jeong et al. (2004a), who implemented a neural network approach that used amino acid type and predicted secondary structure information as input. The dataset used by Jeong and Miyano contained 96 protein chains solved by X-ray crystallography with resolution better than 3Å and was homology-reduced by eliminating sequences that shared greater than 70% similarity over their matched regions. They defined a residue as an interaction residue if the closest distance between the atoms of a protein and its partner RNA was less than 6Å. Terribilini et al. (2006b, 2007) presented RNABindR, which used amino acid sequence identity information to train a Naïve Bayes (NB) classifier to predict RNA-binding residues in proteins. Interface residues were defined using ENTANGLE [Allers and Shamoo (2001)]. They generated and utilized the RB109 dataset (see Methods section) and used sequence-based leave-one-out cross-validation to evaluate classification performance.

Some studies [Jeong and Miyano (2006b); Kumar et al. (2008)] have shown that evolutionary information in the form of position-specific scoring matrices (PSSMs) significantly improves prediction performance over single sequence methods. For a given protein sequence, a PSSM gives the likelihood of a particular residue substitution at a specific position based on evolutionary information. PSSM profiles have been successfully used for a variety of prediction tasks, including the prediction of protein secondary structure [Jones (1999); Pollastri et al. (2002)], protein solvent accessibility [Garg et al. (2005); Nguyen and Rajapakse (2006)], protein function [Jeong et al. (2011)], disordered regions in proteins [Jones and Ward (2003)], and DNA-binding sites in proteins [Ahmad and Sarai (2005)]. Kumar et al. (2008) developed a support vector machine (SVM) model that was trained on 86 RNA-binding protein (RBP) chains and evaluated it using window-based five-fold cross-validation. This dataset of 86 RBPs contained no two chains with more than 70% sequence similarity with one another. A distance-based cutoff of 6Å was used to define interacting residues. Multiple sequence alignments in the form of PSSM profiles were used as input to the SVM classifier. Kumar et al. were able to demonstrate a significant increase in the prediction accuracy with the use of PSSMs. Wang and Brown (2006a,b) developed BindN, an SVM classifier that uses physico-chemical properties, such as hydrophobicity, side chain pK_a, and molecular mass, in addition to evolutionary information in the form of PSSMs, to predict RNA-binding residues. They evaluated the performance of their classifier by using PSSMs and several combinations of the physico-chemical properties, and found that an SVM classifier constructed using all features gave the best predictive performance. Their classifier was evaluated using window-based five-fold cross-validation. BindN+ was developed by Wang et al. (2010a) using PSSMs and several other descriptors of evolutionary information. They used an SVM classifier to build their classifier. The method was evaluated using window-based five-fold cross-validation. Cheng et al. (2008) introduced smoothed PSSMs in RNAProB to incorporate context information from neighboring sequence residues. In a smoothed PSSM, the score for the central residue i is calculated by summing the scores of neighboring residues within a specified window size (see Methods section for additional details). Cheng et al. evaluated their SVM classifier using window-based five-fold cross-validation and parameter optimization on the RB86 [Kumar et al. (2008)], RB109 [Terribilini et al. (2006b)] and RB107 [Wang and Brown (2006b)] datasets, all used in previous studies. Wang et al. (2011) have recently proposed a method that combines amino acid sequence information, including PSSMs and smoothed PSSMs, with physico-chemical properties and predicted solvent accessibility (ASA) as input to an SVM classifier. They utilized a non-redundant dataset of 77 proteins, derived from the RB86 dataset used by Cheng et al. (2008) and Kumar et al. (2008), by ensuring that no two protein chains shared a sequence identity of more than 25%. Interface residues were defined as those residues in the protein with at least one atom separated by $\leq$ 6Å from any atom in the RNA molecule. RISP [Tong et al. (2008)] is an SVM classifier that uses PSSM profiles for predicting RNA-binding residues in proteins. An amino acid was defined as a binding residue if its side chain or backbone atoms fell within a 3.5Å distance cutoff from any atom in the RNA sequence. Tong et al. evaluated their classifier using window-based five-fold cross-validation on the RB147 [Terribilini et al. (2007)] dataset. ProteRNA [Huang et al. (2010)] is another recent SVM classifier that uses evolutionary information and sequence conservation to classify RNA-binding protein residues. Sequence-based five-fold cross-validation on the RB147 dataset was used to evaluate performance. A study that used PSSM profiles, interface propensity, predicted solvent accessibility, and hydrophobicity as features to train an SVM classifier to predict protein-RNA interface residues was conducted by Spriggs et al. (2009). Their method, PiRaNhA, used a non-redundant dataset of 81 known RNA-binding protein (RBP) sequences. It should be noted that the dataset was only weakly redundancy reduced; protein chains with 70% sequence identity over 90% of the sequence length were included in the dataset. An interface residue was defined as any amino acid residue within 3.9Å of the atoms in the RNA. NAPS [Carson et al. (2010)] is a server which uses sequence-derived features such as amino acid identity, residue charge, and evolutionary information in the form of PSSM profiles to predict residues involved in DNA or RNA-binding. It uses a modified C4.5 decision tree algorithm. Zhang et al. (2010b) presented a method that uses sequence, evolutionary conservation (in the form of PSSMs), predicted secondary structure, and predicted relative solvent accessibility as features to train an SVM classifier. Performance was evaluated using sequence-based five-fold cross-validation. This study also analyzed the relationship between the various features used and RNA-binding residues (RBRs).

In summary, the primary differences among the methods listed above are: (i) sequence features used, (ii) classifier used, (iii) interface residue definitions, (iv) number of protein-RNA complexes and redundancy levels in the datasets, and (v) cross-validation technique. Interface residue definitions commonly vary between 3.5Å to 6Å. The datasets constructed range from those which contain protein chains that share no more than 70% sequence identity to more stringent collections which share no more than 25% sequence identity. Cross-validation is either window-based or sequence-based.

2.1.2 Structure-based Methods

Several structure-based methods for predicting RNA-binding sites in proteins have been proposed in literature. KYG [Kim et al. (2006)] is a structure-based method that 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. Several of the scores calculated are averages over residues located within a certain distance (structural neighbors). Amino acid residues were predicted to be interacting if the calculated scores were higher than a certain threshold. An interface residue was defined as an amino acid residue with at least one RNA atom within a distance of 7Å. Studies [Shazman and Mandel-Gutfreund (2008); Shazman et al. (2011)] have shown that structural properties such as surface geometry (patches and clefts) and the corresponding electrostatic properties, patch size, roughness, and surface accessibility can help to distinguish between RNA-binding proteins (RBPs) and non-RBPs as well as RNA-binding surfaces and DNA-binding surfaces. Chen and Lim Chen and Lim (2008) used information from protein structures to determine the geometry of the surface residues, and classified these as either surface patches or clefts. This was done using gas-phase electrostatic energy changes and relative conservation of each residue on the protein surface. After the identification of patches and clefts on the protein surface, residues within these RNA-binding regions were predicted as interface residues if they had relative solvent accessibilities $\geq 5\%$. OPRA [Perez-Cano and Fernandez-Recio (2010b)] is a method that calculates patch energy scores for each residue in a protein by considering energy scores of neighboring surface residues. The energy scores are calculated using interface propensity scores weighted by the accessible surface area (ASA) of the residue. Residues with better patch scores are predicted to be RNA-binding. In this study, interface residues were defined as those that had at least one amino acid atom within a distance of 10Å from any RNA atom. Zhao et al. (2010) introduced DRNA, a method that simultaneously predicts RBPs and RNA-binding sites. A query protein is structurally aligned with known protein-RNA complexes, and if the similarity score is higher than a certain threshold, then the query is predicted as an RBP. Binding energy is calculated using a DFIRE-based statistical energy function, to improve the discriminative ability to identity RBPs versus non-RBPs. The binding sites are then inferred from the predicted protein-RNA complex structure. Residues are defined as interface residues if a heavy atom in the amino acid is < 4.5Å away from any heavy atom of an RNA base.

A number of methods have incorporated structural information along with evolutionary information to predict RNA-binding sites. Maetschke and Yuan (2009) presented a method that uses an SVM classifier with a combination of PSSM profiles, solvent accessible surface area (ASA), betweenness centrality, and retention coefficient as input features. Performance was evaluated on the RB106 and RB144 datasets, which are slightly modified versions of the benchmark datasets created by Terribilini et al. (2006b, 2007). In the Maetschke and Yuan study, an interface residue is defined using a distance cutoff of 5Å. PRINTR [Wang et al. (2008)] is another method that uses SVMs and PSSMs to identify RNA-binding residues. The method was developed on the RB109 dataset using window-based seven-fold cross-validation. A combination of sequence and structure derived features was used, and the best performance was obtained by using multiple sequence alignments combined with observed secondary structure and solvent accessibility information. Li et al. (2010) built a classifier using multiple linear regression with a combination of features derived from sequence alone, such as the physico-chemical properties of amino acids and PSSMs, and structure derived features, such as actual secondary structure, solvent accessibility, and the amino acid composition of structural neighbors. Their method was evaluated using window-based six-fold cross-validation. A recent method proposed by Ma et al. (2011) used an enriched RF classifier with a hybrid set of features that includes secondary structure information, a combination of PSSMs with physico-chemical properties, a polarity-charge correlation, and a hydrophobicity correlation. A dataset of 180 RNA-binding protein sequences was constructed by excluding all protein chains that shared more than 25% sequence identity and proteins with fewer than 10 residues. Residues were defined as interacting based on a distance cutoff of 3.5Å. They tested the performance of their classifier using a window-based nested cross-validation procedure, where an outer cross-validation loop was used for model assessment and an inner loop for model selection. A method that encodes PSSM profiles using information from spatially adjacent residues and uses an SVM classifier as well as an SVM-KNN classifier was proposed by Chen et al. (2011). Interface residues were defined using a distance cutoff of 5Å. The performance of the method was tested using window-based five-fold cross-validation.

Towfic et al. (2010) exploited several structural features (e.g. roughness, CX values) and showed that an ensemble of five NB classifiers that combine sequence and structural features performed better than a NB classifier that only used sequence features. They trained their method, Struct-NB, on the RB147 dataset, and used sequence-based five-fold cross-validation. Struct-NB was trained using residues known to be on the surface of the protein. Liu et al. (2010) used a Random Forest (RF) [Breiman (2001)] classifier to predict RNA-binding residues in proteins by combining interaction propensities with other sequence features and relative accessible surface area derived from the protein structure. They defined a mutual interaction propensity between a residue triplet and a nucleotide, where the target residue is the central residue in the triplet. A dataset of 205 non-homologous RBPs was constructed to evaluate their method. Protein chains with greater than 25% and RNA chains with greater than 60% sequence identity were removed from their initial pool of 1,182 protein-RNA chains.

In summary, the structure-based methods described above differ in terms of the features, classifiers, interface residue definitions, datasets, and cross-validation techniques used. Interface residues are typically defined within a range of 3.5Å all the way up to 10Å. Features used include RNA-binding propensities of surface residues, geometry of surface residues, solvent accessible surface area, and secondary structure, among others.

2.1.3 Assessment of existing methods on standard datasets

In this study, we follow a direct approach for comparing different machine learning methods for predicting protein-RNA interface residues using a unified experimental setting (i.e., all methods are trained and evaluated on the same training and test sets). Therefore, our approach can address questions such as (i) Which feature representation is most useful for this prediction problem? (ii) How does feature encoding using sequence-based or structure-based windows compare in terms of performance? and (iii) Which machine learning algorithm provides the best predictive performance? In our experiments, we used three non-redundant benchmark datasets (RB106, RB144 and RB198; see Table 2.1) to compare several classifiers trained to predict RNA-binding sites in proteins using information derived from a protein's sequence, or its structure. Two versions of the datasets were constructed: (i) Sequence datasets, which contain all the residues in a protein chain, and (ii) Structure datasets, which contain only those residues that have been solved in the protein structure (see Methods section for details). The input to the classifier consists of an encoding of the target residue plus its sequence or spatial (based on the structure) neighbors. Each residue is encoded using either its amino acid identity or its PSSM (position-specific scoring matrix) profile obtained using multiple sequence alignment. In addition to the questions posted above, we also address to what extent (if any) the recent increase in the number of protein-RNA complexes available in Protein Data Bank (PDB) over the past 6 years contributes to improved prediction of RNA-binding residues.

Table 2.1 The number of interface and non-interface residues in the datasets used in this study

	RB106Seq	**RB106Str**	**RB144Seq**	**RB144Str**	**RB199Seq**	**RB199Str**
Non-interface residues	20,172	19,284	27,509	26,128	45,710	43,045
Interface residues	4,534	4,534	6,109	6,109	7,950	7,950

2.1.4 Assessment of methods on an independent test set

Several studies [Chen and Lim (2008); Kim et al. (2006); Perez-Cano and Fernandez-Recio (2010b); Zhao et al. (2010)] have incorporated structural information, such as interaction propensities of surface residues, geometry of the protein surface, and electrostatic properties, to predict RNA-binding residues. Because it is more difficult to implement some of these methods from scratch, we utilized an independent test set of 44 RNA-binding proteins [Puton et al. (2012)] to compare our best performing sequence-based method with results obtained by Puton et al. (2012) using the following structure-based methods: KYG [Kim et al. (2006)], OPRA [Perez-Cano and Fernandez-Recio (2010b)], and DRNA [Zhao et al. (2010)]. We also used information about surface residues to filter the results obtained by our best performing sequence-based method to directly compare this simple structure information with more complex structure-based methods.

2.2 Results and Discussion

For a rigorous comparison of classifiers trained to predict RNA-protein interfacial residues, we first used a sequence-based five-fold cross-validation procedure (see Methods). The input

Table 2.2 Abbreviations used in this manuscript for the six different encodings implemented in this study. NB (Naïve Bayes), SVM (Support Vector Machine)

Classifier	**Sequence**	**Sequence PSSM**	**Smoothed Sequence PSSM**	**Structure**	**Structure PSSM**	**Smoothed Structure PSSM**
NB	IDSeq NB	PSSMSeq NB	SmoPSSMSeq NB	IDStr NB	PSSMStr NB	SmoPSSMStr NB
SVM with Linear Kernel	IDSeq LK	PSSMSeq LK	SmoPSSMSeq LK	IDStr LK	PSSMStr LK	SmoPSSMStr LK
SVM with Radial Basis Function Kernel	IDSeq RBFK	PSSMSeq RBFK	SmoPSSMSeq RBFK	IDStr RBFK	PSSMStr RBFK	SmoPSSMStr RBFK

for each classifier consists of an encoding of the target residue plus its sequence or spatial neighbors. Each residue is encoded using its amino acid identity, its PSSM (position-specific scoring matrix) profile obtained using multiple sequence alignment, or its smoothed PSSM profile. We refer to the classifiers that rely exclusively on sequence as "sequence-based" and those that use structural information only to identify spatial neighbors as "simple structure-based" to distinguish them from structure-based methods (discussed below) that exploit more complex structure-derived information, such as surface concavity or other geometric features. Here, we considered 6 different encodings (see Table 2.2). IDSeq and IDStr encode each amino acid and its sequence or structural neighbors, respectively, using the 20-letter amino acid alphabet; PSSMSeq and PSSMStr encode each amino acid and its sequence or structural neighbors respectively using their PSSM profiles; SmoPSSMSeq and SmoPSSMStr encode each amino acid and its sequence or structural neighbors by using a summation of the values of the PSSM profiles of neighboring residues and itself (see Methods section for details).

Tables 2.3, 2.4, and 2.5 compare performance based on the AUC (Area Under the receiver operating characteristic Curve) of the different feature encodings using three different machine learning classifiers: (i) Naïve Bayes (NB), (ii) Support Vector Machine (SVM) using a linear kernel (polynomial kernel with degree of the polynomial $p = 1$), and (iii) SVM using a radial

Table 2.3 Residue-based evaluation of Sequence Methods on Sequence Data. AUC (averaged over five folds) of sequence methods on sequence data using residue-based evaluation. For each dataset, the rank of each classifier is shown in parentheses. Based on average rank, the best sequence method is the SVM classifier that uses the RBF kernel and PSSMSeq as input. (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel)

	IDSeq NB	**IDSeq LK**	**IDSeq RBFK**	**PSSM Seq NB**	**PSSM Seq LK**	**PSSM Seq RBFK**	**Smo PSSM Seq NB**	**Smo PSSM Seq LK**	**Smo PSSM Seq RBFK**
RB106Seq	0.74 (7)	0.72 (9)	0.73 (8)	0.76 (4.5)	0.78 (2.5)	0.80 (1)	0.75 (6)	0.76 (4.5)	0.78 (2.5)
RB144Seq	0.73 (7.5)	0.72 (9)	0.73 (7.5)	0.74 (6)	0.79 (2.5)	0.80 (1)	0.75 (5)	0.77 (4)	0.79 (2.5)
RB198Seq	0.72 (8)	0.72 (8)	0.72 (8)	0.73 (6)	0.78 (2.5)	0.80 (1)	0.74 (5)	0.77 (4)	0.78 (2.5)
Average	**0.73 (7.5)**	**0.72 (8.7)**	**0.73 (7.8)**	**0.74 (5.5)**	**0.78 (2.5)**	**0.80 (1)**	**0.75 (5.3)**	**0.77 (4.2)**	**0.78 (2.5)**

basis function (RBF) kernel, using residue-based evaluation (See Methods section for details). For each dataset, the rank of each classifier is shown in parentheses. The last row in each table summarizes the average AUC and rank for each classifier. Table 2.3 shows a comparison of the AUC of the different sequence-based methods across the three different sequence datasets (RB106Seq, RB144Seq, and RB198Seq). Following the suggestion of Demšar (2006), we present average ranks of the classifiers to obtain an overall assessment of how they compare relative to each other. Based on average rank alone, an SVM classifier that uses the RBF kernel and PSSMSeq encoding, which has an average AUC of 0.80, outperforms the other methods. Table 2.4 shows a comparison of the AUC of the different sequence-based methods on the structure datasets (RB106Str, RB144Str, and RB198Str). The best performance across all three structure datasets by a sequence-based method is achieved by an SVM classifier using the RBF kernel, which obtains an average AUC of 0.81, using the PSSMSeq encoding. A comparison of the simple structure-based methods on the structure datasets is shown in Table 2.5. The best performing method uses the SmoPSSMStr encoding (with a window size of 3) as input to an SVM classifier constructed with the RBF kernel, achieving an average AUC of 0.80. Tables

Table 2.4 Residue-based evaluation of Sequence Methods on Structure Data. AUC (averaged over five folds) of sequence methods on structure data using residue-based evaluation. For each dataset, the rank of each classifier is shown in parentheses. Based on average rank, the best sequence method is the SVM classifier that uses the RBF kernel and PSSMSeq as input. (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel)

	IDSeq NB	**IDSeq LK**	**IDSeq RBFK**	**PSSM Seq NB**	**PSSM Seq LK**	**PSSM Seq RBFK**	**Smo PSSM Seq NB**	**Smo PSSM Seq LK**	**Smo PSSM Seq RBFK**
RB106Str	0.74 (7.5)	0.73 (9)	0.74 (7.5)	0.76 (5.5)	0.78 (3)	0.81 (1)	0.76 (5.5)	0.77 (4)	0.79 (2)
RB144Str	0.74 (7)	0.73 (9)	0.74 (7)	0.74 (7)	0.79 (2.5)	0.81 (1)	0.75 (5)	0.77 (4)	0.79 (2.5)
RB198Str	0.73 (7)	0.73 (7)	0.73 (7)	0.72 (9)	0.78 (3)	0.80 (1)	0.74 (5)	0.77 (4)	0.79 (2)
Average	**0.74 (7.2)**	**0.73 (8.3)**	**0.74 (7.2)**	**0.74 (7.2)**	**0.78 (2.8)**	**0.81 (1)**	**0.75 (5.2)**	**0.77 (4)**	**0.79 (2.2)**

2.6, 2.7, and 2.8 compare the AUC of the different feature encodings using the three different machine learning classifiers, using protein-based evaluation (See Methods for details). All AUC values obtained using the protein-based evaluation are lower than those obtained using residue-based evaluation. However, the average ranks of the top-performing methods, using the two evaluation methods, were equivalent. Protein-based evaluation returns lower AUC values than residue-based evaluations because the former is a more stringent measure of the performance of a classifier. These measures are reported as average values over a subset of protein families in the dataset. Poor performance on the more challenging members of the dataset is more apparent in this type of evaluation.

Table 2.9 shows the performance of the 6 top ranking sequence-based and simple structure-based methods on structure datasets. The feature encoding that gives best performance across all three structure datasets is PSSMSeq, when used as input to an SVM classifier that uses the RBF kernel. Figure 2.1 shows Receiver Operating Characteristic (ROC) curves and Precision vs Recall (PR) curves for the top performing methods on structure datasets. Notably, classifiers that utilize evolutionary information, i.e., PSSM profiles, have significantly better prediction

Table 2.5 Residue-based evaluation of Structure Methods on Structure Data. AUC (averaged over five folds) of structure methods on structure data using residue-based evaluation. Based on average rank, the best structure method across the three datasets is the SVM classifier that uses the RBF kernel and SmoPSSMStr (with a window size of 3) as input (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel) The rank of each classifier is shown in parentheses.

	IDStr NB	**IDStr LK**	**IDStr RBFK**	**PSSM Str NB**	**PSSM Str LK**	**PSSM Str RBFK**	**Smo PSSM Str NB**	**Smo PSSM Str LK**	**Smo PSSM Str RBFK**
RB106Str	0.76 (3.5)	0.75 (5.5)	0.76 (3.5)	0.71 (8.5)	0.75 (5.5)	0.74 (7)	0.71 (8.5)	0.78 (2)	0.80 (1)
RB144Str	0.77 (3.5)	0.76 (5)	0.77 (3.5)	0.71 (8)	0.75 (6)	0.74 (7)	0.70 (9)	0.79 (2)	0.80 (1)
RB198Str	0.76 (4)	0.76 (4)	0.76 (4)	0.70 (6.5)	0.74 (6.5)	0.74 (6.5)	0.67 (9)	0.78 (2)	0.79 (1)
Average	**0.76 (3.7)**	**0.76 (4.8)**	**0.76 (3.7)**	**0.71 (8.2)**	**0.75 (6)**	**0.74 (6.8)**	**0.69 (8.8)**	**0.78 (2)**	**0.80 (1)**

performance than classifiers that are trained using only the amino acid identities of the target residue and its sequence neighbors.

Supplementary Tables A.1 and A.2 (see Appendix A) highlights the similarities and differences between methods implemented in this study and existing methods in the field.

2.2.1 Representations based on sequence versus structural neighbors

The sequence-based classifiers, IDSeq and PSSMSeq, utilize a sliding-window representation to generate subsequences around residues that are contiguous in the protein sequence. In an attempt to capture the structural context for predicting RNA-binding sites, we constructed the IDStr and PSSMStr encodings which use the spatial neighbors (derived from the 3D structure) of an amino acid as input.

Comparison of the ROC curves of the IDSeq_NB and IDStr_NB classifiers (Figure 2.2a) on the structure dataset (RB144Str) shows that the performance of the IDStr_NB classifier is superior to that of the IDSeq_NB classifier. Similarly, the PR curve (Figure 2.2b) shows

Table 2.6 Protein-based evaluation of Sequence Methods on Sequence Data. AUC (averaged over five folds) of sequence methods on sequence data using protein-based evaluation. Based on average rank, the best sequence method is the SVM classifier that uses the RBF kernel and PSSMSeq as input. (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel) The rank of each classifier is shown in parentheses.

	IDSeq NB	**IDSeq LK**	**IDSeq RBFK**	**PSSM Seq NB**	**PSSM Seq LK**	**PSSM Seq RBFK**	**Smo PSSM Seq NB**	**Smo PSSM Seq LK**	**Smo PSSM Seq RBFK**
RB106Seq	0.69 (6.5)	0.68 (9)	0.69 (6.5)	0.72 (2)	0.71 (3.5)	0.74 (1)	0.69 (6.5)	0.69 (6.5)	0.71 (3.5)
RB144Seq	0.68 (7)	0.67 (9)	0.68 (7)	0.71 (4)	0.73 (2)	0.74 (1)	0.68 (7)	0.70 (5)	0.72 (3)
RB198Seq	0.68 (6.5)	0.67 (8.5)	0.68 (6.5)	0.69 (5)	0.72 (2.5)	0.74 (1)	0.67 (8.5)	0.70 (4)	0.72 (2.5)
Average	**0.68 (6.7)**	**0.67 (8.8)**	**0.68 (6.7)**	**0.71 (3.7)**	**0.72 (2.7)**	**0.74 (1)**	**0.68 (7.3)**	**0.70 (5.2)**	**0.72 (3)**

that the IDStr_NB classifier achieves a higher precision at any given level of recall than the IDSeq_NB classifier. The AUC, a good overall measure of classifier performance, is 0.77 for the IDStr_NB classifier compared to 0.74 for the IDSeq_NB classifier on the RB144Str dataset using a Naïve Bayes classifier. The use of spatial neighbors to encode amino acid identity effectively captures information about residues that are close together in the protein structure. It is possible that this encoding indirectly incorporates surface patch information, which leads to improved performance using the IDStr feature, for any choice of classifier.

It is interesting and somewhat surprising that the AUC for the PSSMStr_NB classifier is 0.71, which is lower than 0.74 of the PSSMSeq_NB classifier. This is possibly due to the fact that evolutionary information is encoded linearly in sequence. The use of sequence windows preserves such information while the use of spatial windows distorts this signal. Figure 2.3 shows ROC curves and PR curves for the PSSMSeq_RBF and PSSMStr_RBF SVM classifiers with a radial basis function (RBF) kernel on the RB144Str dataset. The ROC curve for the PSSMSeq_RBF classifier dominates that of the PSSMStr_RBF classifier at all possible classification thresholds. The PR curve also shows that the PSSMSeq_RBF classifier achieves

Table 2.7 Protein-based evaluation of Sequence Methods on Structure Data. AUC (averaged over five folds) of sequence methods on structure data using protein-based evaluation. For each dataset, the rank of each classifier is shown in parentheses. Based on average rank, the best sequence method is the SVM classifier that uses the RBF kernel and PSSMSeq as input. (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel)

	IDSeq NB	**IDSeq LK**	**IDSeq RBFK**	**PSSM Seq NB**	**PSSM Seq LK**	**PSSM Seq RBFK**	**Smo PSSM Seq NB**	**Smo PSSM Seq LK**	**Smo PSSM Seq RBFK**
RB106Str	0.69 (7.5)	0.69 (7.5)	0.70 (5.5)	0.72 (3)	0.72 (3)	0.74 (1)	0.68 (9)	0.70 (5.5)	0.72 (3)
RB144Str	0.68 (7)	0.67 (9)	0.68 (7)	0.71 (4)	0.73 (2)	0.74 (1)	0.68 (7)	0.70 (5)	0.72 (3)
RB198Str	0.69 (6)	0.68 (8)	0.69 (6)	0.69 (6)	0.72 (2.5)	0.73 (1)	0.67 (9)	0.70 (4)	0.72 (2.5)
Average	**0.69 (6.8)**	**0.68 (8.2)**	**0.69 (6.2)**	**0.71 (4.3)**	**0.72 (2.5)**	**0.74 (1)**	**0.68 (8.3)**	**0.70 (4.8)**	**0.72 (2.8)**

a higher precision for any given level of recall than the PSSMStr_RBF classifier. Similar results were seen for all classifiers on the IDSeq, IDStr, PSSMSeq, and PSSMStr features (see Tables 2.4 and 2.5).

2.2.2 PSSM profile-based encoding of a target residue and its sequence neighbors improves the prediction of RNA-binding residues

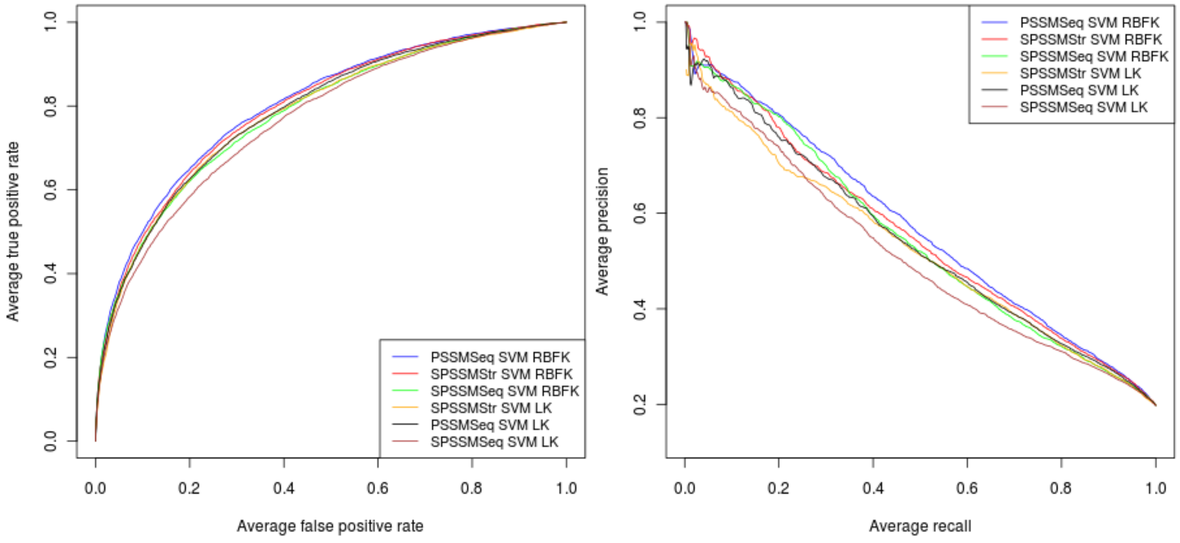
Sequence conservation is correlated with functionally and/or structurally important residues [Capra and Singh (2007); Fodor and Aldrich (2004); Friedberg and Margalit (2002)]. We incorporated information regarding sequence conservation of amino acids in our classifiers by using position-specific scoring matrix (PSSM) profiles. PSSMs have been shown to improve prediction performance in a number of tasks including protein-protein interaction site prediction [Kakuta et al. (2008)], protein-DNA interaction site prediction [Ahmad and Sarai (2005); Ofran et al. (2007)], and protein secondary structure prediction [Jones (1999); Pollastri et al. (2002)]. PSSMs have been previously shown to improve prediction of RNA-binding sites as well [Jeong and Miyano (2006b); Kim et al. (2006); Kumar et al. (2008); Maetschke and Yuan (2009);

Table 2.8 Protein-based evaluation of Structure Methods on Structure Data. AUC (averaged over five folds) of structure methods on structure data using residue-based evaluation. Based on average rank, the best structure method across the three datasets is the SVM classifier that uses the RBF kernel and SmoPSSMStr (with a window size of 3) as input (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel) The rank of each classifier is shown in parentheses.

	IDStr NB	**IDStr LK**	**IDStr RBFK**	**PSSM Str NB**	**PSSM Str LK**	**PSSM Str RBFK**	**Smo PSSM Str NB**	**Smo PSSM Str LK**	**Smo PSSM Str RBFK**
RB106Str	0.72 (3)	0.71 (7)	0.72 (3)	0.71 (7)	0.72 (3)	0.72 (3)	0.69 (9)	0.71 (7)	0.72 (3)
RB144Str	0.71 (6.5)	0.71 (6.5)	0.71 (6.5)	0.71 (6.5)	0.72 (3.5)	0.73 (2)	0.68 (9)	0.72 (3.5)	0.74 (1)
RB198Str	0.72 (3.5)	0.72 (3.5)	0.72 (3.5)	0.68 (8)	0.72 (3.5)	0.72 (3.5)	0.66 (9)	0.71 (7)	0.72 (3.5)
Average	**0.72 (4.3)**	**0.68 (5.7)**	**0.69 (4.3)**	**0.70 (7.2)**	**0.72 (3.3)**	**0.72 (2.8)**	**0.68 (9)**	**0.71 (5.8)**	**0.73 (2.5)**

Wang et al. (2008); Tong et al. (2008)].

In this work, we constructed Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers that utilize PSSM-based encoding of the target residue and its sequence or structural neighbors. The input to the classifiers is a window of PSSM profiles for the target residue and its neighbors in the sequence, in the case of the PSSMSeq classifier, or its spatial neighbors, in the case of the PSSMStr classifier. PSSM-based encoding dramatically improves prediction performance of sequence-based classifiers. Figure 2.4a shows the ROC curves of the IDSeq and PSSMSeq encodings with three different classifiers on the RB144Seq data. IDSeq_NB has an AUC of 0.73 and PSSMSeq_NB has an AUC of 0.74. The SVM classifier (built using a linear kernel) that used IDSeq has an AUC of 0.72 while the one that used PSSMSeq has an AUC of 0.79. The classifiers that used the PSSMSeq encoding also had a higher specificity for almost all levels of sensitivity (Figure 2.4b). Evolutionary information, as encoded by PSSMs, does not improve performance in the structure-based classifiers, as shown by the ROC curves in Figure 2.5a. On the RB144Str data, the SVM classifier built using an RBF kernel has an AUC of 0.74



(a)

(b)

Figure 2.1 Top 6 methods on structure datasets. (a) ROC curves and (b) PR curves for the top six methods on structure datasets

with PSSMStr, and an AUC of 0.77 with IDStr. Additionally, the precision of the IDStr_RBF encoding is higher for all levels of recall than the PSSMStr_RBF encoding, as shown in Figure 2.5b.

The main reason that information from multiple sequence alignments improves prediction accuracy is that it captures evolutionary information about proteins. Multiple sequence alignments reveal more information about a sequence in terms of the observed patterns of sequence variability and the locations of insertions and deletions [Kloczkowski et al. (2002)]. It is believed that more conserved regions of a protein sequence are either those that are functionally important [Lichtarge and Sowa (2002)] and/or are buried in the protein core directly influencing the three dimensional structure of the protein and variable sequence regions are considered to be on the surface of the protein [Jones (1999)]. In protein-RNA interactions, RNA-binding residues (RBRs) in proteins play certain functional roles and are thus likely to be more conserved than non-RBRs. A study by Spriggs and Jones (2009) revealed that RBRs are indeed more conserved than other surface residues.

Table 2.9 Top Six Methods on Structure Data using Residue-Based Evaluation. Comparison of AUC (averaged over five folds) of the top six methods on structure data using residue-based evaluation. Based on average rank, the best method across all three datasets is the SVM classifier that uses the RBF kernel and PSSMSeq as input. (NB - Naïve Bayes, SVM - Support Vector Machine, LK - Linear Kernel, RBFK - Radial Basis Function Kernel) The rank of each classifier is shown in parentheses.

	PSSMSeq RBFK	**Smo PSSMSeq RBFK**	**PSSMSeq LK**	**Smo PSSM-Seq LK**	**Smo PSSMStr RBFK**	**Smo PSSMStr LK**
RB106Str	0.81 (1)	0.79 (3)	0.78 (4.5)	0.77 (6)	0.80 (2)	0.78 (4.5)
RB144Str	0.81 (1)	0.79 (4)	0.79 (4)	0.77 (6)	0.80 (2)	0.79 (4)
RB198Str	0.80 (1)	0.79 (2.5)	0.78 (4.5)	0.77 (6)	0.79 (2.5)	0.78 (4.5)
Average	**0.80 (1)**	**0.79 (3.2)**	**0.78 (4.3)**	**0.77 (6)**	**0.80 (2.2)**	**0.78 (4.3)**

2.2.3 The predicted solvent accessibility feature does not improve performance of the classifiers

Spriggs et al. (2009) combined evolutionary information via PSSMs with the predicted solvent accessibility feature calculated using SABLE [Adamczak et al. (2005)]. They used an SVM classifier with an RBF kernel and optimized C and γ parameters to achieve the best AUC values. We performed an experiment to test whether the addition of the predicted solvent accessibility feature (calculated using SABLE with default parameters as in Spriggs et al. (2009)) would improve the performance of our NB classifier and SVM classifier trained using sequence information. This comparative experiment was performed on our smallest dataset, RB106Seq. We combined predicted solvent accessibility with amino acid identity (IDSeq), sequence PSSMs (PSSMSeq), and smoothed PSSMs (SmoPSSMSeq) using a window size of 25. Table 2.10 shows the average AUC values calculated from a sequence-based five-fold cross-validation experiment. We did not observe any difference in performance by adding the predicted solvent accessibility feature, which is consistent with the study conducted by Spriggs et al. (2009) in which they observed a slight improvement after adding predicted solvent accessibility. In the case of the SVM classifier (using an RBF kernel), adding the predicted accessibility feature to the SmoPSSMSeq feature actually led to a decrease in the AUC value. A possible reason for why addition of the

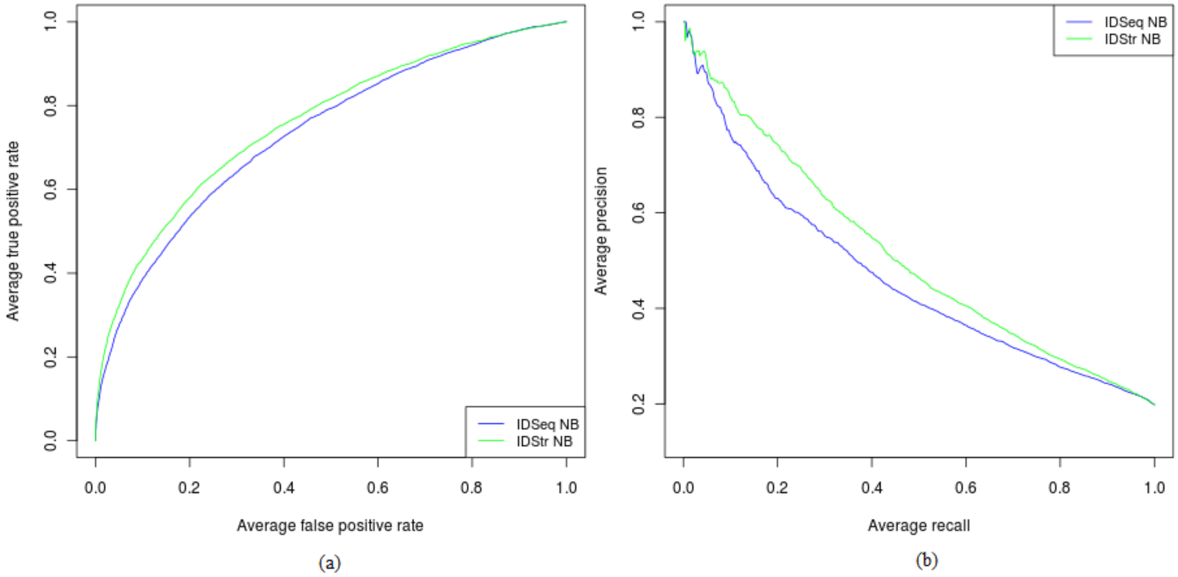


Figure 2.2 Naïve Bayes (NB) classifier on the IDSeq and IDStr features using the RB144Str dataset. (a) ROC curves and (b) PR curves of the NB classifier on the IDSeq and IDStr features. Both curves are generated using the RB144Str dataset.

predicted solvent accessibility feature did not lead to an improvement in performance for our datasets is that this information is already captured by the other features, such as PSSMs.

2.2.4 Classification performance has remained constant as the non-redundant datasets have doubled in size

We have attempted to exploit more information about protein-RNA interactions to improve prediction performance by generating a new larger dataset, RB198 (see Methods section), which includes recently solved protein-RNA complexes. The size of the non-redundant dataset has grown from 106 to 198 proteins (as of May 2010), as more complexes have been deposited in the PDB. In this study, we compared three non-redundant datasets, RB106, RB144, and RB198, derived from data extracted from the Protein Data Bank on June 2004, January 2006, and May 2010, using the same exclusion criteria of no more than 30% sequence identity between any two protein chains, and experimental structure resolution of ≤ 3.5Å. To investigate the effect of increasing the size of the non-redundant training set on prediction performance, we trained all of our classifiers on each of the three datasets and compared performance on each. Figure 2.6

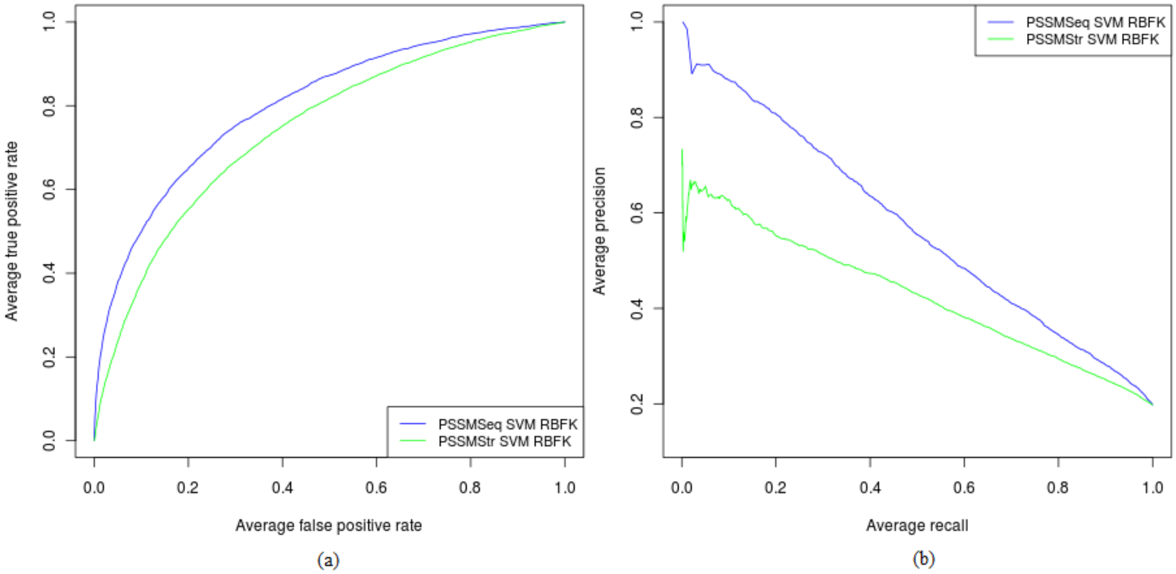


Figure 2.3 Support Vector Machine (SVM) classifier with radial basis function (RBF) kernel on the PSSMSeq and PSSMStr features using the RB144Str dataset. (a) ROC curves and (b) PR curves of the SVM classifier with an RBF kernel on the PSSMSeq and PSSMStr features. Both curves are generated using the RB144Str dataset.

shows the ROC curves for the Naïve Bayes classifier on sequence data for the IDSeq feature. The ROC curves for the RB106Seq, RB144Seq, and RB199Seq datasets, are nearly identical, with AUCs of 0.74, 0.74 and 0.73, respectively. Figure 2.7 shows the ROC and PR curves for the SVM classifier using an RBF kernel on structure datasets for the PSSMStr feature. These ROC curves are nearly identical, also, with AUCs of 0.74 for all three datasets. The PR curve shows that on the RB198Str dataset, the precision values are actually lower than those obtained using the two smaller datasets, RB106Str and RB144Str, for all values of recall.

Taken together, these results show that the prediction performance as estimated by cross-validation has not improved as the non-redundant dataset has doubled in size. There are several possible explanations for these observations: (i) we have reached the limits of predictability of protein-RNA interface residues using local sequence and simple structural features of interfaces; (ii) the data representations used may not be discriminative enough to yield further improvements in predictions; (iii) the statistical machine learning algorithms used may not be sophisticated enough to extract the information needed to improve the specificity and sensitivity of discrimination of protein-RNA interface residues from non-interface residues; (iv) the

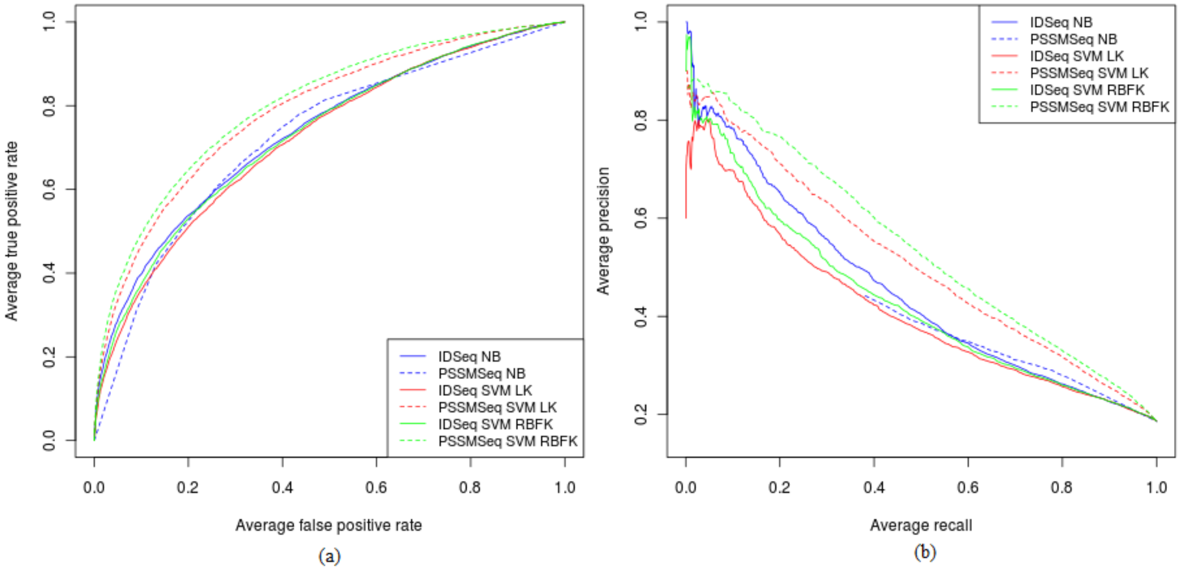


Figure 2.4 A comparison of the IDSeq and PSSMSeq features on the RB144Seq dataset. The PSSMSeq encoding leads to a better prediction performance compared to the IDSeq encoding. (a) ROC curves and (b) PR curves showing the difference between the IDSeq and PSSMSeq features across 3 different classifiers, Naïve Bayes (NB), Support Vector Machine (SVM) with linear kernel (LK), and SVM with radial basis function (RBF) kernel.

coverage of the structure space of protein-RNA interfaces in the available datasets needs to be improved before we can obtain further gains in the performance of the classifiers.

2.2.5 Comparisons with methods that use more complex structural information

In a recent review, Puton et al. (2012) evaluated existing web-based servers for RNA-binding site prediction, including three servers that exploit structure-based information, KYG [Kim et al. (2006)], DRNA [Zhao et al. (2010)], and OPRA [Perez-Cano and Fernandez-Recio (2010b)]. To facilitate direct comparison of our results with that study, we evaluated our best sequence-based method, PSSMSeq_RBFK, on the same dataset used in that study, RB44 which was shared with us by the Bujnicki lab where the Puton et al. Puton et al. (2012) study was carried out. Because our experiments employ a different distance-based interface residue definition (5Å instead of 3.5Å, see Methods), we calculated performance metrics using both definitions. We also calculated both residue-based and protein-based performance measures.

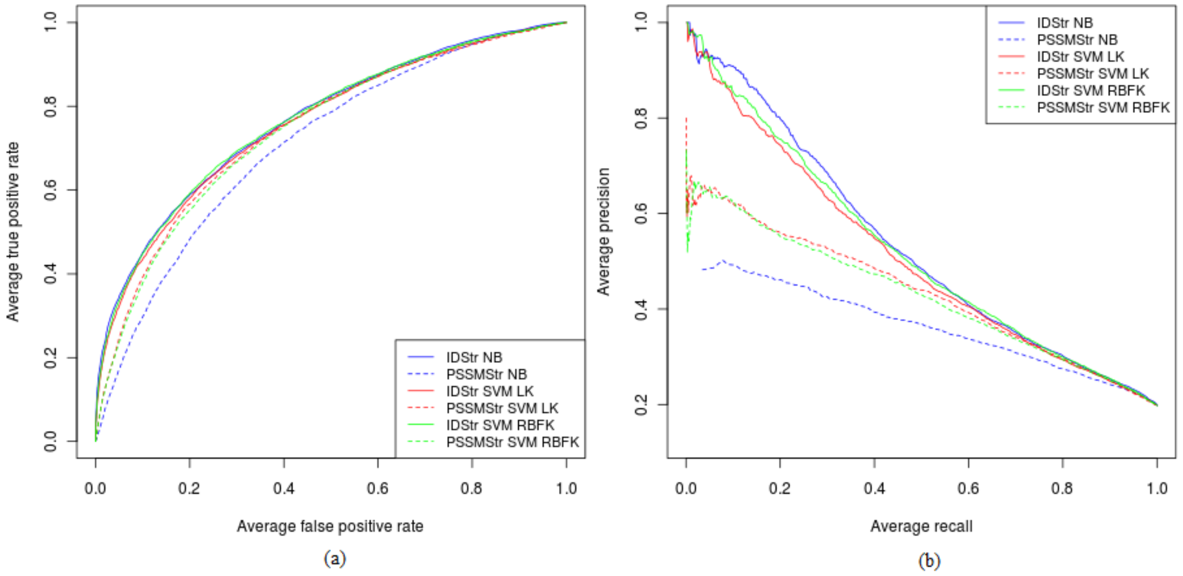


Figure 2.5 A comparison of the IDStr and PSSMStr features on the RB144Str dataset. The IDStr encoding leads to a better prediction performance compared to the PSSMStr encoding. (a) ROC curves and (b) PR curves showing the difference between the IDStr and PSSMStr features across 3 different classifiers, Naïve Bayes (NB), Support Vector Machine (SVM) with linear kernel (LK), and SVM with radial basis function (RBF) kernel.

Table 11 shows the performance of different methods on the RB44 dataset using residue-based evaluation. As shown in the table, when evaluated in terms of Matthews Correlation Coefficient (MCC), PSSMSeq_RBFK achieves performance comparable to or slightly lower than that of the structure-based methods: using the 3.5Å interface residue definition (3.5Å IRs), the MCC for PSSMSeq_RBFK is 0.33, whereas the MCC for the structure-based methods ranges from 0.30 - 0.38; using 5.0Å IRs, the MCC for PSSMSeq_RBFK is 0.38 compared with 0.36 - 0.42 for the structure-based methods.

Although the MCC is valuable as a single measure for comparing the performance of different machine learning classifiers, additional performance metrics such as Specificity and Sensitivity can be of greater practical importance for biologists studying protein-RNA interfaces. For example, a high value of Specificity indicates that a prediction method returns fewer false positives, thus allowing biologists to focus on a smaller number of likely interface residues for experimental interrogation. Among all of the methods compared in Table 11, DRNA [Zhao

Table 2.10 AUC values for different sequence-based features alone and in combination with predicted solvent accessibility. Comparison of AUC (averaged over five-folds) for different sequence features alone and in combination with predicted solvent accessibility (PA) on the RB106Seq dataset (NB - Naïve Bayes, SVM - Support Vector Machine, RBFK - Radial Basis Function Kernel).

Features	**NB**	**SVM RBFK**
IDSeq	0.74	0.73
IDSeq + PA	0.73	0.73
PSSMSeq	0.76	0.80
PSSMSeq + PA	0.76	0.80
SmoPSSMSeq	0.75	0.78
SmoPSSMSeq + PA	0.75	0.75

et al. (2010)] achieved the highest Specificity values of 0.75 (using 3.5Å IRs) and 0.94 (using 5.0Å IRs). Similarly, when protein-based evaluation is used (Table 12), DRNA [Zhao et al. (2010)] achieved the highest Specificity values of 0.94 (using 3.5Å IRs) and 0.98 (using 5.0Å IRs).

Among classifiers compared using 5.0Å IRs and residue-based evaluation, PSSMSeq-RBFK-Surface returned the best MCC of 0.42. PSSMSeq_RBFK_Surface takes predictions from PSSMSeq_RBFK and considers whether a predicted interface residue is a surface residue or not. If a residue is predicted as an interface residue but is not a surface residue, then it is marked as a non-interface residue (label = '0'). On the other hand, if it is a predicted interface residue and is also a surface residue, then the residue remains an interface residue. Surface residues were calculated using NACCESS [Hubbard and Thornton (1993)]. In our study, residues that have $>$ 5% relative accessible surface area (RSA) are defined as surface residues [Jones et al. (2001)]. PSSMSeq_RBFK_Surface achieved Specificity = 0.51 and Sensitivity = 0.78. KYG had similar performance, achieving Specificity = 0.55, Sensitivity = 0.67, and MCC = 0.41. In contrast, when classifiers are compared using 3.5Å IRs and residue-based evaluation, KYG and DRNA have the highest MCC of 0.38, consistent with the results published in Puton et al. (2012). However, PSSMSeq_RBFK has the highest Sensitivity of 0.84 followed by 0.83 for PSSMSeq_RBFK_Surface. Predictors that achieve high values of Sensitivity return fewer false negative values.

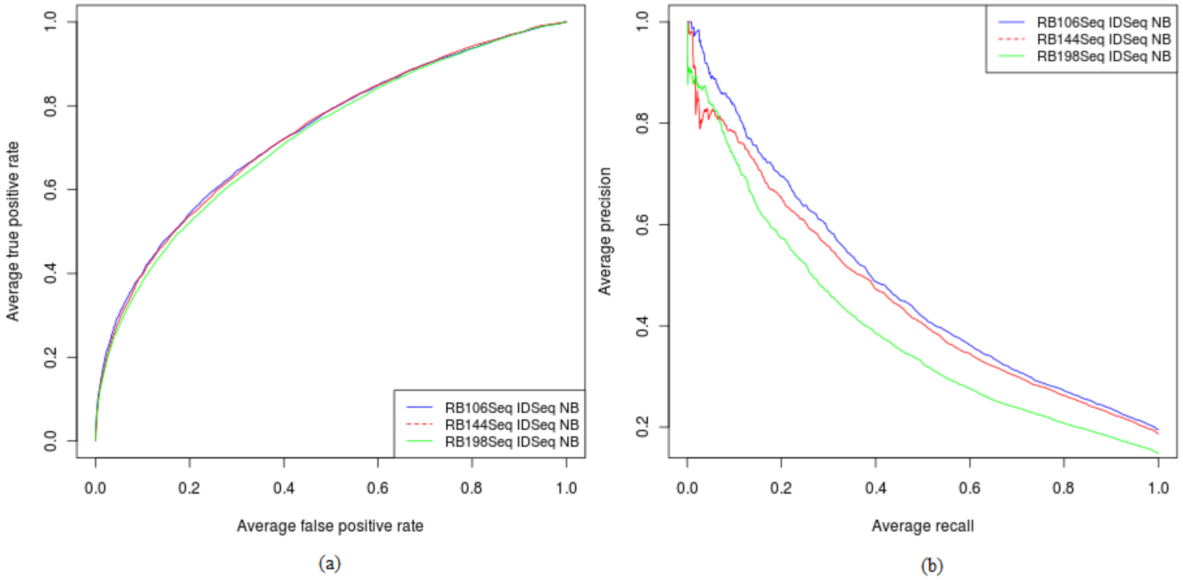


Figure 2.6 A comparison of the performance of the Naïve Bayes (NB) classifier on the 3 different sequence datasets using the IDSeq feature. (a) ROC curves and (b) PR curves showing the comparison of the performance of the NB classifier using the IDSeq feature on RB106Seq, RB144Seq, and RB198Seq datasets. Prediction performance has not improved as the non-redundant datasets have grown larger.

When we utilized protein-based evaluation and 5.0Å IRs, KYG returned the best MCC of 0.36. It achieved Specificity = 0.54, Sensitivity = 0.63, and Fmeasure = 0.56. PSSM-Seq_RBFK_Surface had similar performance, achieving MCC = 0.35, Specificity = 0.48, Sensitivity = 0.74, and Fmeasure = 0.57. On the other hand, when classifiers are compared using 3.5Å IRs, unlike the case of residue-based evaluation, DRNA does not emerge as a top method. It has low values of MCC = 0.19, Sensitivity = 0.23, and Fmeasure = 0.21. However, it has a Specificity of 0.94. This is because we assign Specificity = 1 in cases where there are zero true positive and false positive predictions (see Performance Measures for more details). The poor performance of DRNA can be explained by the fact that, in 32 out of the 44 proteins in the dataset, DRNA returns zero true positive and zero false positive predictions. In these cases, it returns a few false negative predictions and a much larger number of true negative predictions which result in an Fmeasure of 0 and MCC of 0 for 32 out of 44 proteins which pulls down the average performance over the 44 proteins to values that are considerably lower than their residue-based counterparts..

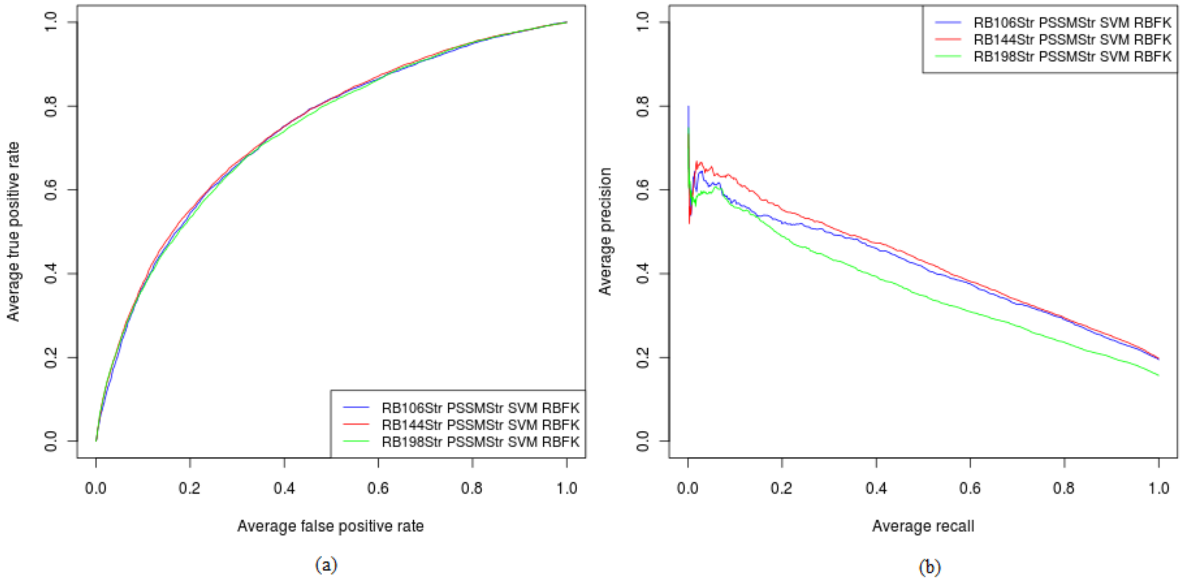


Figure 2.7 A comparison of the performance of the Support Vector Machine (SVM) classifier with a radial basis function (RBF) kernel on 3 different structure datasets using the PSSMStr feature. (a) ROC curves and (b) PR curves showing the comparison of the performance of the SVM classifier with an RBF kernel using the PSSMStr feature on RB106Str, RB144Str, and RB198Str datasets.

Taken together, these results indicate that the performance of different methods is affected by the type of evaluation procedure used, i.e., residue-based or protein-based evaluation. Generally, the performance of the sequence-based classifier, PSSMSeq_RBFK, and the simple structure-based classifier, PSSMSeq_RBFK_Surface, is comparable to that of several structure-based methods that exploit more complex structure features, when evaluated based on MCC. They also outperform structure-based methods in terms of Sensitivity, at the cost of Specificity.

An unexpected result of these studies is the finding that the interface residue definition can have a substantial impact on the performance of methods for predicting RNA-binding sites in proteins. For all of the methods compared in Table 2.11 and Table 2.12, using a 5Å instead of 3.5Å definition resulted in an increase in MCC, and Specificity, with a decrease in Sensitivity. Moreover, the differences in performance between methods compared using the same interface residue definition, are substantially smaller than the differences in performance obtained for a single method, using different interface residue definitions. Thus, the interface residue definition is an important factor that must be taken into consideration when comparing different methods for predicting RNA-binding residues.

Table 2.11 Residue-based evaluation of Methods on the RB44 Dataset. Performance measures computed on the RB44 dataset (IR - Interface Residue definition in Å).

Method	IR	Specificity	Sensitivity	Fmeasure	MCC
PSSMSeq_RBFK	3.5	0.33	0.84	0.47	0.33
	5.0	0.47	0.80	0.59	0.38
PSSMSeq_RBFK Surface	3.5	0.36	0.83	0.51	0.37
	5.0	0.51	0.78	0.62	0.42
KYG	3.5	0.40	0.73	0.52	0.38
	5.0	0.55	0.67	0.60	0.41
DRNA	3.5	0.75	0.27	0.40	0.38
	5.0	0.94	0.23	0.37	0.39
OPRA	3.5	0.40	0.54	0.45	0.30
	5.0	0.57	0.51	0.54	0.36

Table 2.12 Protein-based evaluation of Methods on the RB44 Dataset. The values reported below are averages over the 44 proteins in the dataset (IR - Interface Residue definition in Å).

Method	IR	Specificity	Sensitivity	Fmeasure	MCC
PSSMSeq_RBFK	3.5	0.32	0.80	0.44	0.27
	5.0	0.45	0.76	0.55	0.30
PSSMSeq_RBFK Surface	3.5	0.35	0.79	0.47	0.32
	5.0	0.48	0.74	0.57	0.35
KYG	3.5	0.39	0.68	0.49	0.33
	5.0	0.54	0.63	0.56	0.36
DRNA	3.5	0.94	0.23	0.21	0.19
	5.0	0.98	0.19	0.21	0.19
OPRA	3.5	0.51	0.46	0.36	0.21
	5.0	0.64	0.45	0.43	0.25

2.3 Conclusions

Studying the interfacial residues of protein-RNA complexes allows biologists to investigate the underlying mechanisms of protein-RNA recognition. Because experimental methods for identifying RNA-binding residues in proteins are, at present, time and labor intensive, reliable computational methods for predicting protein-RNA interface residues are valuable.

In this study, we evaluated different machine learning classifiers and different feature encodings for predicting RNA-binding sites in proteins. We implemented Naïve Bayes and Support Vector Machine classifiers using several sequence and simple structure-based features and evaluated performance using sequence-based k-fold cross-validation. Our results from this set of experiments indicate that using PSSM profiles outperforms all other sequence-based methods. This is in agreement with previously published studies [Kumar et al. (2008); Spriggs et al. (2009); Wang et al. (2008); Tong et al. (2008); Wang and Brown (2006b)], which demonstrated increased accuracy of prediction of RNA-binding residues by using PSSM profiles. Taken together, these results indicate that determinants of protein-RNA recognition include features that can be effectively captured by amino acid sequence (and sequence conservation) information alone. However, exploiting additional features of structures (e.g. geometry, surface roughness, CX protrusion index, secondary structure, side chain environment) can result in improved performance as suggested in the studies of Liu et al. (2010), Ma et al. (2011), Towfic et al. (2010), and Wang et al. (2008). We observed that the performance of methods utilizing the PSSMSeq feature is comparable to that of three state-of-the-art structure-based methods [Kim et al. (2006), Zhao et al. (2010), Perez-Cano and Fernandez-Recio (2010b)] in terms of MCC. Nonetheless, structure-based methods achieve higher values of Specificity than methods that rely exclusively on sequence information.

In conclusion, we suggest that for rigorous benchmark comparisons of methods for predicting RNA-binding residues, it is important to consider: (i) the rules used to define interface residues, (ii) the redundancy of datasets used for training, and (iii) the details of evaluation procedures, i.e., cross-validation, performance metrics used, and residue-based versus protein-based evaluation.

Our benchmark datasets and implementation of the best performing sequence-based method for predicting protein-RNA interface residues are freely accessible at

http://einstein.cs.iastate.edu/RNABindR/.

2.4 Methods

2.4.1 Datasets

We used homology-reduced benchmark datasets for evaluating our classifiers. All three datasets (RB106, RB144 and RB198) used in this study contain protein chains extracted from structures of protein-RNA complexes in the PDB, after exclusion of structures whose resolution is worse than 3.5Å and protein chains that share greater than 30% sequence identity with one or more other protein chains. RB106 and RB144 were derived from RB109 and RB147 [Terribilini et al. (2006b, 2007)], respectively, by eliminating three chains in each dataset that are shorter than 40 residues [Maetschke and Yuan (2009)]. RB199 [Lewis et al. (2010)] is a more recently extracted dataset (May 2010) that contains 199 unique protein chains. To be included in the dataset, proteins must include $\geq$ 40 amino acids and $\geq$ 3 RNA-binding amino acids and the RNA in the complex must be $\geq$ 5 nucleotides long. Upon further examination of RB199, it was discovered that one chain, 2RFK_C, had been included erroneously, and so we consider instead the dataset RB198 which does not include that chain. An amino acid residue is considered an interface residue (RNA-binding residue) if it contains at least one atom within 5Å of any atom in the bound RNA.

For all three datasets, we constructed two different versions of the data, referred to as sequence data and structure data. The rationale for creating two different versions of the same dataset was to ensure fair comparison of the sequence and simple structure-based methods. To achieve this, the sequence and structure methods must be evaluated on exactly the same datasets. The sequence data (RB106Seq, RB144Seq, and RB198Seq) consists of all residues in the protein chain, regardless of whether those residues appear in the solved protein structure. On the other hand, the structure data (RB106Str, RB144Str, and RB198Str) includes only those residues that appear in the solved structure of the protein in the PDB. Because of this difference, the total number of residues in the sequence data is greater than the total number of residues in the structure data. Interface residues are labeled with '1' and non-interface residues are labeled '0'. Those residues that appear in the sequence only (i.e., have not been solved in the structure) are labeled as non-interface residues. Table 1 shows the number of interface and non-interface residues for the datasets used in this study.

RB44 [Puton et al. (2012)] is a dataset of 44 RNA-binding proteins released between January 1st and April 28th 2011 from the PDB. No two protein chains in the dataset share greater than 40% sequence identity.

2.4.2 Data Representation

In this study, we use three different encodings for amino acids. First, amino acid identity (ID) is simply the one letter abbreviation for each of the twenty amino acids. The second encoding is a position-specific scoring matrix (PSSM) vector for each amino acid. For each protein sequence in the dataset, the PSSM is generated by running PSI-BLAST [Altschul et al. (1997)] against the NCBI nr database for three iterations with an E-value cutoff of 0.001 for inclusion in the next iteration. The third encoding is the smoothed PSSM [Cheng et al. (2008)].

We employ two methods for capturing the context of an amino acid within the protein. First, sequence-based windows are constructed by using a sliding window approach in which the input to the classifier is the target amino acid and the surrounding n residues in the protein sequence. This captures the local context of the amino acid within the protein sequence. Second, structure-based windows are designed to capture the structural context of each amino acid, based on spatial neighboring residues in the protein three dimensional structure. We define the distance between two amino acids to be the distance between the centroids of the residues. The structure-based window consists of the target residue and the nearest n residues based on this distance measure.

We use both sequence and simple structural features as input to the different classifiers that we have used. Features derived from protein sequence include the amino acid sequence itself (IDSeq), PSSMSeq, the position-specific scoring matrices (PSSMs) and SmoPSSMSeq, the smoothed PSSMs. IDSeq uses a window of 25 contiguous amino acid residues, with 12 residues on either side of the target residue, that is labeled ’0’ or ’1’ depending on whether it is a non-interface or interface residue. PSSMSeq encodes evolutionary information about amino acids. The PSSM is an $n \times 20$ matrix that represents the likelihood of different amino acids occurring at a specific position in the protein sequence, where n is the length of the protein sequence. The PSSMs are generated by PSI-BLAST using three iterations and an E-value of 0.001. PSSMSeq also uses a window size of 25 to encode information about the target residue. All individual values in the PSSM are normalized using the logistic function, $y = \frac{1}{1+e^{-x}}$, where y is the normalized value and x is the original value. Each target residue is represented by 500 (25 × 20) features. The smoothed PSSM concept (SmoPSSMSeq) was first introduced by [Cheng et al. (2008)] and was shown to perform significantly well in predicting interface residues for the protein-RNA problem. In the construction of a smoothed PSSM, the score for a target residue i is obtained by summing up the scores of neighboring residues. The number of scores to be summed up is determined by the size of the smoothing window. For example, if the smoothing window size is 5, then we sum up scores for residues at positions $i-2$ to $i+2$ to get the score for residue i. We experimented with a smoothing window size of 3, 5 and 7 and obtained the best performance with a smoothing window size of 3 (data not shown).

IDStr, PSSMStr, and SmoPSSMStr are structural features equivalent to the above sequence features. The major difference between structural and sequence features is that contiguous residues for structural features are listed as those residues that are close to each other (in space) within the structure of the protein, regardless of whether they are contiguous in the protein sequence.

2.4.3 Classifiers

The Naïve Bayes (NB) classifier is based on Bayesian statistics and makes the simplifying assumption that all attributes are independent given the class label. Even though the independence assumption is often violated, NB classifiers have been shown to perform as well as or better than more sophisticated methods for many problems. In this work, we used the NB implementation provided by the Weka machine learning workbench [Witten and Frank (2005)].

Let X denote the random variable corresponding to the input to the classifier and C denote the binary random variable corresponding to the output of the classifier. The NB classifier assigns input x the class label '1' (interface) if:

$$\frac{P(C=1|X=x)}{P(C=0|X=x)} \geq 1$$

and the class label '0' (non-interface) otherwise. Because the inputs are assumed to be independent given the class, using Bayes' theorem we have:

$$\frac{P(C=1|X=x)}{P(C=0|X=x)} = \frac{P(C=1)\prod_{i=1}^{n} P(X_i=x|C=1)}{P(C=0)\prod_{i=1}^{n} P(X_i=x|C=0)}$$

The relevant probabilities are estimated from the training set using the Laplace estimator [Mitchell (1997)].

The Support Vector Machine (SVM) classifier finds a hyperplane that maximizes the margin of separation between classes in the feature space. When the classes are not linearly separable in the feature space induced by the instance representation, SVM uses a kernel function K to map the instances into a typically high dimensional kernel-induced feature space. It then computes a linear hyperplane that maximizes the separation between classes in the kernel-induced feature space. In practice, when the classes are not perfectly separable in the feature space, it is necessary to allow some of the training samples to be misclassified by the resulting hyperplane. More precisely, the SVM learning algorithm [Keerthi et al. (2001); Platt (1999)] finds the parameters w, b, and slack variables ξ_i by solving the following optimization problem:

$$Min_{w,b,\xi_i}\left(\frac{1}{2}w^T w\right) + C\sum_{i=1}^{n}\xi_i \text{ subject to } y_i(w^T\Phi(x_i)+b) \geq 1-\xi_i, \xi_i \geq 0,\ i=1,\,2,\ldots,n$$

where $w \in \mathbb{R}^d$ is a weight vector, b is a bias and Φ is a mapping function. The larger the value of C, the higher the penalty assigned to errors. We use both the polynomial kernel with $p = 1$ (Equation 2.1) and radial basis function (RBF) kernel with $\gamma = 0.01$ (Equation 2.2) in our study. For our experiments, we used the SVM algorithm implementation (SMO) available in Weka [Witten and Frank (2005)]. We used default parameters for the kernels ($p = 1$, $\gamma = 0.01$, and $C = 1.0$) without any optimization in our experiments.

$$K(x_i, x_j) = (x_i \,.\, x_j + 1)^p \text{ where the degree of the polynomial } p \text{is a user-specified parameter} \tag{2.1}$$

$$K(x_i, x_j) = exp(-\gamma\,\|x_i - x_j\|^2) \text{ where } \gamma \text{ is a training parameter} \tag{2.2}$$

We trained all three classifiers on the different sequence- and structure-based features that we constructed. We balanced the training datasets for the SVM classifiers by employing undersampling of the majority class (i.e., non-interface residues). We also changed nominal attributes (IDSeq and IDStr) to binary attributes using the Weka unsupervised filter *NominalToBinary* for input to the SVM classifier.

2.4.4 Performance Measures

All the statistics reported in this work are for the positive class (i.e., interface residues). To assess the performance of our classifiers we report the following measures described in Baldi et al. (2000): Receiver Operating Characteristic (ROC) curve, Precision-Recall (PR) curve, Area Under the ROC Curve (AUC), Specificity, Sensitivity, Fmeasure and Matthews Correlation Coefficient (MCC):

$$Specificity = \frac{TP}{TP+FP} \text{ (Precision)}$$

$$Sensitivity = \frac{TP}{TP+FN} \text{ (Recall)}$$

$$Fmeasure = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

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

We denote true positives by TP, true negatives by TN, false positives by FP and false negatives by FN. The measures describe different aspects of classifier performance. Intuitively, Specificity corresponds to the probability that a positive class prediction is correct; Sensitivity corresponds to the probability that the predictor detects the instances of the positive class. Often it is possible to trade off Specificity against Sensitivity. In the extreme case, a predictor that makes 0 positive predictions (TP + FP = 0, and hence TP = 0 and FP = 0) trivially achieves a Specificity of 1. However, such a predictor is useless in practice because it fails to identify any instances of the positive class, and hence has a Sensitivity as well as MCC of 0. An ideal predictor has both Specificity *and* Sensitivity equal to 1 and Fmeasure as well as MCC equal to 1.

The ROC curve plots the proportion of correctly classified positive examples, True Positive Rate (TPR), as a function of the proportion of incorrectly classified negative examples, False Positive Rate (FPR), for different classification thresholds. In comparing two different classifiers using ROC curves, for the same FPR, the classifier with higher TPR gives better performance measures. Each point on the ROC curve represents two particular values of TPR and FPR obtained using a classification threshold θ. The ROCR package [Sing et al. (2005)] in R was used to generate all ROC curves and PR curves. PR curves give a more informative picture of an algorithm's performance when dealing with unbalanced datasets [Davis and Goadrich (2006)]. In our case, we have many more negative examples (non-interface residues) than positive examples (interface residues) in the dataset. In PR curves, we plot precision (specificity) as a function of recall (sensitivity or TPR).

To evaluate how effective a classifier is in discriminating between the positive and negative instances, we report the AUC on the test set, which represents the probability of a correct classification [Baldi et al. (2000)]. That is, an AUC of 0.5 indicates a random discrimination between positives and negatives (a random classifier), while an AUC of 1.0 indicates a perfect discrimination (an optimal classifier).

The above performance measures are computed based on a sequence-based k-fold cross-validation procedure. k-fold cross-validation [Mitchell (1997)] is an evaluation scheme for estimating the generalization accuracy of a predictive algorithm (i.e., the accuracy of the predictive model on the test set). In a single round of sequence-based cross-validation, m protein sequences ($m = D/k$ where D is the number of sequences in the dataset) are randomly chosen to be in the test set and all the other sequences are used to train the classifier. Sequence-based cross-validation has been shown to be more rigorous than window-based cross-validation [Caragea et al. (2007a)], because the procedure guarantees that training and test sets are disjoint at the sequence level. Window-based cross-validation has the potential to bias the classifier because portions of the test sequence are used in the training set. In this work, we report the results of sequence-based five-fold cross-validation.

We report our results using two performance evaluation approaches. The first approach, called protein-based evaluation, provides an assessment of the reliability of predicted interfaces in a given protein. The second approach, which we call residue-based evaluation, provides an assessment of the reliability of prediction on a given residue. Let S represent the dataset of sequences. We randomly partition S into k equal folds $S_1, ..., S_k$. For each run of a cross-validation experiment, $k-1$ folds are used for training the classifier and the remaining fold is used for testing the classifier. Let $S_i = (s_{1_i}, ..., s_{r_i})$ represent the test set on the i-th run of the cross-validation experiment (r_i is the number of sequences in the test set S_i). In protein-based evaluation, we calculate for each sequence $s_{ji} \in S_i$ the true positives (TP_{ji}), true negatives (TN_{ji}), false positives (FP_{ji}), and false negatives (FN_{ji}). These values are then used to compute the true positive rate (TPR_{ji}) and false positive rate (FPR_{ji}) for each protein s_{ji} in the test set S_i. The TPR and FPR values for the i-th cross-validation run are then obtained as: $TPR_i = \frac{\sum_j TPR_{ji}}{r_i}$ and $FPR_i = \frac{\sum_j FPR_{ji}}{r_i}$. We then report the average TPR and FPR of the classifier over k-folds as $TPR_{protein} = \frac{\sum_i TPR_i}{k}$. Other performance measures for protein-based evaluation are obtained in an analogous fashion. The residue-based measures are estimated as follows: $TP_i = \sum_{j=1}^{r_i} TP_{ji}$, $TN_i = \sum_{j=1}^{r_i} TN_{ji}$, $FP_i = \sum_{j=1}^{r_i} FP_{ji}$ and $FN_i = \sum_{j=1}^{r_i} FN_{ji}$. These values are then used to calculate TPR_i $(= \frac{TP_i}{TP_i+FN_i})$ and FPR_i $(= \frac{FP_i}{FP_i+TN_i})$ for the i-th cross-validation run. We then report the average TPR of the classifier over the k-folds as $TPR_{residue} = \frac{\sum_i TPR_i}{k}$. Other residue-based performance measures are obtained in an analogous fashion.

2.4.5 Statistical Analysis

We used the non-parametric statistical test proposed by Demšar (2006) to compare the performance of the different prediction methods across the three benchmark datasets, RB106, RB144, and RB198. First we computed the ranks of the different methods for each dataset separately, with the best performing algorithm getting the rank of 1, the second best rank of 2 and so on. Demĺar proposes the Friedman test [Friedman (1940)] as a non-parametric test that compares the average ranks of the different classifiers. For the results of the Friedman test to be statistically sound, the number of datasets should be greater than 10 and the number of classifiers should be more than 5 [Demšar (2006)]. Because we have only three datasets, the Friedman test is not applicable (and thus, was not performed) and we relied on average rank across the three datasets to compare the performance of the different methods. As noted by Demĺar, average rank of the classifier provides a fair means of comparing alternative classifiers.

Competing interests

The authors declare that they have no competing interests.

2.5 Acknowledgements

This work was funded in part by the National Institutes of Health grant GM066387 to Vasant Honavar and Drena Dobbs and in part by a research assistantship funded by the Center for Computational Intelligence, Learning, and Discovery. The authors sincerely thank Rafael Jordan for helpful comments and Jeffry Sander for useful discussions. We also thank Lukasz Kozlowski, Janusz M. Bujnicki, and other members of their group for sharing their data with us, enabling us to compare our methods with other structure-based methods as well as include more performance metrics in our comparisons. The work of Yasser El-Manzalawy was supported by National Science Foundation funding. The work of Vasant Honavar while working at the National Science Foundation was supported by the National Science Foundation. Any opinions, findings, and conclusions contained in this article are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Author's contributions

VH and DD conceived of the study and contributed to experimental design and writing. RW carried out the implementation, experiments, and analysis with assistance from CC, FT and YE-M. MT and BL prepared the datasets used in the study and performed preliminary experiments. RW prepared the initial manuscript. All authors read and approved the manuscript.

CHAPTER 3. RNABINDRPLUS: A PREDICTOR THAT COMBINES MACHINE LEARNING AND SEQUENCE HOMOLOGY-BASED METHODS TO IMPROVE THE RELIABILITY OF PREDICTED RNA-BINDING RESIDUES IN PROTEINS

Paper originally appeared in PLoS ONE 9(5): e97725, 2014[1]

Rasna R. Walia, Li C. Xue, Katherine Wilkins, Yasser El-Manzalawy, Drena Dobbs, and Vasant Honavar

Abstract

Protein-RNA interactions are central to essential cellular processes such as protein synthesis and regulation of gene expression and play roles in human infectious and genetic diseases. Reliable identification of protein-RNA interfaces is critical for understanding the structural bases and functional implications of such interactions and for developing effective approaches to rational drug design. Sequence-based computational methods offer a viable, cost-effective way to identify putative RNA-binding residues in RNA-binding proteins. Here we report two novel approaches: (i) HomPRIP, a sequence homology-based method for predicting RNA-binding sites in proteins; (ii) RNABindRPlus, a new method that combines predictions from HomPRIP with those from an optimized Support Vector Machine (SVM) classifier trained on a benchmark dataset of 198 RNA-binding proteins. Although highly reliable, HomPRIP cannot make predictions for the unaligned parts of query proteins and its coverage is limited by the availability of close sequence homologs of the query protein with experimentally determined RNA-binding sites. RNABindRPlus overcomes these limitations. We compared the performance of HomPRIP and RNABindRPlus with that of several state-of-the-art predictors on

two test sets, 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 and 0.37, respectively, and outperforming all other methods, including those that make use of structure-derived features of proteins. More importantly, RNABindRPlus outperforms all other methods for any choice of tradeoff between precision and recall. An important advantage of both HomPRIP and RNABindRPlus is that they rely on readily available sequence and sequence-derived features of RNA-binding proteins. A webserver implementation of both methods is freely available at `http://einstein.cs.iastate.edu/RNABindRPlus/`.

3.1 Introduction

Protein-RNA interactions play key roles in many vital cellular processes including translation [Galicia-Vazquez et al. (2009); Standart and Jackson (1994)] , post-transcriptional regulation of gene expression [Grigull et al. (2004); Tadros et al. (2007)], RNA splicing [Blencowe (2006); Muers (2008)], and viral replication [Denison (2008); Nagy and Pogany (2011)]. Recent evidence points to the role of non-coding RNAs (ncRNAs) in a number of human diseases [Esteller (2011); Khalil and Rinn (2011); Tsai et al. (2011); Van Roosbroeck et al. (2013)] such as Alzheimer's [Schonrock and Götz (2012); Tan et al. (2013)] and various cancers [Huarte and Rinn (2010); Mitra et al. (2012); Cheetham et al. (2013); Kechavarzi and Janga (2014)]. Reliable identification of protein-RNA interfaces is critical for understanding the structural bases, the underlying mechanisms, and functional implications of protein-RNA interactions. Such understanding is essential for the success of efforts aimed at identifying novel therapies for genetic and infectious diseases.

Despite extensive structural genomics efforts, the number of solved protein-RNA structures substantially lags behind the number of possible protein-RNA complexes [Puton et al. (2012)]. Because of the cost and effort involved in the experimental determination of protein-RNA complex structures [Ke and Doudna (2004); Wu et al. (2005)] and RNA-binding sites in proteins [Hellman and Fried (2007); Ule et al. (2005)], considerable effort has been directed at developing reliable computational methods for predicting RNA-binding residues in proteins.

Computational approaches to protein-RNA interface prediction fall into two broad categories [Puton et al. (2012); Walia et al. (2012)]: (i) Sequence-based methods, which use an encoding of sequence-derived features of a target residue and its neighboring residues in sequence (sequence neighbors) to make predictions, and (ii) Structure-based methods, which use an encoding of structure-derived features of a target residue and its neighboring residues in sequence or structure to make predictions. Sequence-based methods [Carson et al. (2010); Cheng et al. (2008); Jeong et al. (2004b); Jeong and Miyano (2006b); Kumar et al. (2008); Ma et al. (2011); Spriggs et al. (2009); Wang et al. (2011); Wang and Brown (2006a); Wang et al. (2010a); Wang and Brown (2006b); Terribilini et al. (2006b)] have exploited features such as amino acid sequence identity, physicochemical properties of amino acids, predicted solvent accessibility, position-specific scoring matrices (PSSMs), and interface propensities, among others. Structure-based methods [Kim et al. (2006); Maetschke and Yuan (2009); Perez-Cano and Fernandez-Recio (2010b); Towfic et al. (2010)] have used features such as amino acid doublet propensities of surface residues, geometry (patches or clefts) of the protein surface, roughness, and atomic protrusion (CX) values, to make predictions of RNA-binding residues in proteins.

Two recent comprehensive surveys of machine learning methods for predicting interfacial residues in protein-RNA complexes [Puton et al. (2012); Walia et al. (2012)] came to a somewhat surprising conclusion that the performance of sequence-based methods, especially those that use PSSMs to encode protein sequences, is comparable to that of structure-based methods, i.e., methods that take advantage of three-dimensional structure of the target protein, when available. *MCC* (Matthews Correlation Coefficient) values for the best methods ranged from 0.38 to 0.46. The difference in performance of the best performing methods among those available was relatively small, and in several cases, not statistically significant [Walia et al. (2012)].

Homology-based methods have proven successful in many bioinformatics tasks, including protein structure prediction [Martiŋ-Renom et al. (2000)], protein function annotation [Andrade (1999); Zehetner (2003)], protein interaction prediction [Matthews et al. (2001)], protein-protein docking [Mukherjee and Zhang (2011); Xue et al. (2014) and protein-protein interface prediction, based on either sequence homology [Xue et al. (2011)] or structure homology [Jordan et al. (2012); Konc and Janezic (2010); Zhang et al. (2011, 2010a)]. Homology-based methods have been shown to outperform other methods whenever close sequence or structural homologs of query proteins (used as templates) can be reliably identified [Kauffman and Karypis (2009); Xue et al. (2011); Jordan et al. (2012)]. Based on their analysis of a dataset of 261 protein-RNA complexes, Spriggs and Jones [Spriggs and Jones (2009)] concluded that RNA-binding residues are more conserved than other surface residues in RNA-binding proteins. To the best of our knowledge, however, there have been no studies that have examined the extent to which RNA-binding residues are indeed conserved among homologous proteins, or used sequence homology to reliably predict RNA-binding residues in protein.

Against this background, we explore whether sequence homology can be used to accurately predict RNA-binding residues in proteins and whether the resulting sequence homology-based approach can be combined with a state-of-the-art machine learning method to enhance the reliability of the predicted RNA-binding residues. Specifically, we: (i) introduce a novel sequence homology-based approach for prediction RNA-binding residues in proteins, HomPRIP, which accurately predicts the RNA-binding residues in a query protein based on the known RNA-binding residues of sequence homologs of the query protein (whenever such homologs are available); and (ii) propose RNABindRPlus, a novel two-stage predictor that uses logistic regression to optimally combine the predictions from HomPRIP and an optimized SVM classifier, SVMOpt, trained to predict RNA-binding interface residues using only sequence derived features of the query protein. We demonstrate that RNABindRPlus substantially outperforms existing sequence-based and structure-based methods. Both HomPRIP and RNABindRPlus have been implemented in a webserver that can be used to reliably predict RNA-binding residues in proteins, even when the structure of the query protein is unavailable.

3.2 Results and Discussion

3.2.1 Rationale for Homology-Based Approach

If RNA-binding residues are conserved across homologous proteins, we can use a simple sequence homology-based approach to predict RNA-binding residues in a query protein: Identify close sequence homologs of the query protein; infer the RNA-binding residues of the query protein based on the known RNA-binding residues of homolog(s) that are aligned with the query protein. The greater the extent to which RNA-binding residues are conserved across homologous protein-RNA complexes, the greater is the reliability with which the RNA-binding residues of a query protein can be predicted based on the known RNA-binding residues of its sequence homologs.

3.2.2 Conservation Analysis of RNA-Binding Residues in Protein-RNA Complexes

Following the approach of Xue et al. [Xue et al. (2011)], we define an interface conservation score $IC(Q, H)$ that 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 (see Methods for details). The IC score measures the degree to which RNA-binding residues of Q are conserved in (and hence can be predicted from the known interface residues of) the protein H. We calculated the pairwise IC scores of proteins in a non-redundant dataset of 216 RNA-binding proteins (RBPs) extracted from the PDB (Protein Data Bank, Berman et al. (2000)) as of October 2010 (NR216, see Methods). Our analysis showed that RNA-binding residues of a protein are highly conserved among its close sequence homologs (data not shown).

Whenever a query protein has a sufficiently high IC score with respect to its putative sequence homolog, we can predict its RNA-binding residues based on the known RNA-binding residues of its sequence homolog. However, examination of the precise definition of the IC score of a protein with respect to its putative sequence homolog (see Methods) shows that computing it requires knowledge of the RNA-binding residues of both the query protein and its homolog. How can we then use the IC score, $IC(Q, H)$, of a query protein Q with respect to a putative sequence homolog H to determine whether we can reliably *predict* the *unknown* RNA-binding residues of Q based on the *known* RNA-binding residues of H? Fortunately, as shown below, we can estimate $IC(Q, H)$ using available information, e.g., the sequence alignment of Q with H. Specifically, we construct a regression model to predict the IC score for the query protein from its alignment with its sequence homolog(s) with known RNA-binding residues.

3.2.3 Predicting the Interface Conservation Score of a Query Protein

We used Principal Components Analysis (PCA) to explore the relationship between six key sequence alignment statistics (see Methods), that are indicative of the quality of the alignment of a protein with its putative sequence homologs, and the IC score of the protein. Our analysis showed that a large fraction (90.6%) of the variance of the IC score is accounted for by the first two principal components. Figure 3.1 shows the projection of 6-dimensional alignment statistics of a protein and its sequence homolog(s) onto a 2-dimensional plane defined by the first two principal components. The resulting 2-dimensional interface conservation space can be partitioned into three regions based on the IC score: (i) Dark Zone, which contains query-homolog pairs with low IC scores (blue data points); (ii) Twilight Zone, which contains query-homolog pairs with intermediate IC scores (yellow, orange, and green data points); and (iii) Safe Zone, which contains query-homolog pairs with high IC scores (red data points).

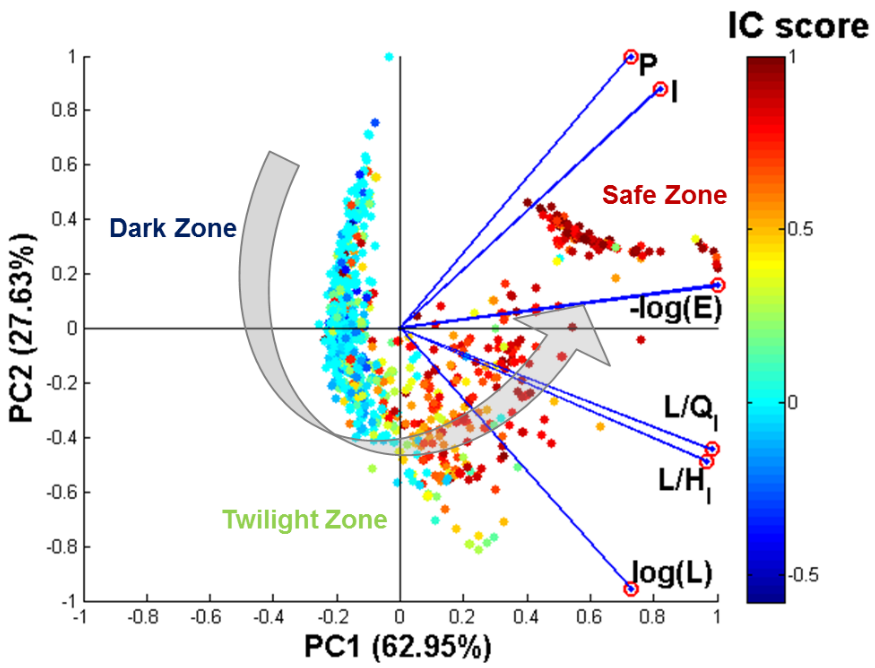


Figure 3.1 Principal Components Analysis (PCA) of interface conservation scores and sequence alignment statistics. Data points in the plot correspond to the projection of a 6-dimensional vector representing the pairwise alignment of a query and homolog sequence onto a 2-dimensional space defined by the first and second principal components. Blue lines with red circles at their tips represent the axes of the original 6-dimensional space for the 6 variables used in PCA analysis: -log(E) (where is the E-value), Identity Score (I), Positive Score (P), log(L) (where L is local alignment length), alignment length fractions (L/Q_l and L/H_l, where Q_l and H_l are the lengths of the query and homolog proteins, respectively). Each data point is colored according to its computed score, with higher score (red/orange) indicating higher interface conservation and lower scores (blue/green) indicating lower interface conservation. The large gray arrow indicates the direction of increasing degree of interface conservation, from Dark to Twilight to Safe Zone.

Based on the results of the PCA analysis which shows that the *Positive Score* and *Identity Score* (I) are highly correlated with each other, we chose to include only the Positive Score (P) along with $log(E)$, $log(L)$ (where E is the E-value, L is the Local Alignment Length), and $F_{QH} = \frac{L}{Q_l} \times \frac{L}{H_l}$ (where Q_l and H_l are lengths of the query protein Q and its homolog H, respectively) in the regression model that predicts the IC score $IC(Q,H)$:

$$\hat{IC}\left(Q,H\right) = \beta_0 + \beta_1 log\left(E\right) + \beta_2 P + \beta_3 F_{QH} + \beta_4 log\left(L\right)$$

All the parameters (Table 3.1) of the regression model are significant (p-values < 0.0001) and the model has an adjusted $R^2 = 0.61$. F_{QH} explains the largest fraction of Type II SS (Sum of Squares) error in the predicted IC score and hence is a good proxy for the IC score.

Table 3.1 The Linear Model for Interface Conservation

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**
β_0	-0.532	0.042	8.70
β_1	0.001	0.000	1.11
β_2	0.005	0.000	12.54
β_3	0.600	0.014	97.55
β_4	0.089	0.007	8.60

3.2.4 HomPRIP: A Sequence Homology-Based RNA-Binding Site Predictor

Now that we have a means of predicting the IC score, $IC(Q,H)$, of a query protein Q with respect to its putative sequence homolog H from the BLAST alignment scores of Q with H, we can proceed to use the predicted IC scores to choose homologs of the query protein to be used to infer the unknown RNA-binding residues of the query protein. HomPRIP, our sequence homology-based protein-RNA interface predictor operates as follows: Given a query protein Q, HomPRIP uses a BLAST search against the proteins in the Protein-RNA Interface Database [Lewis et al. (2010)], PRIDB, to identify a set of sequence homologs of Q, $Homologs(Q)$, with known RNA-binding residues. Each sequence homolog $H_i \in Homologs(Q)$ is assigned a weight w_i, which is the predicted IC score, $\hat{IC}(Q,H)$. A weighted nearest neighbor classifier is used to infer the RNA-binding residues of the query protein based on the known interface residues of its closest homologs (see Methods). The reliability of the predicted RNA-binding residues in each case can be estimated based on the predicted IC scores of the homologs used to arrive at the prediction.

3.2.5 Evaluation of HomPRIP Predictions: Reliability and Coverage

In previous work, we used RB198, a non-redundant dataset of protein-RNA complexes [Lewis et al. (2010)] to assess the performance of alternative approaches to predicting RNA-binding residues in proteins [Walia et al. (2012)]. For the purpose of comparison with previous approaches, we used each of the proteins in the RB198 dataset as a query protein to HomPRIP. HomPRIP searched for putative sequence homologs of the query proteins in RB198 against the nr_RNAprot_s2c database (see Datasets). Homologs that shared greater than 95% sequence similarity with the query proteins were discarded. This ensures that the query protein itself is excluded from being one of the homologs. HomPRIP was able to find at least one Safe, Twilight, or Dark Zone homolog for only 152 out of the 198 proteins in the RB198 dataset. The prediction performance of HomPRIP was evaluated using several standard metrics (see Methods for details). As shown in Table 3.2, for 45% of proteins in RB198, HomPRIP was able to find Safe Zone homologs and, as expected, very reliably predict their RNA-binding residues (with *MCC* of 0.83, *Specificity* of 0.87, and *Sensitivity* of 0.85). For 27% of the proteins, HomPRIP could find only Twilight Zone homologs and for 5%, only Dark Zone homologs. When predictions are based only on Twilight Zone homologs, the performance of HomPRIP drops to an *MCC* of 0.5, *Specificity* of 0.64, and *Sensitivity* of 0.49. When predictions are based only on Dark Zone homologs, HomPRIP has an *MCC* of 0.17, *Specificity* of 0.37, and *Sensitivity* of 0.12. On the 152 proteins that had at least one homolog (from any zone), HomPRIP was able to predict RNA-binding residues with an *MCC* of 0.69, *Specificity* of 0.79, *Sensitivity* of 0.69, and an *F-measure* of 0.73.

The prediction coverage of any sequence homology-based method for predicting RNA-binding residues of proteins is limited by the availability of homologs with known RNA-binding residues. Thus, HomPRIP fails to predict RNA-binding residues of query proteins that do not have at least one homolog with experimentally determined RNA-binding residues. For this reason, HomPRIP fails to return any predictions for 23% of proteins in the RB198 dataset. In addition, HomPRIP cannot make predictions on parts of a query protein sequence that are not aligned with any of its homologs. On the other hand, predictors trained using machine learning offer 100% coverage, although the increased coverage may come at the expense of the reduced reliability of predictions. To explore whether improved predictions can be obtained by combining a sequence homology-based method with a machine learning method, we developed RNABindRPlus, a hybrid predictor that combines HomPRIP predictions with those from an optimized Support Vector Machine (SVM) classifier, SVMOpt (Figure 3.2).

Table 3.2 Performance of HomPRIP on RB198. The performance is shown for the Safe, Twilight, and Dark Zones, separately. Prediction coverage is the fraction of queries that can be predicted by HomPRIP in a given zone.

Homology Zone	**Prediction Coverage**	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
Safe Zone	89/198=45%	0.87	0.85	0.86	0.83
Twilight Zone	54/198=27%	0.64	0.49	0.55	0.50
Dark Zone	9/198=5%	0.37	0.12	0.18	0.17
All Zones	152/198=77%	0.79	0.69	0.73	0.69

3.2.6 Hybrid Method: RNABindRPlus

A recent study [Walia et al. (2012)] compared the performance of Naïve Bayes (NB) and Support Vector Machine (SVM) classifiers trained to predict RNA-binding residues of proteins, from features of a sliding window of 25 amino acid residues centered on the target residue, using three different sequence-based feature representations (amino acid identity, position specific scoring matrices, and smoothed PSSMs [Cheng et al. (2008)]). The study concluded that an SVM classifier, SVM-RBF, which used a radial basis function (RBF) kernel and a PSSM profile to encode the target residue and its sequence neighbors, outperformed all other sequence-based RNA-binding site predictors and was competitive with predictors that use structure-derived features. The study used the default parameters ($C = 1.0$ and $\gamma = 0.01$) for the RBF kernel. In the current study, we used an optimized version of the SVM classifier, which we refer to as SVMOpt. The SVM classifier utilized by RNABindRPlus has the hyper parameters, C and γ, as well as the window size optimized (see Methods) for performance on the RB198 dataset.

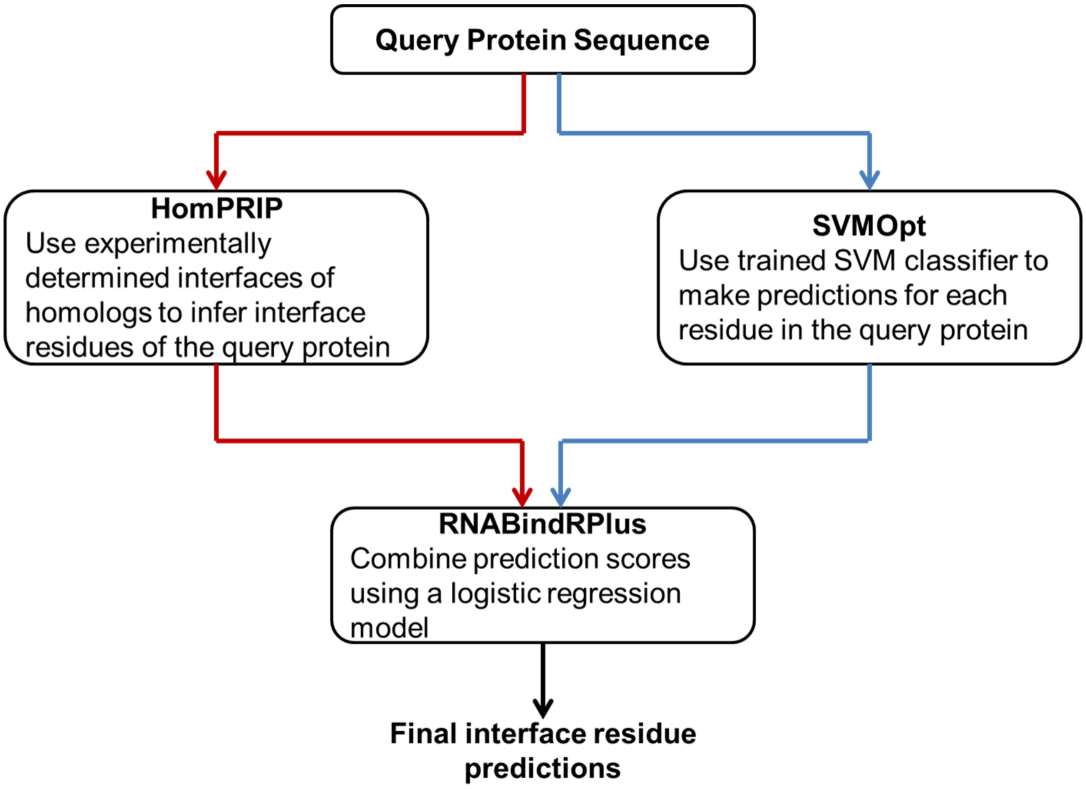

Query Protein Sequence
HomPRIP
Use experimentally determined interfaces of homologs to infer interface residues of the query protein
SVMOpt
Use trained SVM classifier to make predictions for each residue in the query protein
RNABindRPlus
Combine prediction scores using a logistic regression model
Final interface residue predictions

Figure 3.2 RNABindRPlus flowchart. Flowchart showing the different components of RNABindRPlus.

The best combination of parameters was found to be $C = 1.0$, $\gamma = 0.0625$, and a window size of 21 (data not shown). To predict whether or not a given amino acid is an RNA-binding residue, RNABindRPlus combines the prediction scores from HomPRIP with SVMOpt using a logistic regression classifier.

3.2.7 Performance of HomPRIP and RNABindRPlus

To rigorously compare the performance of HomPRIP and RNABindRPlus with each other and with available state-of-the-art methods (see below), we used two independent test sets:

- RB44 [Puton et al. (2012)] (see Datasets), an independent benchmark test set of 44 protein chains extracted from protein-RNA complexes deposited in the PDB between January 2011 and April 2011. The performance of a variety of methods for predicting RNA-binding residues in proteins was benchmarked on this dataset by Puton et al. [Puton et al. (2012)]. Note that the datasets RB198 and RB44 share no common members.

- RB111, a more recently generated test set of 111 protein chains extracted from protein-RNA complexes deposited in the PDB between June 2010 to December 2010, and May 2011 to March 2014. Sequences in RB111 share less than 40% sequence similarity with sequences in RB198 and RB44.

Out of the 44 proteins in the RB44 dataset, HomPRIP was able to make predictions on 28 proteins. Table 3.3 compares the performance measures of different methods on these 28 proteins. HomPRIP achieved an MCC of 0.63 as compared to RNABindRPlus, which had an MCC of 0.60 and the Metapredictor [Puton et al. (2012)] and PiRaNhA [Spriggs and Jones (2009)], both of which had an MCC of 0.51. Other sequence- and structure-based methods tested had even lower values of MCC. This result shows that when HomPRIP can identify homologs with known interfaces, it can outperform other methods.

Out of the 28 proteins, HomPRIP found Safe Zone homologs for 11 proteins, Twilight Zone homologs for 15 proteins, and Dark Zone homologs for 2 proteins. Table 3.4 lists the proteins from RB44 that have homologs in the different homology zones. Not surprisingly, HomPRIP achieved the best results with $Specificity$, $Sensitivity$, F-$measure$, and MCC of 0.88, 0.80, 0.84 and 0.77, respectively on the 11 query proteins for which Safe Zone homologs could be found. On this subset of 11 proteins, HomPRIP substantially outperforms RNABindRPlus, which had $Specificity$, $Sensitivity$, F-$measure$, and MCC values of 0.79, 0.67, 0.72, and 0.61, respectively (Table 3.5). For 15 query proteins that had Twilight Zone homologs, HomPRIP had a higher $Specificity$ of 0.83 than RNABindRPlus (0.73). However, RNABindRPlus had higher values of $Sensitivity$, F-$measure$, and MCC (Table 3.5). On the subset of 2 proteins that have Dark Zone homologs, RNABindRPlus achieved higher values of $Specificity$, $Sensitivity$, F-$measure$, and MCC than HomPRIP (0.83, 0.54, 0.65, and 0.57 versus 0.45, 0.18, 0.26, and 0.13, respectively). Thus, although HomPRIP has higher values of performance metrics on query proteins that have Safe Zone homologs, RNABindRPlus has superior performance on

Table 3.3 Evaluation of Methods on the 28 proteins from the RB44 dataset. The first 11 methods are sequence-based methods. The last 3 methods are structure-based methods (indicated by **). Methods in each category are sorted in descending order of MCC. The highest value in each column is shown in bold font.

Method	Reference	Specificity	Sensitivity	F-measure	MCC
HomPRIP	This paper	**0.84**	0.62	**0.71**	**0.63**
RNABindRPlus	This paper	0.76	0.67	**0.71**	0.60
SVMOpt	This paper	0.58	0.72	0.64	0.48
Metapredictor	Puton et al. (2012)	0.74	0.54	0.62	0.51
PiRaNhA	Murakami et al. (2010)	0.66	0.65	0.65	0.51
BindN+	Wang et al. (2010a)	0.56	0.75	0.64	0.47
PPRInt	Kumar et al. (2008)	0.49	**0.77**	0.60	0.39
PRBR	Ma et al. (2011)	0.58	0.45	0.51	0.34
RNABindR	Terribilini et al. (2007)	0.60	0.39	0.48	0.32
BindN	Wang and Brown (2006a)	0.50	0.50	0.50	0.28
NAPS	Carson et al. (2010)	0.43	0.58	0.49	0.23
KYG**	Kim et al. (2006)	0.55	0.66	0.60	0.41
OPRA**	Perez-Cano and Fernandez-Recio (2010b)	0.61	0.48	0.53	0.37
PRIP**	Maetschke and Yuan (2009)	0.47	0.71	0.56	0.33

Table 3.4 HomPRIP Performance by Zone on RB28. All measures are highest for proteins with Safe Zone homologs and lowest for those with Dark Zone homologs.

Homology Zone	Proteins	Specificity	Sensitivity	F-measure	MCC
Safe Zone	2L5D_A, 2XD0_A, 2XZN_J, 3IZV_M, 3IZW_I, 3J00_G, 3J01_5, 3PIP_F, 3PIP_G, 3PIP_T, 3Q2T_A	0.88	0.80	0.84	0.77
Twilight Zone	2XXA_D, 2XZM_B, 2XZM_C, 2XZM_G, 2XZM_I, 2XZM_M, 3IZV_X, 2RRA_A, 2XZM_E, 2XZM_Q, 2XZN_L, 2XZM_8, 2XZM_S, 2XZM_U, 3IZW_R	0.83	0.55	0.66	0.58
Dark Zone	2XZM_D, 3PDM_P	0.45	0.18	0.26	0.13

query proteins that have homologs in the Twilight and Dark Zones.

On the RB111 dataset, HomPRIP was able to make predictions on 49 proteins (Table 3.6). Table 3.7 compares the performance measures of different methods on these 49 proteins. Not surprisingly, HomPRIP achieves the highest values of all performance metrics on these 49 proteins (*Specificity* of 0.85, *Sensitivity* of 0.85, *F-measure* of 0.85 and *MCC* of 0.83), because it can find Safe Zone homologs for all of them. The second best method on this subset of RB111 is RNABindRPlus, achieving a *Specificity* of 0.64, *Sensitivity* of 0.54, *F-measure* of 0.59, and *MCC* of 0.55.

These results confirm that HomPRIP's prediction performance is dependent upon the degree of sequence similarity between the query protein and its putative sequence homologs with known RNA-binding residues. More importantly, it demonstrates that the homology zones are good indicators of the reliability of HomPRIP's predictions. When Safe Zone homologs are available for query proteins, HomPRIP has the highest predictive performance. In contrast,

Table 3.5 HomPRIP, RNABindRPlus, and SVMOpt Performance by Zone on RB28.

Safe Zone	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
HomPRIP	0.88	0.80	0.84	0.77
RNABindRPlus	0.79	0.67	0.72	0.61
SVMOpt	0.63	0.68	0.65	0.48
Twilight Zone	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
HomPRIP	0.83	0.55	0.66	0.58
RNABindRPlus	0.73	0.69	0.71	0.60
SVMOpt	0.54	0.76	0.63	0.47
Dark Zone	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
HomPRIP	0.45	0.18	0.26	0.13
RNABindRPlus	0.83	0.54	0.65	0.57
SVMOpt	0.68	0.64	0.66	0.52

the performance of RNABindRPlus is similar across proteins from different homology zones, although it is slightly lower than that of HomPRIP on query proteins in the Safe Zone.

3.2.8 What Factors Lead to Superior Performance for RNABindRPlus?

As noted by Walia et al. [Walia et al. (2012)], predictors that use PSSMs outperform those that use amino acid identity when evaluated using a standardized experimental setup (same datasets, same cross-validation procedure). Each score in a PSSM is a log-likelihood ratio of an amino acids' appearance in a specific column of a multiple sequence alignment against a background distribution, representing the degree of conservation of the amino acid in that specific position; the higher the score, the higher the degree of conservation. Therefore, PSSMs capture important evolutionary information by exploiting the large number of available protein sequences, which are much easier to obtain than protein structures.

RNABindRPlus combines our homology-based method, HomPRIP, with SVMOpt, an optimized SVM classifier that uses a radial basis function (RBF) kernel with the sequence PSSM features. We believe that RNABindRPlus achieves a superior performance because it benefits from (i) the interface conservation information contributed by HomPRIP; (ii) residue conservation information encoded in PSSMs; and (iii) the hidden interaction patterns extracted by

Table 3.6 Proteins with Safe Zone Homologs in RB111. There are 49 proteins in RB111 for which HomPRIP can find homologs and return predictions.

Homology Zone	**Proteins**
Safe Zone	2XGJ_A, 2XS2_A, 2YSY_A, 3AGV_A, 3AMT_A, 3B0U_X, 3KFU_A, 3KFU_F, 3LWR_A, 3NMR_A, 3R2C_A, 3RC8_A, 3S14_A, 3S14_B, 3T5N_A, 3V22_V, 3V2C_Y, 3ZD6_A, 4AFY_A, 4ARC_A, 4ATO_A, 4B3G_A, 4BTD_2, 4BTD_D, 4BTD_G, 4BTD_S, 4BTD_X, 4DH9_Y, 4DWA_A, 4E78_A, 4ERD_A, 4IFD_A, 4IFD_H, 4K4Z_A, 4KJ5_5, 4KJ5_G, 3NVI_A, 3OIN_A, 3R9X_B, 3RW6_A, 3ULD_A, 3VYX_A, 4AM3_A, 4B3O_A, 4BA2_A, 4F02_A, 4F1N_A, 4FXD_A, 4GV3_A

SVMOpt from the training set.

3.2.9 Case Study: Accurate Identification of RNA-Binding Residues in the Human Immunorecognition Protein, RIG-I

RNA-protein interactions play key roles in the innate immune system in mammals, which is the first line of defense against invading viral and bacterial pathogens [Iwasaki (2012)]. One class of cytosolic RNA-binding proteins, the RIG-I-Like receptors (RLRs), function as RNA sensors that can identify viral RNA as "non-self" by binding to specific molecular motifs in viral RNAs and activating cellular signaling pathways that stimulate host antiviral immune responses and suppress viral replication [Leung et al. (2012)]. The crystal structure of the RIG-I C-terminal domain (CTD) bound to 5'pp dsRNA has been published [Wang et al. (2010b)], but was not included in the RB44 or RB198 benchmark datasets.

Figure 3.3 shows the predictions of HomPRIP, SVMOpt, and RNABindRPlus on the RIG-I CTD (PDB Id: 3NCU, chain A). All of the homologs used by HomPRIP for making the prediction were in the Safe Zone. This example illustrates how RNABindRPlus combines the predictions from HomPRIP and SVMOpt to provide better overall predictions. RNABindRPlus returns the lowest number of false positive predictions and has the highest MCC (0.75), compared to HomPRIP (0.73) and SVMOpt (0.39). RNABindRPlus also has the highest

Table 3.7 Evaluation of Methods on 49 proteins from the RB111 dataset. The first 7 methods are sequence-based methods. The last 2 methods are structure-based methods (indicated by **). Methods in each category are sorted in descending order of MCC. The highest value in each column is shown in bold font.

Method	Reference	Specificity	Sensitivity	F-measure	MCC
HomPRIP	This paper	**0.85**	**0.85**	**0.85**	**0.83**
RNABindRPlus	This paper	0.64	0.54	0.59	0.55
SVMOpt	This paper	0.27	0.51	0.35	0.28
BindN+	Wang et al. (2010a)	0.28	0.48	0.36	0.28
RNABindR v2.0	Walia et al. (2012)	0.19	0.67	0.30	0.24
PPRInt	Kumar et al. (2008)	0.21	0.56	0.31	0.23
BindN	Wang and Brown (2006a)	0.18	0.39	0.24	0.14
KYG**	Kim et al. (2006)	0.20	0.46	0.28	0.19
PRIP**	Maetschke and Yuan (2009)	0.19	0.49	0.27	0.19

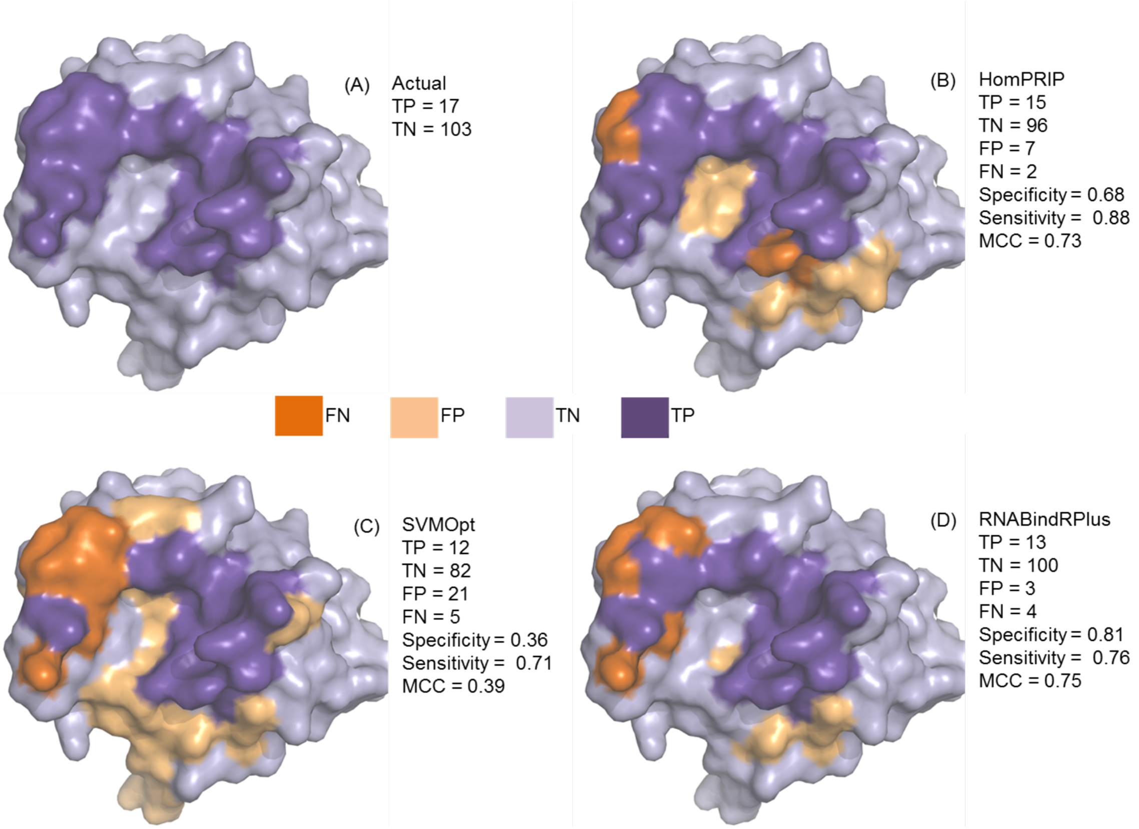


Figure 3.3 PDB ID: 3NCU, Chain A: RIG-I. (A) Actual interface residues, (B) Predictions made by HomPRIP, (C) Predictions made by SVMOpt, and (D) Predictions made by RNABindRPlus.

Specificity of 0.81 compared to HomPRIP (0.68) and SVMOpt (0.36) whereas HomPRIP has the highest *Sensitivity* of 0.88 compared to RNABindRPlus (0.76) and SVMOpt (0.71). For many biological applications, high *Specificity* is desirable, because it allows researchers to identify a short list of residues for targeted mutations designed to alter the affinity or specificity of RNA-binding. As with most classifiers, RNABindRPlus can be tuned to favor even higher specificity, at the expense of lower sensitivity.

3.2.10 RNABindRPlus Outperforms Other Predictors of RNA-binding Residues

On the RB44 dataset, we compared the performance of RNABindRPlus with eight sequence-based methods (see Table 3.8 for method descriptions) and three structure-based methods (see Table 3.9 for method descriptions). These methods were chosen based on a recent study [Puton et al. (2012)] of the performance of readily available sequence- and structure-based predictors of RNA-binding sites in proteins. The Puton et al. study used webservers implementing these methods and concluded that the top performing sequence-based methods were a Metapredictor (which combines predictions from PiRaNhA, BindN+, and PPRInt), PiRaNhA [Murakami et al. (2010)], and BindN+ [Wang et al. (2010a)]. The top performing structure-based methods were KYG [Kim et al. (2006)] and DRNA [Zhao et al. (2010)]. In our comparisons, we used the predictions returned by the same webservers (data shared with us by the Bujnicki group) with one exception. We did not compare our methods with the structure-based version of DRNA because the DRNA webserver uses structural homologs that may be exactly the same as the query protein, which could give the DRNA webserver an unfair advantage over other methods. DRNA can predict i) whether or not a protein is RNA-binding, and ii) which amino acids are RNA-binding. In the Puton et al. study, if a protein was predicted as non-RNA binding by DRNA, the case was considered to be one for which DRNA did not predict any RNA-binding residues [Puton et al. (2012)]. However, in our experiments, we considered only the prediction of the RNA-binding residues, regardless of whether or not a protein was predicted to bind RNA. In addition, we included comparisons with another structure-based method, PRIP [Maetschke and Yuan (2009)].

On the RB111 dataset, we compared the performance of RNABindRPlus with four sequence-based methods (BindN [Wang and Brown (2006a)], BindN+ [Wang et al. (2010a)], PPRInt [Kumar et al. (2008)], and RNABindR v2.0 [Walia et al. (2012)]) and two structure-based methods (KYG [Kim et al. (2006)] and PRIP [Maetschke and Yuan (2009)]). The Metapredictor [Puton et al. (2012)], PiRaNhA [Murakami et al. (2010)], and NAPS [Carson et al. (2010)] servers were all inaccessible at the time of running the experiments on RB111.

Because several methods return only binary predictions, we do not report Area under the ROC Curve (AUC) values, but instead compare the different methods based on *Specificity*, *Sensitivity*, *F-measure* and *MCC*.

The performance of different methods on the RB44 dataset is summarized in Table 3.10. Among all methods that return predictions for every query protein in the dataset (i.e., excluding HomPRIP), RNABindRPlus achieved the highest *MCC* value of 0.55. The next highest

Table 3.8 Sequence-based Methods for Predicting RNA-binding sites in Proteins.

Method	Reference	Description
BindN	Wang and Brown (2006a)	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. Accessible at: http://bioinfo.ggc.org/bindn/
BindN+	Wang et al. (2010a)	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. Accessible at: http://bioinfo.ggc.org/bindn+/
Metapredictor	Puton et al. (2012)	A predictor that combines the output of PiRaNhA, PPRInt, and BindN+ to make predictions of RNA-binding residues using a weighted mean. Accessible at: http://iimcb.genesilico.pl/meta2/. The Metapredictor is not available as of March 2014.
NAPS	Carson et al. (2010)	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. Accessible at: http://prediction.bioengr.uic.edu/. The webserver cannot be accessed as of March 2014.
PiRaNhA	Murakami et al. (2010)	An SVM classifier that makes use of PSSM profiles, interface propensity, predicted solvent accessibility, and hydrophobicity to predict protein-RNA interface residues. Accessible at: http://bioinformatics.sussex.ac.uk/PIRANHA/. The webserver cannot be accessed as of March 2014.
PPRInt	Kumar et al. (2008)	An SVM classifier trained on PSSM profiles. Accessible at: http://www.imtech.res.in/raghava/pprint/
PRBR	Ma et al. (2011)	An enriched random forest classifier trained on predicted secondary structure, a combination of PSSMs with physico-chemical properties, a polarity-charge correlation, and a hydrophobicity correlation. Accessible at: http://www.cbi.seu.edu.cn/PRBR/
RNABindR	Terribilini et al. (2007)	A Naïve Bayes classifier that uses the amino acid sequence identity to predict RNA-binding residues in proteins. Previously accessible at: http://bindr.gdcb.iastate.edu/RNABindR/. It is no longer maintained.
RNABindR v2.0	Walia et al. (2012)	An SVM classifier that uses sequence PSSMs to predict RNA-binding residues in proteins. Accessible at: http://einstein.cs.iastate.edu/RNABindR/.

MCC of 0.48 was obtained by PiRaNhA [Murakami et al. (2010)], and then by SVMOpt and the Metapredictor [Puton et al. (2012)], both with an *MCC* of 0.47. Notably, in terms of *MCC*, the best performing structure-based method was KYG Kim et al. (2006) with a value of 0.42, considerably lower than the top sequence-based methods. The highest *Specificity* was obtained by the Metapredictor (0.74) followed by RNABindRPlus with 0.72. The highest *Sensitivity* was obtained by BindN+ (0.73) [Wang et al. (2010a)] followed by SVMOpt and PPRInt [Kumar et al. (2008)] with 0.72. RNABindRPlus had the highest *F-measure* value of 0.67. A comparison of the ROC curves (Fig. 3.4a) shows that the performance of RNABindRPlus ($AUC = 0.86$) is superior to that of the other two methods (both have an $AUC = 0.82$). Similarly, the PR curves (Fig. 3.4b) show that RNABindRPlus achieves a higher precision at all levels of recall than the other two methods.

The performance of different methods on the RB111 dataset is summarized in Table 3.11. RNABindRPlus achieved the highest *MCC* value of 0.37, followed by SVMOpt and BindN+ [Wang et al. (2010a)], both with an *MCC* of 0.24. The best performing structure-based method on this dataset was KYG [Kim et al. (2006)], with an *MCC* of 0.19, which is considerably lower than the top sequence-based methods. The highest *Specificity* was obtained by RNABindRPlus (0.47) followed by a tie between SVMOpt and BindN+ [Wang et al. (2010a)] (0.25). The highest *Sensitivity* was obtained by RNABindR v2.0 [Walia et al. (2012)] (0.63) followed by PPRInt [Kumar et al. (2008)] (0.48). RNABindRPlus had the highest *F-measure* value of 0.37. A comparison of the ROC curves (Fig. 3.5a) shows that the performance of RNABindRPlus ($AUC = 0.82$) is superior to that of the other methods. Similarly, the PR curves (Fig. 3.5b) show that RNABindRPlus achieves a higher precision at all levels of recall than the other five methods.

Interestingly, the performance of all methods is better on the RB44 dataset than on the RB111 dataset. One possible explanation for this is that RB44 is composed mostly of ribosomal protein chains (36/44), whose roles are structural rather than enzymatic. In contrast, RB111 contains a much smaller proportion of ribosomal protein chains (10/111) and many more enzymes, including CRISPR nucleases, RNA helicases, and RNA methylases. This suggests that training custom classifiers on specific functional or structural classes of RNA-binding proteins

Table 3.9 Structure-based Methods for Predicting RNA-binding sites in Proteins.

Method	Reference	Description
KYG	Kim et al. (2006)	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. Accessible at: http://cib.cf.ocha.ac.jp/KYG/
OPRA	Perez-Cano and Fernandez-Recio (2010b)	Uses patch energy scores calculated using interface propensity scores weighted by the accessible surface area of a residue to predict RNA-binding sites. The program is available upon request from the authors.
PRIP	Maetschke and Yuan (2009)	Uses an SVM classifier and a combination of PSSM profiles, solvent accessible surface area (ASA), betweenness centrality, and retention coefficient as input features. Not accessible via the web server, but results can be obtained via correspondence with the author.

could provide improved performance.

Taken together, these results demonstrate that the hybrid sequence-based method, RNABindRPlus, has substantially higher MCC values than other methods evaluated here. Moreover, RNABindRPlus outperforms all other methods at any level of precision and recall. An unexpected result is that the top sequence-based methods, e.g. RNABindRPlus, BindN+, and SVMOpt, all have much higher MCC values than any of the structure-based methods.

3.2.11 HomPRIP and RNABindRPlus Webservers

A webserver implementation of HomPRIP and RNABindRPlus is freely available at `http://einstein.cs.iastate.edu/RNABindRPlus/`. Users can submit a single or multiple proteins in FASTA format or upload a file containing proteins in FASTA format. Results returned include the RNA-binding residue predictions from HomPRIP, SVMOpt, and RNABindRPlus, as well as the prediction scores from each method. The server also returns a file containing the putative homologs and corresponding predicted IC scores for the query protein(s). Users can utilize the IC scores to determine whether their query protein(s) have Safe, Twilight, or Dark Zone homologs. A text file containing all potential homologs (i.e., the corresponding protein-RNA complexes with solved structures) and their sequence similarity to the query protein is also returned to the user.

3.3 Materials and Methods

3.3.1 Datasets

We utilized five datasets in our experiments.

1. nr_RNAprot_s2c: We built a BLAST database using RNA-binding proteins from PRIDB [Lewis et al. (2010)] (as of May 2013) with a resolution of 3.5Å or better. There are 210,796 residues and 907 proteins in this database. In our experiments, this dataset was used with BLASTP-2.2.27+ [Altschul et al. (1997)] to search for putative sequence homologs.

Table 3.10 Evaluation of Methods on the RB44 dataset. The first 10 methods are sequence-based methods. The last 3 methods (indicated by **) are structure-based methods. Methods in each category are sorted in descending order of MCC. The highest value in each column is shown in bold font.

Method	**Reference**	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
RNABindRPlus	This paper	0.72	0.63	**0.67**	**0.55**
SVMOpt	This paper	0.58	0.72	0.64	0.47
PiRaNhA	Murakami et al. (2010)	0.64	0.63	0.64	0.48
Metapredictor	Puton et al. (2012)	**0.74**	0.49	0.59	0.47
BindN+	Wang et al. (2010a)	0.54	**0.73**	0.62	0.43
PPRInt	Kumar et al. (2008)	0.50	0.72	0.59	0.38
RNABindR	Terribilini et al. (2007)	0.62	0.39	0.48	0.33
PRBR	Ma et al. (2011)	0.58	0.41	0.48	0.31
BindN	Wang and Brown (2006a)	0.50	0.51	0.50	0.28
NAPS	Carson et al. (2010)	0.43	0.58	0.49	0.22
KYG**	Kim et al. (2006)	0.56	0.67	0.61	0.42
OPRA**	Perez-Cano and Fernandez-Recio (2010b)	0.57	0.51	0.54	0.36
PRIP**	Maetschke and Yuan (2009)	0.46	0.68	0.55	0.31

Table 3.11 Evaluation of Methods on the RB111 dataset. The first 6 methods are sequence-based methods. The last 2 methods (indicated by **) are structure-based methods. Methods in each category are sorted in descending order of MCC. The highest value in each column is shown in bold font.

Method	**Reference**	**Specificity**	**Sensitivity**	**F-measure**	**MCC**
RNABindRPlus	This paper	**0.47**	0.37	**0.42**	**0.37**
SVMOpt	This paper	0.25	0.44	0.32	0.24
BindN+	Wang et al. (2010a)	0.25	0.43	0.31	0.24
RNABindR v2.0	Walia et al. (2012)	0.18	**0.63**	0.28	0.22
PPRInt	Kumar et al. (2008)	0.18	0.48	0.26	0.18
BindN	Wang and Brown (2006a)	0.16	0.39	0.23	0.14
KYG**	Kim et al. (2006)	0.19	0.47	0.27	0.19
PRIP**	Maetschke and Yuan (2009)	0.17	0.45	0.24	0.15

2. NR216: We constructed a maximal non-redundant dataset of RNA-binding proteins (RBPs) using the following steps. We retrieved 9,649 protein chains from the set of all protein-RNA complexes in the PDB [Berman et al. (2000, 2002)] as of October 2010. Out of this redundant set of protein chains, we obtained 242 non-redundant protein chains using PISCES [Wang and Dunbrack (2003)] with the following criteria: (i) sequence identity $\leq$ 30%; (ii) resolution of 3.5Å or better; (iii) sequence length $\geq$ 40 amino acids; (iv) non-X-ray entries were excluded; (v) CA-only entries were excluded. Further, we removed chains with interfaces containing fewer than 5 residues. An amino acid residue is considered an interface residue if it contains at least one heavy atom within 5Å of any atom in the bound RNA. This definition of interface residues is used throughout this paper. The final dataset contained 216 non-redundant RBP chains with 8,420 interface residues and 48,129 non-interface residues (those residues that do not appear in the 3D structure of a complex are not counted, since we cannot determine if they are interface or not). We used NR216 for analyzing interface conservation in RNA-binding proteins.

3. RB198: RB199 [Lewis et al. (2010)] is a dataset that contains 199 non-redundant RNA-binding protein chains. It was created by using the PISCES server [Wang and Dunbrack (2003)] to generate a set of proteins with $<$ 30% sequence identity and a resolution of 3.5Å or better from all protein-RNA complexes in the PDB as of May 2010. To be included in the dataset, proteins must include $\geq$ 40 amino acids and $\geq$ 3 RNA-binding amino acids and the RNA in the complex must be $\geq$ 5 nucleotides long. RB198 is identical to RB199 except that one chain (2RFK_C) was omitted because it does not contain any interface residues based on the definition provided above. To maintain consistency with previous studies, both RB198 and RB199 include another chain (3EX7_A) which has no interface residues and one chain with only 2 interface residues (2J01_4). In this dataset, we consider residues that are not solved in the structure as non-interface residues. We used this dataset for cross-validation experiments and for training the final machine learning classifiers.

4. RB44: This is a non-redundant benchmark dataset compiled by Puton et al. [Puton et al.

(2012)] containing RNA-protein complexes deposited in the PDB [Berman et al. (2000, 2002)] between January 1st and April 28th 2011. It is composed of 44 protein chains that share $< 40\%$ sequence identity. We used this dataset as an independent test set. None of the protein chains in RB44 share any global similarity with RB198 at a sequence similarity threshold of 40%.

5. RB111: This is a dataset compiled as of March 2014 that contains 111 non-redundant RNA-binding protein chains. It was created using the PISCES server [Wang and Dunbrack (2003)] to generate a set of proteins with $< 30\%$ sequence identity and a resolution of 3.5Å or better from all protein-RNA complexes deposited in the PDB between June 2010 and December 2010, and between May 2011 to March 2014. The dataset excludes any non-X-ray entries as well as CA-only entries. All protein chains in this dataset include ≥ 40 amino acids and ≥ 3 RNA-binding amino acids. We used this dataset as a newer, independent test set. None of the protein chains in RB111 share any global similarity with RB198 or RB44 at a sequence similarity threshold of 40% (tool used for this: CD-HIT [Li and Godzik (2006); Fu et al. (2012)]).

3.3.2 Sequence Conservation Analysis

We analyzed interface residues in structural homologs of each protein in a non-redundant dataset of 216 RNA-binding proteins, NR216. We extracted homologs for each of the 216 proteins from the nr_RNAprot_s2c database using BLASTP with an $E\text{-}value \leq 10$. The structures and interface residues for proteins in NR216 and their homologs from nr_RNAprot_s2c were experimentally determined. From the resulting set of homologs, sequences that are likely to be copies of the query sequence and hence likely to introduce an undesirable bias in the estimation of sequence conservation were eliminated to obtain a dataset of 8,970 query/homolog pairs. For each query-homolog pair, (Q, H), we calculated the interface conservation score, $IC(Q, H)$, which is a measure of the degree of conservation of interface residues between the query protein, Q and its homolog(s), H. The higher the IC score, the more conserved are the interface residues between homologs and the query protein.

We studied the functional relationship of the IC score with six alignment statistics, four of which are returned by BLAST [Altschul et al. (1997)] and two of which are derived from BLAST statistics: (i) *Positive score* (P), (ii) *Identity score* (I), (iii) *E-value* (E), (iv) *Local Alignment Length* (L), (v) $\frac{L}{Q_l}$ and (vi) $\frac{L}{H_l}$ (where Q_l and H_l are lengths of the query protein Q and its homolog H, respectively). The last two measures tell us the extent of sequence homology between a query sequence, Q and its homolog, H. The E-value is the expected number of random hits when a query sequence is searched against a database of a particular size. The smaller the *E-value*, the greater the chance that a hit is a biologically relevant homolog. *Identity score* measures the sequence identity shared by two amino acid sequences. BLASTP also returns a *Positive score* for a specific position, which calculates the observed substitutions that preserve the physicochemical properties of the original residue. A substitution of one residue type for another is labeled positive when the corresponding entry in the scoring matrix has a positive score. We represented each query-homolog alignment pair as a data point in a six-dimensional space defined by the six alignment statistics.

We used Principal Components Analysis (PCA), a dimensionality reduction technique, to visualize the relationship between the six sequence alignment statistics and the IC score. We also constructed a regression model to quantitatively describe interface conservation as a function of sequence alignment statistics.

3.3.3 HomPRIP: A Sequence Homology-Based RNA-Binding Site Predictor

Given a query protein sequence, Q, HomPRIP searches the nr_RNAprot_s2c database to identify homologous sequences that correspond to the protein components of experimentally determined protein-RNA complexes. The query protein itself is not utilized as one of the homologs. If at least one Safe Zone homolog is found, HomPRIP uses it to predict the interface residues of the query protein, Q. Otherwise, the search is repeated for homologs in the Twilight and Dark zones. HomPRIP reports the homology zones (Safe, Twilight, or Dark, see Table 3.12) accordingly, and uses the zone as an indicator of prediction confidence. Homologs that share $> 95\%$ sequence identity with the query protein are discarded. This ensures conservative performance estimates for the method. If HomPRIP cannot find homologs in any of the three zones, it does not return any predictions for the query protein.

Table 3.12 Boundaries of Safe, Twilight, and Dark Zones used by HomPRIP.

Homology Zones	***IC* score Cutoff**
Safe Zone	0.70
Twilight Zone	0.20
Dark Zone	0.15

HomPRIP assigns a prediction score to each residue of the query protein sequence based on the label of the residue in the corresponding position in its homolog(s) (after pairwise sequence alignment). Specifically, the prediction score (PS) for the j^{th} residue of the query protein is calculated as:

$$PS_j = \frac{\sum_{i=1}^{k} w_i S_{ij}}{\sum_{i=1}^{k} w_i}, \; j = 1, 2, ..., L$$

where L is the length of the query protein, Q and k is the number of close homologs. S_{ij} is the vote of a homolog H_i ($H_i \in Homologs(Q)$) for the j^{th} position of the alignment and is equal to 1 if the corresponding residue in the homolog is an interface residue and 0 otherwise. w_i is $\hat{IC}(Q, H)$, the IC score predicted by the regression model for the i^{th} homolog of Q. The prediction score, PS_j, is converted into a binary prediction (“1” represents an interface residue and “0” represents a non-interface residue) as follows:

$$Prediction_j = \begin{cases} 1 & \text{if } PS_j \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$$

3.3.4 SVMOpt: Support Vector Machine Classifier

From the Walia et al. [Walia et al. (2012)] study, we picked the best performing feature, PSSMs, and the best classifier, SVM-RBF (SVM with the RBF kernel), and optimized the cost parameter C and the RBF kernel parameter, gamma, as well as the window sizes. We tuned these parameters using a three-dimensional grid search over the range $C = 2^{-5}, ..., 2^{15}$ and $\gamma = 2^{-15}, ..., 2^{3}$ and window sizes ranging from 15 to 27. For finding the optimal values for C, γ, and the window size, we divided RB198 into training, validation, and test sets by splitting it into 6 parts. 165 chains were used for training and validation sets, and 33 chains were used as the held-out test set. Specifically, the optimization process was as follows: (i) Pick values for C, gamma, and the window size, (ii) Train the model using the training set, (iii) Evaluate the performance of the model on the validation set, (iv) Repeat steps (i) - (iii) using different training parameters, (v) Select the best model (parameter values) and train it using all the data from the training and validation sets, and (vi) Assess the final model using the held-out test set. Sequence-based 5-fold cross-validation was used in the optimization experiments, so steps (ii) and (iii) were repeated for each fold. We call the optimized classifier SVMOpt. The PSSMs were constructed by running PSI-BLAST [Altschul et al. (1997)] against the NCBI nr database for three iterations with an E-value cutoff of 0.001 for inclusion in the next iteration.

3.3.5 Hybrid Method: RNABindRPlus

The prediction scores from HomPRIP and SVMOpt were combined using a second stage logistic regression model. The Weka implementation of logistic regression [le Cessie and van Houwelingen (1992)] was used with the default ridge parameter of $1.0E-8$. The input to the logistic regression model is a 2D vector representing the prediction scores from HomPRIP and SVMOpt. In cases where HomPRIP failed to return predictions (i.e., no homologs for query proteins are found or the target residue is not aligned with any residues in the homolog(s)), a missing input value (represented as '?') is fed to the logistic regression model. We refer to this hybrid model as RNABindRPlus.

3.3.6 Performance Evaluation

We used several different measures of classifier performance. On the RB198 dataset, performance measures were obtained by carrying out sequence-based 5-fold cross-validation. Sequence-based 5-fold cross-validation randomly divides protein chains in RB198 into 5 sets and alternatively uses 4 sets as the training set and 1 set as the test set. The average performance on the 5 test sets is used as the final evaluation of the classifier. Sequence-based cross-validation has been shown to be more rigorous than window-based cross-validation [Caragea et al. (2007a)], because it ensures disjoint training and test sets at the sequence level instead of at the residue level. The predicted label for each residue is compared to the actual label and the residue is classified as a true positive (TP), false positive (FP), true negative (TN), or false negative (FN). We report the performance measures as defined in Baldi et al. [Baldi et al. (2000)].

Overall performance measures are calculated as follows:

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

$$Sensitivity = \frac{TP}{TP+FN} \quad (= \text{Recall})$$

$$F\text{-}measure = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

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

The measures describe different aspects of classifier performance. $Sensitivity$ is the probability of correctly predicting the interface residues of a given protein. $Specificity$ is the probability that a predicted interface residue in any given protein is in fact an interface residue. $F-measure$ is the harmonic mean of precision and recall, where the best score is 1 and the worst score is 0. The *Matthews correlation coefficient* (MCC) measures how predictions correlate with true interface and non-interfaces. All machine learning methods have an inherent trade-off between specificity and sensitivity that is controlled through the classification threshold. Predictors that make no positive predictions trivially achieve a specificity of 1. However, such methods are not useful, because they do not return any true positive predictions.

A *Receiver Operating Characteristic* (ROC) curve is useful for comparing classifiers across all classification thresholds. Where possible, we show the ROC curve and report *Area under the ROC curve* (AUC). The ROC curve plots the proportion of correctly classified positive examples, *True Positive Rate* (TPR), as a function of the proportion of incorrectly classified negative examples, *False Positive Rate* (FPR), for different classification thresholds. When comparing the performance of two classifiers, for the same FPR, the one with a higher TPR performs better. The ROCR package [Sing et al. (2005)] in R was used to generate all ROC curves and *Precision-Recall* (PR) curves. When data are unbalanced (fewer interface residues than non-interface residues) PR curves give a more informative picture of an algorithm's performance than ROC curves. In PR curves, we plot precision as a function of recall, with respect to different prediction score cutoffs. We also report the AUC value, which is the probability that a classifier gives a higher score to a positive instance than to a negative instance. An AUC of 0.5 indicates a random discrimination between the positive and negative class while an AUC of 1.0 indicates perfect discrimination.

3.4 Conclusions

We have shown that HomPRIP, a sequence homology-based method, can reliably predict RNA-binding residues when close sequence homologs of the query protein, with known RNA-binding residues, can be found. A sequence-based machine learning classifier, SVMOpt, returns reliable predictions for any query protein, regardless of whether structures of protein-RNA complexes containing homologous protein sequences are available. When Safe Zone homologs for a query protein can be found, HomPRIP is the method of choice. For other query proteins, RNABindRPlus, which combines HomPRIP with SVMOpt, has superior performance because it exploits the strengths of both methods. RNABindRPlus outperforms several state-of-the-art methods, both sequence-based and structure-based, for predicting RNA-binding sites in proteins. An important advantage of RNABindRPlus is that it is a purely sequence-based approach. A webserver implementation is freely available at http://einstein.cs.iastate.edu/RNABindRPlus/.

3.5 Acknowledgments

We thank members of the Dobbs and Honavar groups for useful discussions and Carla Mann for assistance in generating the RB111 dataset. The authors sincerely thank Janusz M. Bujnicki and members of his group, Lukasz Kozlowski and Tomas Puton, for sharing their experimental data with us. The authors also thank Stefan Maetschke for running PRIP on our datasets and sending us the results and Yuedong Yang for assistance with running DRNA on the RB44 dataset.

Author Contributions

RRW, VH, and DD conceived of the study and contributed to experimental design and writing. RRW carried out the implementation, experiments, and analysis with assistance from LCX, KW, and YE-M. RRW prepared the initial manuscript. All authors read and approved the manuscript.

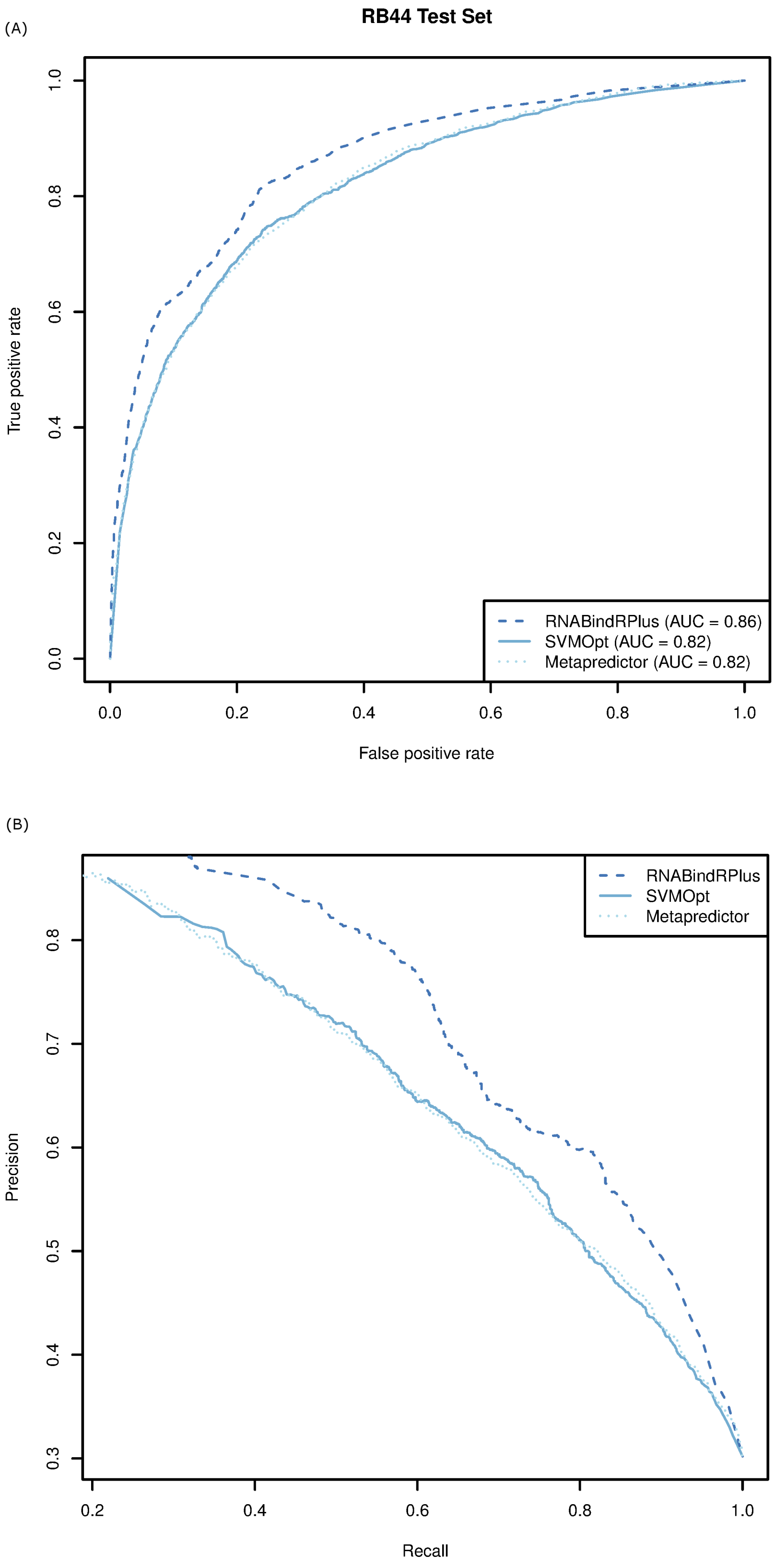


Figure 3.4 Comparison of SVMOpt, RNABindRPlus, and the Metapredictor on the RB44 dataset. (A) ROC curves and (B) PR curves with a 5Å distance cut-off for interface residues.

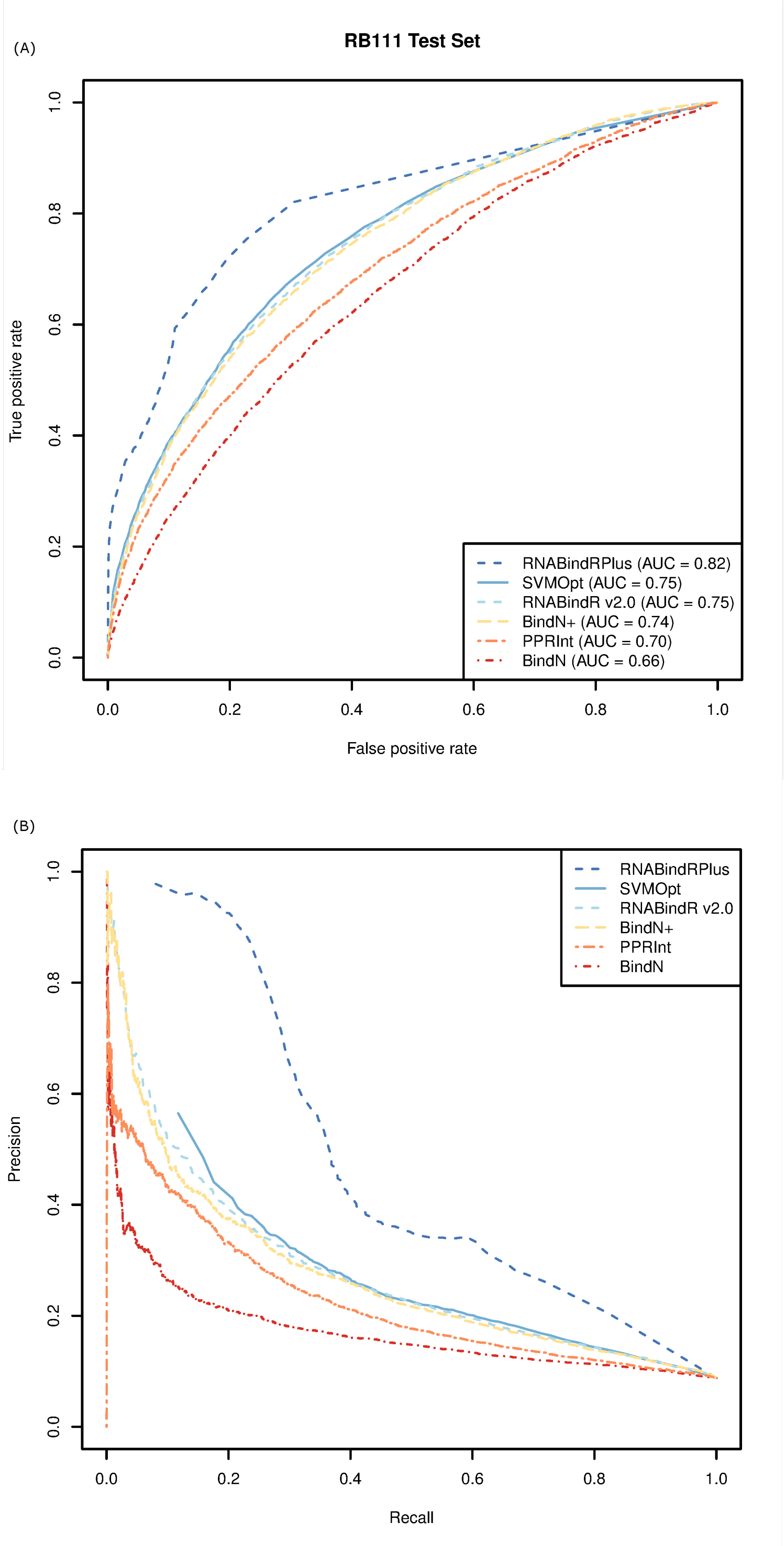


Figure 3.5 Comparison of SVMOpt, RNABindRPlus, RNABindR v2.0, BindN, BindN+ and PPRInt on the RB111 dataset. (A) ROC curves and (B) PR curves with a 5Å distance cut-off for interface residues.

CHAPTER 4. DISCOVERING INTERACTION MOTIFS IN THE INTERFACES OF RNA-PROTEIN COMPLEXES

Manuscript in preparation

Rasna R. Walia, Vasant Honavar, and Drena Dobbs

Abstract

RNA-protein interactions are crucial for many cellular processes. With the increasing numbers of characterized RNA-protein complexes, it is possible to gain deeper insight into the principles of RNA-protein recognition. Previous investigations have identified several types of RNA-binding domains in proteins and characterized the RNA-binding propensities of specific amino acids, but little work has been done to identify sequence motifs that correspond to interfaces in RNA-protein complexes. In this study, we introduce a new way to extract groups of interacting residues from both the RNA and protein sides of RNA-protein complexes in the PDB. We generate a comprehensive list of 33,594 RNA-protein interaction motifs (RPIMs), and demonstrate that RNA-protein interfaces share similarities that can be identified at the primary sequence level.

4.1 Introduction

RNA molecules play many diverse roles in biology, in part due to the diversity of the three-dimensional structures they can adopt [Fritsch and Westhof (2010)]. In order to carry out their biological functions, RNA molecules bind other macromolecules and/or small molecules. In many cases, RNAs interact with one or several proteins to form ribonucleoprotein (RNP) complexes. RNP complexes and RNA-protein interaction networks are involved in myriad essential functions in living systems, ranging from storage and propagation of genetic information, to structural and catalytic roles in ribosomes and spliceosomes, to regulatory roles in non-coding RNA (ncRNA) mediated transcriptional and post-transcriptional gene regulation [Kishore et al. (2010); Geisler and Coller (2013); Licatalosi and Darnell (2010); Khalil and Rinn (2011); Kuersten et al. (2013)].

RNA-binding proteins (RBPs) recognize specific sequence or structural characteristics of their RNA targets. RNA-binding capacity is attributed to RNA-binding domains (RBDs) that bind to RNA recognition elements (RREs) in the RNA [Anko and Neugebauer (2012)]. The RNA recognition motif (RRM) is the most abundant RNA-binding domain in higher vertebrates, and imparts diverse biological functions to RRM-containing proteins. The RRM domain is typically recognized at the primary sequence level as 90 amino acids long, containing two conserved sequences of eight and six amino acids, called RNP1 and RNP2, respectively. Even though the structures of the RRM domain are well characterized, its mode of RNA-protein recognition is not clear because of the high variability of its interactions [Maris et al. (2005); Clery et al. (2008); Daubner et al. (2013)]. The second most abundant domain found in RNA-binding proteins is the double-stranded RNA-binding domain (dsRBD). It is found in proteins that play essential roles in RNA interference, RNA processing, RNA localization, RNA editing, and transcriptional repression [Stefl et al. (2005); Chen and Varani (2013)]. The dsRBDs is a small protein domain containing 70 amino acids with a conserved $\alpha\beta\beta\beta\alpha$ protein topology, which binds double-stranded RNA (dsRNA) [Masliah et al. (2013)]. Two other commonly occurring RNA-binding domains include the K-homology domain (KH) [Valverde et al. (2008)] and zinc finger (ZF) domains [Font and Mackay (2010)]. Not all RBPs have recognizable RNA-binding domains [Hogan et al. (2008); Riley and Steitz (2013)]. Additionally, RNA-binding domains themselves may not directly correspond to residues that bind RNA, but instead are often determinants of a particular protein fold [Li and Li (2005)]. Most RNA-binding domains consist of approximately 100 amino acid residues; typically only 10-20% of those make direct contacts with RNA [Stefl et al. (2005); Chen and Varani (2013)].

The number of high resolution structures of RNA-protein complexes available for defining RNA-protein interaction sites is relatively small at present. The Protein Data Bank (PDB) contains a total of 1,820 RNA-protein complexes, compared to more than 3000 DNA-protein complexes, as of October 15, 2014. With the advent of next-generation sequencing technologies, several high-throughput experimental methods have been developed to elucidate the binding preferences of RBPs for different RNA sequences [Ray et al. (2009); Paz et al. (2014); Kazan et al. (2010)]. Databases such as *RBPmotif* and *CISBP-RNA* [Ray et al. (2013)], CLIPZ [Khorshid et al. (2010)], and doRiNA [Anders et al. (2011)] catalogue binding sites of RBPs derived from RNAcompete [Ray et al. (2009)], PAR-Clip [Hafner et al. (2010a,b)], and related experiments [Keene et al. (2006); McHugh et al. (2014)]. Despite this impressive progress, none of the published methods for defining RNA- binding domains in proteins or consensus recognition motifs in RNA sequences take into account information from both the RNA-binding protein and its RNA target.

A few previous studies have attempted to elucidate interacting residues in protein-protein or DNA-protein interactions. Li and Li [Li et al. (2004); Li and Li (2005)] studied protein-protein binding motif pairs in which each "side" of the binding site consisted of short sequences of continuous residues, and where the two sides are spatially close to one another. They defined a "binding motif pair" as consisting of two such contiguous amino acid sequence motifs, each derived from a different chain in a protein-protein complex. Their idea was to use binding motif pairs to represent correlated patterns of interactions in protein-protein binding sites. Sathyapriya et al. [Sathyapriya et al. (2008)] used bipartite graph representations to identify clusters of interacting residues in DNA-protein complexes. They included residues from both the DNA and protein sides of the interface to obtain insights into DNA-protein interactions by looking at groups of interacting residues instead of pairwise interactions. Their analysis shed light on some of the recognition determinants of DNA-protein interactions, by characterizing their nature and specificity.

Other than unpublished efforts in our own group [Muppirala (2013b)], we are not aware of any previous attempts to identify similar interacting motifs in RNA-protein complexes, i.e., motifs that consist of short contiguous strings of residues extracted from both the RNA and protein sides of RNA-protein complexes. We present a new method for identifying such motifs and have generated a comprehensive collection of novel RNA-protein interaction motifs

(RPIMs) that contain components from both the RNA and protein partners of structurally characterized RNA-protein complexes. Our goal in developing this resource was to answer the following question: Do interfaces in RNA-protein complexes share similarities that can be recognized at the primary sequence level?

4.2 Methods

4.2.1 Dataset

We examined every RNA-protein complex in the Protein Data Bank (PDB) on March 30, 2014. There were a total of 1,720 RNA-protein complexes, out of which 790 are ribosomal complexes. We extracted interacting pairs of RNA and protein chains that contain at least one interacting amino acid-ribonucleotide pair using a 5Å distance cutoff, i.e., those in which at least one heavy atom in an amino acid residue in the protein chain lies within 5Å of a heavy atom in the ribonucleotide in the RNA chain. This resulted in 23,017 interacting pairs of RNA-protein chains. This number includes redundant RNA and protein chains, i.e., we did not exclude RNA or protein sequences that were similar to other sequences the dataset (to extract all distinct motifs in characterized RNA-protein complexes). PDB IDs for all 23,017 interacting pairs are provided online at `https://github.com/Dobbs-Lab/RPIMotif`.

4.2.2 Extracting RPIMs from Contact Matrices using the BWLabel Algorithm

In order to extract RNA-protein interaction motifs (RPIMs) that contain components from both the RNA and protein partners of structurally characterized RNA-protein complexes, we first generated contact matrices as described in Results. The *bwlabel* algorithm [Haralick and Shapiro (1991)] was used to identify clusters of interacting ribonucleotides and amino acids, i.e., relatively small RPIMs that consist of contiguous amino acids from the protein and contiguous ribonucleotides from the RNA, all of which participate directly in the RNA-protein interface, based on a 5Å cutoff distance for interacting residues. Each RPIM was defined as corresponding to a single four-connected component as follows: For each position i, j in the RNA-protein contact matrix, the *bwlabel* algorithm checks the labels of the top, right, bottom, and left neighbors, to determine which cluster (RPIM) entry i, j should be assigned to.

4.2.3 Representing Subgraphs

The result of running *bwlabel* on the RNA-protein contact matrices is a list of corresponding clusters of connected components. The most basic cluster consists of one protein node and one RNA node. We define RNA-protein interaction motifs (RPIMs) as the clusters generated by running the *bwlabel* algorithm.

To facilitate comparison of resulting RPIMs in order to identify similarities and difference among them, we represent RPIMs using a modified version of the graph description language, DOT. For each amino acid residue in a cluster, the identity of the amino acid, as well as the number of amino acids, are represented by nodei(aa_id,type aa), where aa_id is the one letter symbol of the amino acid, $i = 0..m - 1$, and m is the number of amino acids in the cluster. Similarly, for each nucleotide in a cluster, the identity of the nucleotide and the number is represented by nj(nt_id,type nt), where nt_id is the one letter symbol of the nucleotide,$j = 1..n$, and n is the number of nucleotides in the cluster. The connectivity information of the subgraph appears near the end of the description of amino acids and nucleotides in the cluster. For example, node0(W,type aa),n0(G,type nt),n1(U,type nt),n2(G,type nt),W_node0–G_n0,W_node0–U_n1,W_node0–G_n2 represents a subgraph with one protein node, W, and three RNA nodes, rG, rU and rG. The subgraph represented by the description is given in Figure 4.1.

4.3 Results and Discussion

4.3.1 Identifying Small RNA-Protein Interaction Motifs (RPIMs)

To systematically identify RNA-protein interaction motifs (RPIMs) that contain components from both the RNA and protein partners in characterized complexes, we generated a dataset of 23,017 interacting RNA-protein pairs from all RNA-protein complex structures in the PDB (see section 4.2). Because our goal was to identify relatively small RPIMs, we initially focused on motifs that consist of short *contiguous* strings of interacting residues.

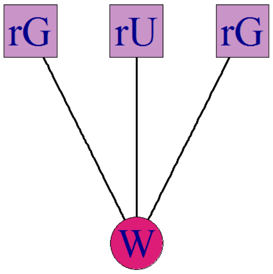
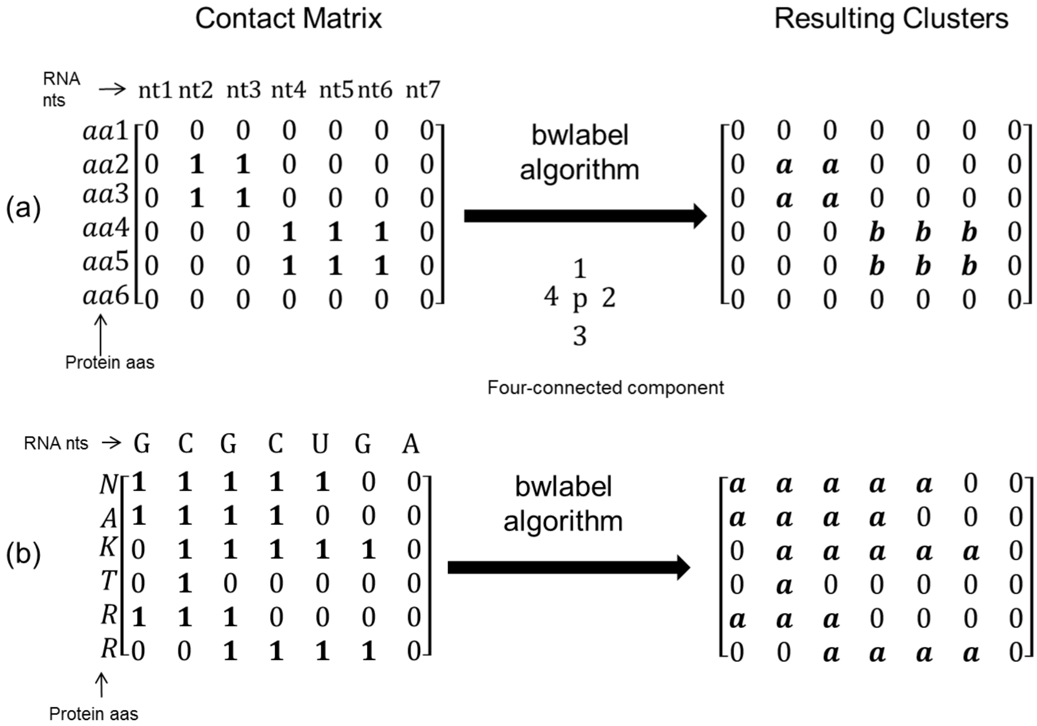


Figure 4.1 Example RPIM. Protein nodes are represented using circles and RNA nodes using squares. The bipartite graph represents the motif described by: node0(W,type aa),n0(G,type nt),n1(U,type nt),n2(G,type nt),W_node0–G_n0,W_node0–U_n1,W_node0–G_n2. This RPIM is found in ribosomal proteins and rRNAs such as 1JJ2_O0, 1K73_QA, 1N8R_QA, 1S1I_P3, and 4G5U_8A (PDBID_ proteinchain, rnachain).

We represent each interacting RNA-protein pair using an adjacency or contact matrix, A, of size $m \times n$, where m is the length of the protein and n is the length of the RNA (Figure 4.2(a)). The rows of the contact matrix represent amino acids, and the columns represent ribonucleotides. Position i, j in the contact matrix is labeled '1' if amino acid i and ribonucleotide j have a heavy atom within 5Å of each other; otherwise it is labeled '0'. We used *bwlabel* [Haralick and Shapiro (1991)], an image analysis algorithm, to identify four-connected components (clusters of contiguous interfacial residues; see section 4.2 for details) in the contact matrices. Figure 4.2(a) shows a simple example matrix and the two clusters, a and b, identified after running bwlabel. Fig 4.2(b) shows part of the contact matrix for the structure of the bacteriophage P22 transcriptional antitermination complex (PDB 1A4T, RNA chain A, protein chain B). The resulting single cluster, a (on the right), corresponds to a single RPIM that consists of the RNA sequence "GCGCUG" and the amino acid sequence, "NAKTRR", with which it interacts.



$$\begin{array}{c|ccccccc} & nt1 & nt2 & nt3 & nt4 & nt5 & nt6 & nt7 \\ \hline aa1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ aa2 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ aa3 & 0 & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ aa4 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ aa5 & 0 & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ aa6 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{array} \xrightarrow{\text{bwlabel algorithm}} \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{a} & \boldsymbol{a} & 0 & 0 & 0 & 0 \\ 0 & \boldsymbol{a} & \boldsymbol{a} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \boldsymbol{b} & \boldsymbol{b} & \boldsymbol{b} & 0 \\ 0 & 0 & 0 & \boldsymbol{b} & \boldsymbol{b} & \boldsymbol{b} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\begin{array}{ccc} & 1 & \\ 4 & \mathrm{p} & 2 \\ & 3 & \end{array}$$

$$\begin{array}{c|ccccccc} & G & C & G & C & U & G & A \\ \hline N & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 \\ A & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 \\ K & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \\ T & 0 & \mathbf{1} & 0 & 0 & 0 & 0 & 0 \\ R & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 & 0 & 0 & 0 \\ R & 0 & 0 & \mathbf{1} & \mathbf{1} & \mathbf{1} & \mathbf{1} & 0 \end{array} \xrightarrow{\text{bwlabel algorithm}} \begin{bmatrix} \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & 0 & 0 \\ \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & 0 & 0 & 0 \\ 0 & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & 0 \\ 0 & \boldsymbol{a} & 0 & 0 & 0 & 0 & 0 \\ \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & 0 & 0 & 0 & 0 \\ 0 & 0 & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & \boldsymbol{a} & 0 \end{bmatrix}$$

Figure 4.2 Identifying RPIMs in RNA-protein contact matrices using the bwlabel algorithm. (a) A simple example illustrating an RNA–protein contact matrix (left) and the two resulting clusters of interfacial residues (right) after running bwlabel to identify four–connected components. Rows represent amino acids (Protein aas) and columns represent ribonucleotides (RNA nts). A "1" appears in position i,j of the contact matrix if the corresponding protein and RNA residues interact, otherwise there is a "0" (see text for details). After running the bwlabel algorithm, the interacting residues are grouped into two clusters or RPIMs, a and b. (b) A partial contact matrix taken from the bacteriophage P22 transcriptional antitermination complex (PDB 1A4T, RNA chain A, protein chain B) (left). In this example, using bwlabel to identify four–connected components results in a single cluster of interacting residues (RPIM) denoted by a. In this example, the RNA and protein residues that make up the RPIM are GCGCUG and NAKTRR, respectively, Note that the final "A" residue in the RNA sequence is not included in this RPIM.

4.3.2 A Large Number of RPIMs Occur Frequently in RNA-Protein Complexes

Using the procedure described above, we identified 33,594 unique RPIMs (RNA-protein interaction motifs) in the dataset of 1,720 RNA-protein complexes extracted from the PDB on March 30, 2014. For completeness, this number includes RPIMs that consist of only a single amino acid residue and a single ribonucleotide (we refer to these as “two-node” RPIMs). Table 4.1 and Figure 4.3 summarize the number of distinct RPIMs that occur at various frequencies within the dataset after excluding this high-frequency class. Notably, 20% of RPIMs appear ≥ 7 times in the dataset of RNA-protein complexes, but about 61% of the 33,594 RPIMs are observed only once or twice.

Table 4.1 RPIM Frequencies

Number of distinct RPIMs	**Observed frequency in dataset of 1,720 complexes**
15,684	Occur only 1 time
17,910	Occur $\geq$ 2 times
8,585	$\geq$ 5 times
5,413	$\geq$ 10 times
4,252	$\geq$ 15 times
3,598	$\geq$ 20 times
3,106	$\geq$ 25 times
2,751	$\geq$ 30 times
1,893	$\geq$ 50 times
767	$\geq$ 100 times
33,594	Total number of distinct RPIMs

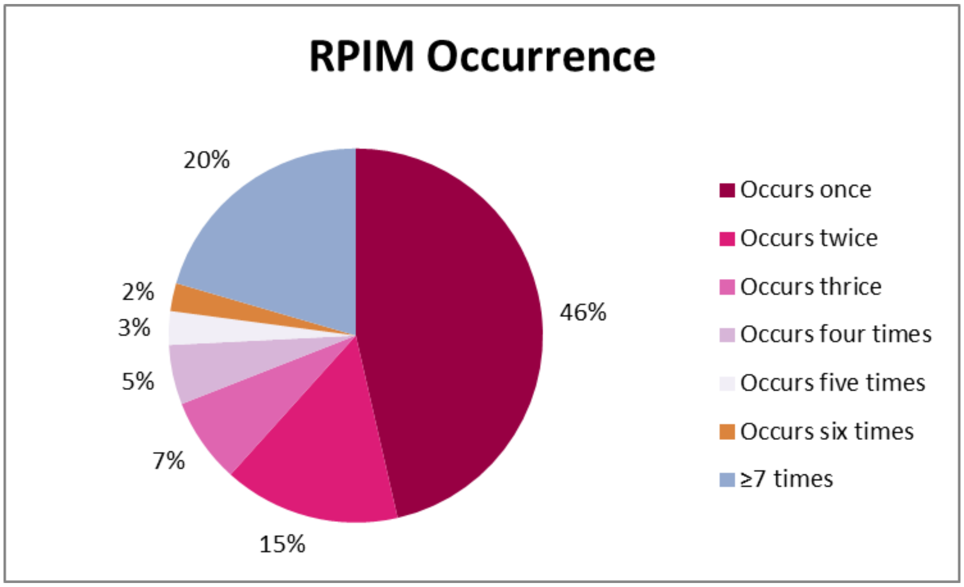


Figure 4.3 Pie chart depicting the percentage of RPIMs that occur a certain number of times.

4.3.3 Composition of the Most Abundant RPIMs

Several studies [Jones et al. (2001); Bahadur et al. (2008); Treger and Westhof (2001); Gupta and Gribskov (2011)] have demonstrated that, as expected, the positively charged residues Arg and Lys have the highest propensity to be found in interfacial regions of RNA-binding proteins. Our results recapitulate this. For example, Table 4.2 shows that among two-node RPIMs, the Arg-Guanine (R:rG) pair occurs most frequently, a total of 15,216 times, followed by the Arg-Adenine (R:rA) pair, which occurs 13,609 times in our dataset. The third most abundant amino acid-ribonucleotide pair is Lys-Guanine (K:rG).

Examples of three-node RPIMs that occur frequently in our dataset are shown in Table 4.3. Note that all of these RPIMs also have Arg and Lys as the RNA-binding residues at the interface.

Examples of some of the most frequently observed longer RPIMs and their counts are shown in Table 4.4, along with the PDB ids in which they occur.

Table 4.2 Counts of two-node RPIMs. The left side of the motif represents the amino acid (e.g. R = Arg) and the right side represents the ribonucleotide (e.g. rG = Guanine).

RPIMs	**Counts**
R:rG	15,216
R:rA	13,609
K:rG	11,927
R:rC	11,106
K:rC	9,344
R:rU	8,756
K:rA	8,619
K:rU	6,909

4.3.4 Amino Acid and Ribonucleotide Composition of RPIMs

Figures 4.4 and 4.5 show the overall amino acid and ribonucleotide content, respectively, of the set of 33,594 RPIMs we extracted. The numbers are derived by counting all the residues found in the RPIMs, and then calculating the percentage of each different amino acid and nucleotide. As expected, positively charged residues Arg and Lys are the most common amino acids in the interaction motifs. Also, the polar residues Asn, Ser, Gln, and Thr can form multiple hydrogen bonds with RNA and are commonly found at RNA-binding interfaces. Gly frequently occurs in binding sites, due to its small size and conformational flexibility [Gupta (2011)]. Although Asp is generally disfavored in RNA-binding sites, a potential role for Asp in RNA-loop recognition has been proposed [Iwakiri et al. (2012)]. Aromatic residues are preferred in the RNA-protein interface due to stacking interactions with bases [Baker and Grant (2007); Gupta (2011)]. The ribonucleotide guanine is universally reported to make the most hydrogen bonds with proteins, and is seen more frequently than the other three common ribonucleotides in interfaces [Gupta and Gribskov (2011)]. Generally, ribonucleotides do not show differing propensities for being in protein-binding versus non-binding regions [Perez-Cano and Fernandez-Recio (2010b); Gupta and Gribskov (2011)] but when nucleobase-specific and –nonspecific interactions are considered separately, distinct preferences emerge.

Table 4.3 Most abundant three-node RPIMs. The left side of the motif represents the amino acid (e.g. R = Arg) and the right side represents the ribonucleotide (e.g. rC).

RPIMs	**Count**
R:rCC	5,453
R:rCG	5,306
R:rAG	5,304
R:rGC	5,244
R:rGG	5,145
K:rCC	4,381
R:rGA	4,278
K:rGG	3,978
K:rGC	3,903
K:rCG	3,737
R:rAA	3,628
R:rUG	3,443
R:rCU	3,372
R:rGU	3,310
K:rGA	3,236
R:rAC	3,233
R:rCA	2,757
K:rGU	2,677
R:rUC	2,617
K:rAG	2,563
K:rUG	2,398
R:rAU	2,232

4.3.5 Case Studies

The question we sought to address in this work is: Do interfaces in RNA-protein complexes share similarities that can be recognized at the primary sequence level? Although we have not yet systematically analyzed all of the motifs identified in this work, analyses of a few examples suggest that the answer is yes. Below we discuss two specific examples: (i) an RPIM found in Ebola VP40 and it also occurs in the human 60S ribosome, and (ii) an RPIM found in the RRM of the spliceosomal U1A protein- PIE RNA complex.

Table 4.4 Examples of longer, abundant RPIMs. The left side of the motif represents the amino acid (e.g. R = Arg) and the right side represents the ribonucleotides (e.g. rGUGU). The PDB ID column shows example RNA-protein pairs in which the RPIMs are found, e.g., 1FJG_DA, where 1FJG is the PDB ID, D is the protein chain, and A is the RNA chain. A lot of these examples are ribosomal RNA-protein pairs.

RPIMs	**Count**	**PDB IDs**
R:rGUGU	283	1FJG_DA; 1HNW_DA; 1N32_DA
SGR:rAUGG	255	1HNW_CA; 2B64_CA; 4K0L_CA
LGG:rAAC	251	1IBL_BA; 2B64_BA; 3I8G_EA
PNSALRK:rCCAGCA	235	1N36_LA; 3J10_OA; 3BBN_LA
RGG:rUUA	232	1XNR_LA; 2V46_LA; 3IZZ_FD
NQ:rCCGC	217	3UZM_OA; 3V22_LA; 4KD8_LA
PNSA:rGCG	209	2UXB_LA; 3D5A_LA; 4L6K_LA
DGKK:rUGA	204	1VS7_GA; 2AVY_GA; 2AW7_GA
GRVPLH:rGCGC	201	2HGR_FA; 2I2P_CA; 3J18_CA
RPISK:rUGGC	195	1PNX_QA; 4EJA_QA; 3O2Z_F1

4.3.5.1 Ebola Matrix Protein, VP40

A timely and intriguing example is provided by the VP40 matrix protein of the deadly Ebola virus. VP40 is one of only seven proteins encoded by Ebola. In transfected cells, VP40 alone is both necessary and sufficient for the assembly and release of virus-like particles [Bharat et al. (2012)]. It is a multifunctional protein that can adopt at least three different functional conformations. The RNA-binding form of VP40 is an octameric ring (PDB 1H2C; Figure 4.6(a)), which regulates viral transcription in infected cells [Bornholdt et al. (2013)]. For its role in membrane trafficking, VP40 forms a dimeric butterfly-shaped structure (PDB 4LDB; Figure 4.6(b)), which readily oligomerizes to form identical linear filaments. As a matrix protein for building nascent virions, VP40 forms a hexameric structure (PDB 4LDD; Figure 4.6(c)).

We identified 4 different RPIMs in the octameric VP40 structure, 3 of which consist of only two nodes (N:rG ,R:rG, Y:rG). The fourth RPIM (THFGKA:rGA) is longer and occurs only once time in the VP40 complex structure. Based on the analysis of the structure [Gomis-Ruth et al. (2003)], Phe125 (F in the RPIM) and Arg134 (not included in the contiguous sequence motif) are the most important residues for recognition of the guanine (rG in the RPIM) in the single-stranded binding loop. When we searched for other structures that might share this

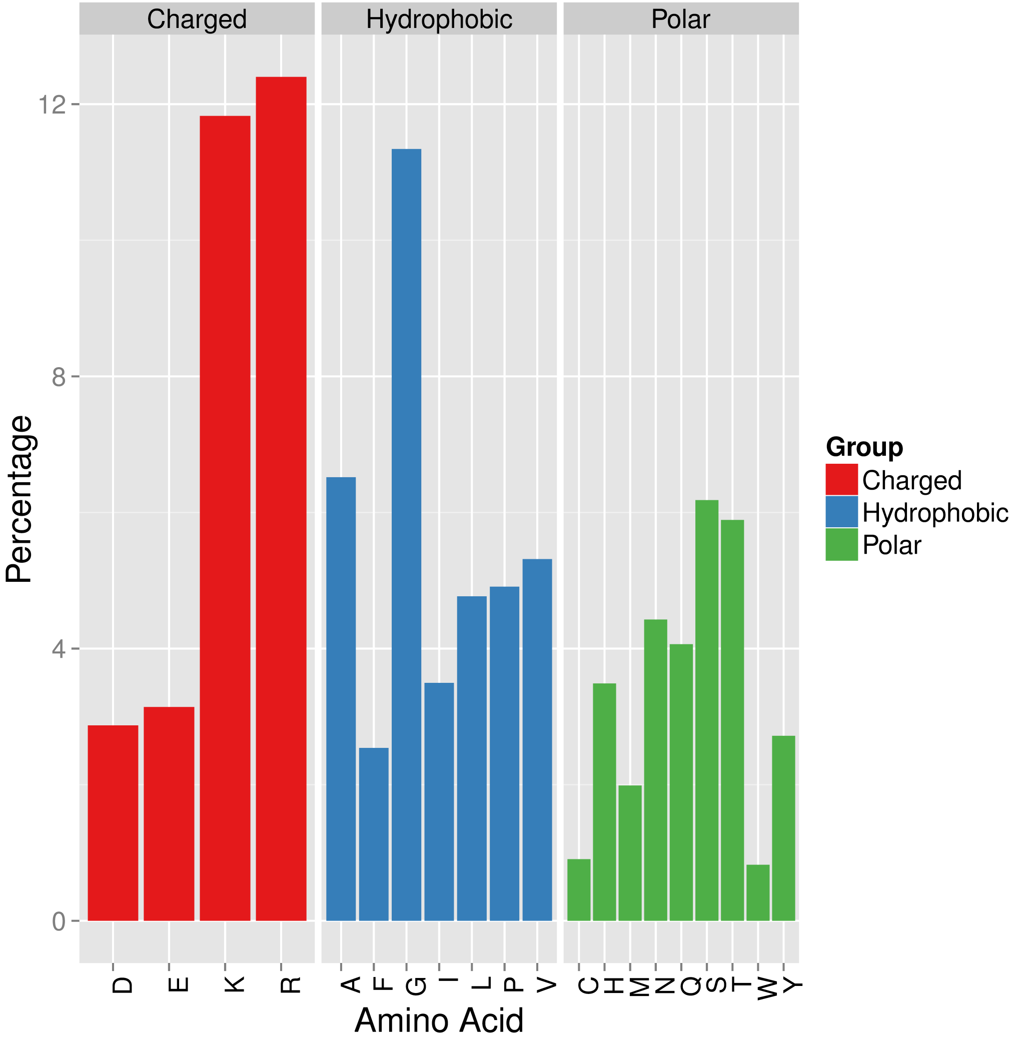


Figure 4.4 Amino acid content of RPIMs.

RPIM, we identified the same motif within the human 60S ribosomal subunit, in which the L18 protein interacts with the rRNA, binding to a single-stranded loop (ES15L) (Figure 4.7) [Anger et al. (2013)].

The L18-rRNA interface contains a close match (HFGKA:rG) to the RPIM in Ebola VP40 (THFGKA:rGA) (Figure 4.7b). Thus, Ebola VP40 protein and the human L18 ribosomal protein share a recognizable RNA-binding sequence motif.

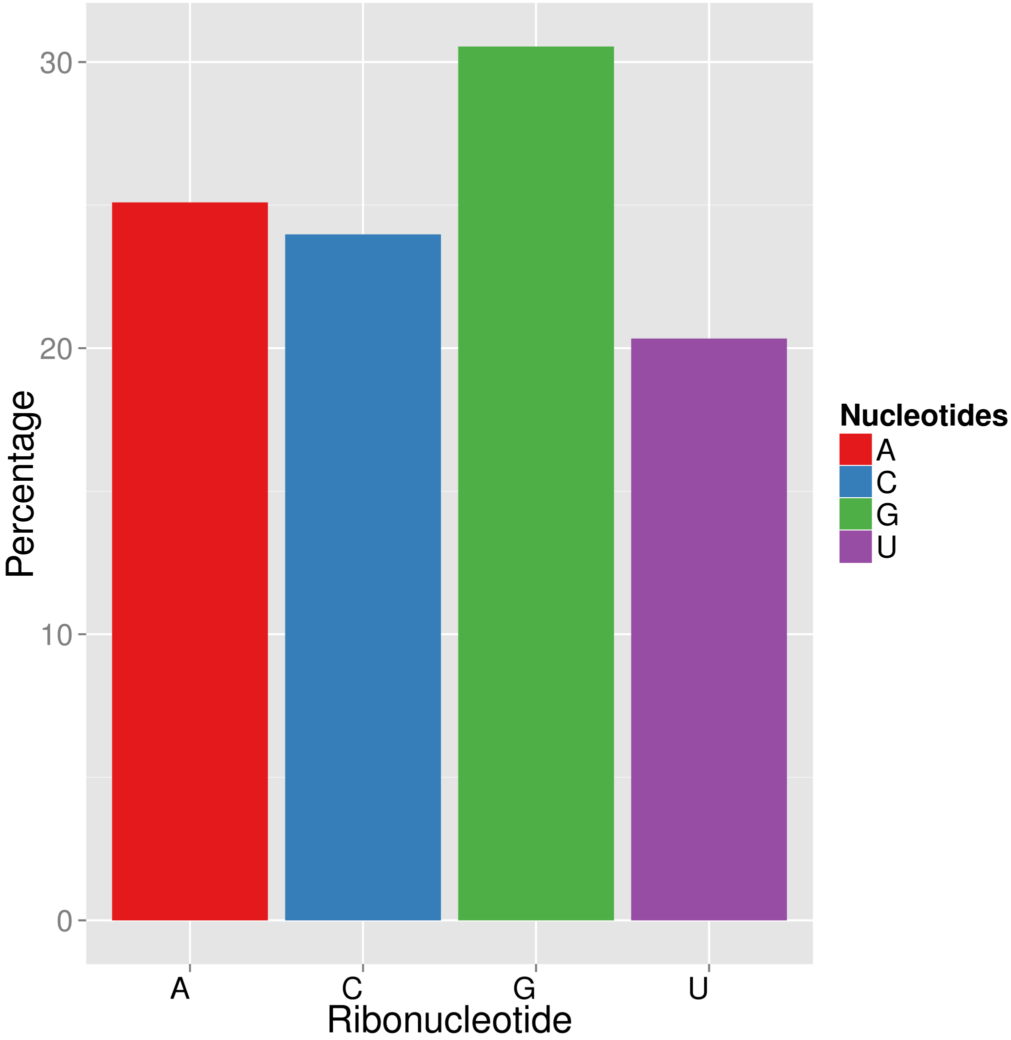


Figure 4.5 Ribonucleotide content of RPIMs.

4.3.5.2 Proteins with RRM Domains

The RNA recognition motif (RRM) is the most abundant RNA-binding domain in eukaryotes [Clery et al. (2008); Chen and Varani (2013); Maris et al. (2005)]. At the primary sequence level, the domain is composed of 80 − 90 amino acids. It consists of two α-helices packed against four anti-parallel β-strands with a $\beta\alpha\beta\beta\alpha\beta$ topology. In mammalian spliceosomal complexes, the U1A protein contains RRM domains. These have a very high binding specificity and recognize the “AUUGCAC” sequence within an RNA hairpin with a sub-nanomolar K_d [Chen and Varani (2013)]. We were able to find this exact RNA sequence in the RPIMs

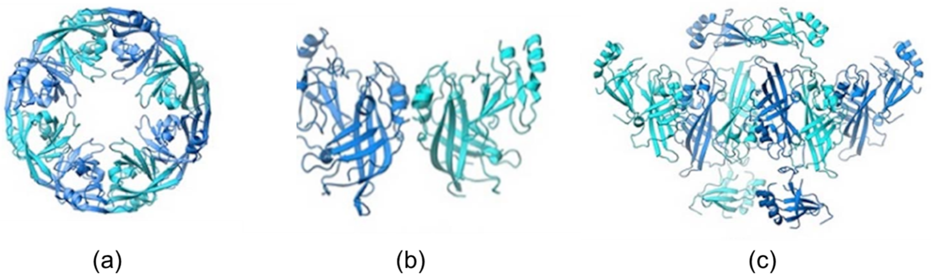


Figure 4.6 Different functional conformations of Ebola VP40. (a) Octameric form, (b) dimeric form, and (c) hexameric form.

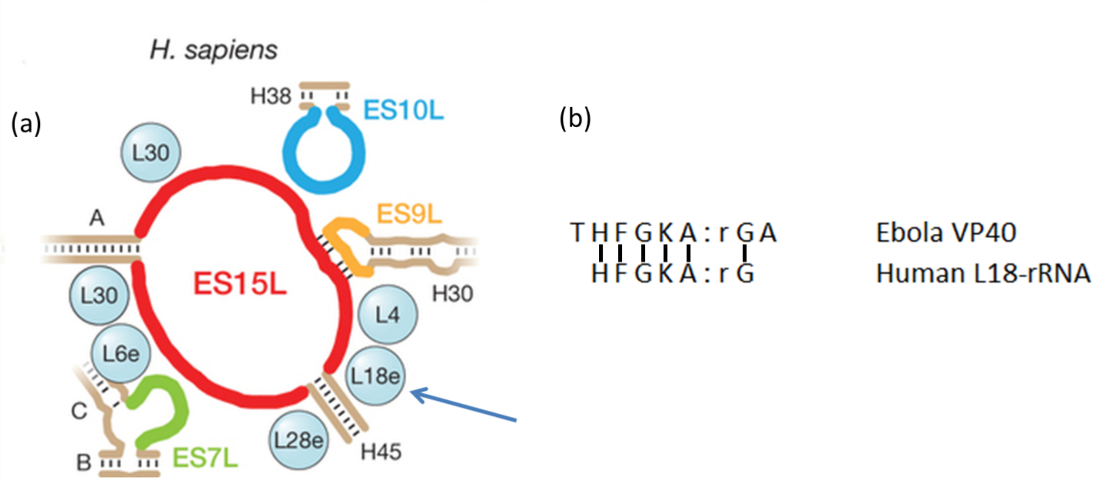


Figure 4.7 (a) Human ribosomal protein L18 (arrow) interacts with the ss loop ES15L of 18s rRNA (red).(Figure source: Figure 4, Anger 2013) (b) The alignment of the RPIMs found in Ebola VP40 and in L18-rRNA.

extracted from the NMR complex structure of the U1A protein-PIE RNA complex (PDB 1DZ5, protein chain B, RNA chain C, and protein chain A, RNA chain D), and identify the amino acids that bind to that sequence within a 5Å cutoff distance. The RPIMs we find are (with specific RNA sequence italicized): (i) HTIYINNLNEKI:rC*AUUGCAC* in 1DZ5_BC; (ii) SLKMRGQAFVIF:rC*AUUGCAC* in 1DZ5_BC; and (iii) SLKMRGQAFVI:rC*AUUGCAC* in 1DZ5_AD. Protein chains A and B share 100% sequence identity, and RNA chains C and D share 100% sequence identity. For the purpose of this discussion, we will use protein chain B and RNA chain C of 1DZ5. Figure 4.8 shows that the RNA sequence recognized by chain B of 1DZ5 is within an RNA hairpin. The RPIMs are able to identify the RNA sequence and amino acid residues involved in recognizing that sequence.

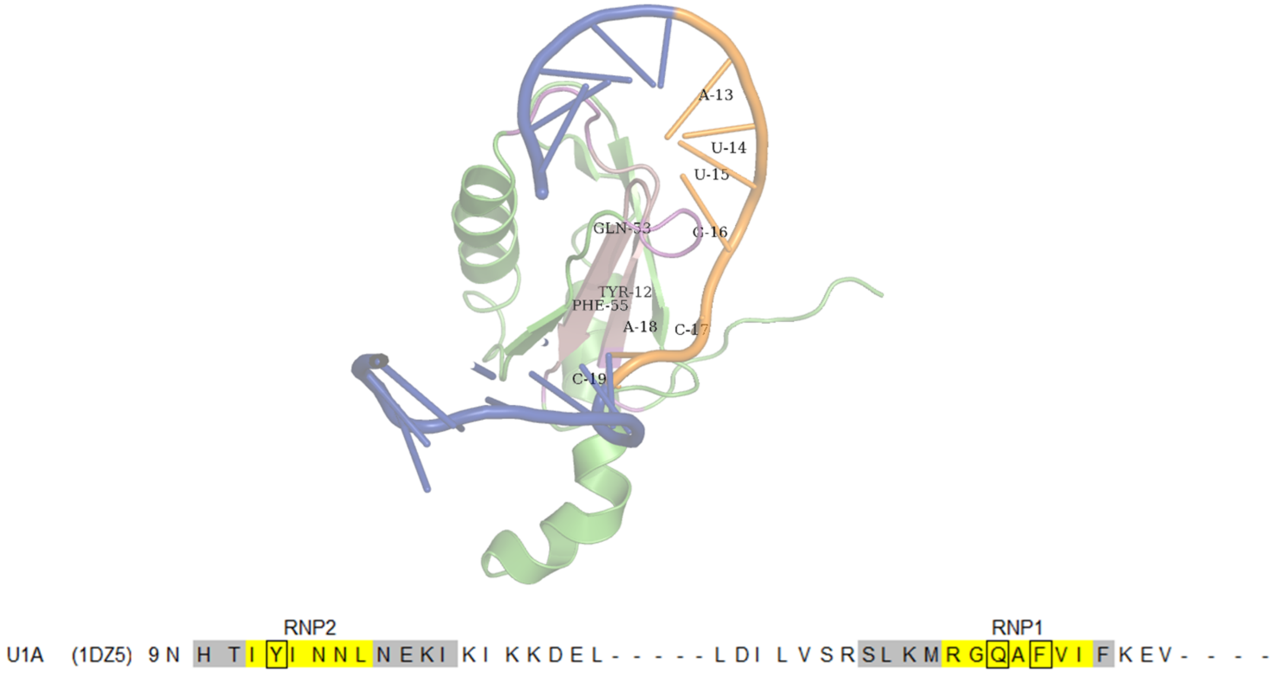

A-13
U-14
U-15
G-16
GLN-53
TYR-12
PHE-55
A-18
C-17
C-19
RNP2
RNP1
U1A (1DZ5) 9 N H T I Y I N N L N E K I K I K K D E L - - - - - L D I L V S R S L K M R G Q A F V I F K E V - - - -

Figure 4.8 Crystal structure of U1A-PIE RNA complex (PDB 1DZ5), showing only protein chain B (green), RNA chain C (blue). The magenta and brown regions of the protein are the interacting amino acid residues. The beta strands in brown are RNP1 (with conserved residues Gln53 and Phe55 marked) and RNP2 (with conserved residue Tyr12 marked), respectively. We show that the ribonucleotide sequence "AUUGCAC" is in the RNA hairpin loop (orange region of RNA, with residues marked). The regions in the protein sequence highlighted in yellow are RNP2 and RNP1. Tyr in RNP2 is an aromatic residue important for primary RNA-binding. Similarily, Gln and Phe in RNP1 are aromatic residues important for primary RNA-binding (Maris, Dominguez, and Allain 2005). The grey highlighted regions in the sequence are the protein sides of the RPIMs we found, interacting with RNA sequence "CAUUGCAC".

Interestingly, we were able to find part of the RPIM shown in Figure 4.8 in another unrelated protein, 4C4W_AD (LKMRGQ:rGCA), which is the structure of a rare non-standard sequence k-turn bound by the L7Ae protein. Kt-23 is a rare RNA kink-turn (k-turn) and L7Ae is a member of a strongly conserved family of proteins that bind a range of k-turn structures [Huang and Lilley (2014)]. The RPIM we find is LKMRGQ:rGCA, which is contained within two of the larger RPIMs found in 1DZ5. Protein chain B of 1DZ5 shares 98% sequence identity with protein chain A of 4C4W. Therefore, it is not surprising that the two share a partial RPIM. This suggests that the mode of interaction between the RNAs and proteins in these two RNA-protein pairs is similar, in part due to their similarity at the sequence level. We also found the RPIM LKMRGQ:rGCA in the crystal structure of the U1A spliceosomal protein complexed with an RNA hairpin (PDB 1URN, protein chain A, RNA chain P).

4.4 Conclusions

We have extracted 33,594 distinct sequence-based RNA-protein interaction motifs (RPIMs) from characterized RNA-protein complexes in the PDB using an image analysis algorithm, *bwlabel* [Haralick and Shapiro (1991)]. This work represents the first such attempt to extract motifs from both the RNA and protein sides of interacting RNA-protein partners. Our preliminary analysis shows that interfaces in RNA-protein complexes share similarities that can be recognized at the primary sequence level.

Future work with RPIMs is aimed at addressing the following interesting questions: (i) Do apparently dissimilar RNA-protein complexes share sequence or structural similarities at their interfaces? (ii) Are RPIMs over-represented in well-characterized RNA-binding domains, such as those annotated in NDB [Coimbatore Narayanan et al. (2014)], Prosite [Sigrist et al. (2013)], or RBPDB [Cook et al. (2011)]? (iii) Can RPIMs be used to define or classify novel RNA-binding domains? Chapter 5 in this thesis addresses the question: Can RPIMs be used to accurately predict whether or not a specific RNA and protein are likely to interact?

4.5 Acknowledgments

We thank members of the Dobbs and Honavar groups for useful discussions, and Carla Mann for critical reading of the manuscript.

CHAPTER 5. PREDICTING RNA-PROTEIN INTERACTION PARTNERS USING RNA-PROTEIN INTERACTION MOTIFS

Manuscript in preparation

Rasna R. Walia, Vasant Honavar, and Drena Dobbs

Abstract

RNA-protein interactions play important roles in cellular processes like protein synthesis, RNA processing, and transcriptional and post-transcriptional regulation of gene expression. Understanding the molecular mechanisms by which proteins recognize and bind RNA is essential for comprehending the functional implications of these interactions, but the recognition "code" that mediates interactions between proteins and RNAs is not yet understood. Success in deciphering this code would dramatically impact the development of new therapeutic strategies for intervening in devastating diseases such as AIDS and cancer.

With the recent advent of high-throughput *in vitro* and *in vivo* methods for identifying RNA-protein interactions, it has been possible to characterize the RNA partners for a rapidly expanding set of RNA-binding proteins, and to identify the protein partners of several RNAs. However, these methods have different advantages and pitfalls. Computational methods that can predict RNA-protein partners are thus viable and cost-effective approaches for reducing the experimental search space for researchers and for identifying potential targets for clinical interventional in genetic and infectious diseases.

We introduce a new and innovative method for predicting RNA-protein interaction partners, RPIMotif. Ten-fold cross-validation results on the training dataset demonstrate that RPIMotif has a high true positive rate and low false positive rate. We compared RPIMotif-ct, which combines RNA-protein interaction motifs (RPIMs) with conjoint triad features, with two existing methods, RPISeq and lncPro. On an independent test dataset containing 11,281 positive examples and 971 negative examples, RPIMotif-ct had a true positive rate of 0.96, false positive rate of 0.29, and MCC of 0.64, compared to RPISeq with true positive rate of 0.97, false positive rate of 0.63, and MCC of 0.40. On the subset of RNA-protein pairs on which lncPro could make predictions, RPIMotif-ct achieved a true positive rate of 0.96, false positive rate of 0.27, and MCC of 0.63, compared to 0.96, 0.63, 0.38 for RPISeq, and 0.77, 0.30, and 0.30 for lncPro. The major advantages of RPIMotif-ct over other methods is that it is a purely sequence-based method that can rapidly make predictions for RNA and protein sequences of any length. The RPIMotif-ct software and related datasets are freely available at https://github.com/Dobbs-Lab/RPIMotif.

5.1 Introduction

It is estimated that most genomes of higher organisms encode as many RNA-binding proteins as DNA-binding transcription factors [Anko and Neugebauer (2012)], yet the former are much less well characterized. Basic questions such as how individual proteins recognize specific RNA molecules still require answers [Cirillo et al. (2014)]. Recent advances in high-throughput technologies (e.g. RIP-Seq, PAR-Clip) [Ascano et al. (2013); Konig et al. (2012); Rinn and Chang (2012)] have elucidated the RNA targets of several RNA-binding proteins and have begun to increase the currently sparse data on RNA-protein interaction networks. Sequence-based computational methods offer a viable and cost-effective way to predict RNA-protein interaction partners, both guiding and complementing the efforts of wet-lab scientists.

Only a few methods for computationally predicting RNA-protein interaction partners have been developed so far. One of the first methods [Pancaldi and Bahler (2011)] used Support Vector Machines (SVM) and Random Forest (RF) classifiers to predict RNA-binding protein (RBP) targets in yeast. Their method uses several features of the RNA and protein partners, e.g., mRNA half-life and isoelectric point of the protein, which cannot be extracted from the sequences alone, making it difficult to apply to many proteins and RNAs of interest. Additionally, Pancaldi and Bahler's method is not yet available as a webserver.

catRAPID [Bellucci et al. (2011)] uses features extracted from sequence information, i.e., predictions of secondary structure, hydrogen bonding, and van der Waals' contributions, to estimate the propensity of binding of a pair of protein and RNA molecules. The catRAPID method was tested on a set of non-RNA binding proteins (known protein- and DNA-binding proteins) randomly paired with RNAs to evaluate its ability to identify non-interacting molecules. A drawback of the catRAPID method is that it can make predictions only on proteins of length 50 – 750 amino acids (aas) and RNAs of length 50 – 1200 nucleotides (nts). Although catRAPID has a webserver available, it is not possible to test large datasets of RNA-protein pairs for interaction. The catRAPID omics [Agostini et al. (2013)] tools allows large-scale interaction predictions of a protein against a whole transcriptome or an RNA against the nucleotide-bindng proteome of a model organism. However, catRAPID omics is not a convenient tool to use for the pairwise prediction of RNA-protein interactions.

RPISeq [Muppirala et al. (2011)] uses conjoint triad features [Shen et al. (2007)] extracted from protein sequences and conjoint tetrads extracted from RNA sequences, without a requirement for non-sequence features, to predict RNA-protein interaction partners. An advantage of RPISeq is that the method is available as a webserver, and a standalone version is available for running large datasets. A major drawback of RPISeq is that it has a high false positive rate [Muppirala (2013b); Cirillo et al. (2014)].

lncPro [Lu et al. (2013)] is another recently published method developed to predict lncRNA-protein interactions. The method uses secondary structure propensities, hydrogen bonding, and van der Waals' propensities from the RNA and protein sequences, transforms them into numerical vectors, combines the vectors using matrix multiplication, and then use the combined vectors to predict whether or not a specific RNA and protein pair will interact. The method is available both as a webserver and a standalone program. lncPro is limited to making predictions on lncRNAs no longer than 4095 nucleotides, because of the limitations of its RNA secondary structure prediction program.

Also recently, Livi and Blanzieri [Livi and Blanzieri (2014)] developed a suite of methods to predict protein-specific mRNA-binding using sequence alone (the Oli method), a binding motif score (the OliMo method), and a predicted secondary structure for the RNA (the OliMoSS method). All three methods are trained using only sequence-based features of the RNA and are designed to predict RNA-binding partners for a single protein of interest, i.e., a separate SVM classifier is trained for each RBP. The scripts for running the different Oli methods are available from the authors.

Sequence-based methods have been shown to be highly successful in predicting functional residues in proteins, including RNA-binding residues [Puton et al. (2012); Walia et al. (2012)], DNA-binding residues [Ofran et al. (2007); Wang and Brown (2006b); Wang et al. (2010a)] and interfacial residues in protein-protein complexes [Xue et al. (2011)], as well as epitopes [EL-Manzalawy and Honavar (2010)], glycosylation sites [Caragea et al. (2007b)] and others. This suggests that sequence-based methods may also be effective for the more challenging "partner prediction" problem. Indeed, Muppirala et al.'s study [Muppirala et al. (2011)] demonstrated that there is strong enough signal in sequences of putative RNA-protein interaction partners to differentiate pairs that do or do not interact. For the interface prediction problem, sequence-based methods that exploit evolutionary information (in the form of position-specific scoring matrices) or partner-specific approaches that employ a combination of machine learning and sequence homology-based methods actually outperform methods that require structural information [Walia et al. (2012, 2014)]. To date, no methods that utilize sequence or structural features from *both* the RNA and protein RNA-protein complexes have been proposed, for either interface prediction or partner prediction.

We propose a new method, RPIMotif, for predicting RNA-protein interaction partners using RNA-protein interaction motifs (RPIMs) extracted from the sequences of characterized RNA-protein complexes deposited in the Protein Data Bank (PDB). We compare the performance of RPIMotif to that of RPISeq [Muppirala et al. (2011)] and lncPro [Lu et al. (2013)], using an independent test set that includes both positive and negative examples. Compared to lncPro, RPIMotif has a higher true positive rate and comparable false positive rate. Compared to RPISeq, RPIMotif has a significantly reduced false positive rate. We also implement a hybrid method that exploits RPIMs in combination with the conjoint triad features used by RPISeq. We show that this hybrid method, RPIMotif-ct, achieves a high accuracy, with both a higher true positive rate and lower false positive rate than other available methods.

5.2 Methods

5.2.1 Datasets

For training our classifier, we utilized the RPI2241 dataset compiled by Muppirala et al. [Muppirala et al. (2011)]. The dataset contains 2,241 experimentally validated RNA-protein pairs extracted from RNA-protein complexes in the PDB. No two proteins or RNAs share greater than 30% sequence identity. Only proteins with $\geq$ 25 amino acids and RNAs with $\geq$ 15 nucleotides were included in the dataset. An RNA-protein pair is defined as interacting if it contains at least one amino acid and ribonucleotide that lie within an 8Å distance cut-off. An equal number of non-interacting RNA-protein pairs obtained from Muppirala et al. [Muppirala et al. (2011)] was utilized as the negative dataset.

We evaluated the performance of classifiers using an independent test set. To obtain a positive set of interacting RNA-protein pairs, we used the NPInter v2.0 database [Wu et al. (2006); Yuan et al. (2014)], which contains experimentally validated functional interactions between non-coding RNAs and other biomolecules. We extracted 11,281 unique RNA-protein pairs from the "binding RNA-protein" and "regulatory RNA-protein" categories. The set of negative examples was obtained by pairing 971 non RNA-binding proteins [Kumar et al. (2011)] with RNAs from fRNAdb [Kin et al. (2007)], a comprehensive database of non-coding RNA sequences. Specifically, we used the ncRNA dataset from fRNAdb, which excludes miRNAs and short reads. In summary, our independent test set, RPI12252, contains 11,281 positive examples, i.e., known RNA-protein pairs and 971 negative examples, i.e., proteins which are not known to bind to RNAs (data available at `https://github.com/Dobbs-Lab/RPIMotif`).

5.2.2 Data Representation

We utilized RNA-protein interaction motifs (RPIMs) extracted from RNA-protein complexes deposited in the PDB as of March 30, 2014 (see Chapter 4). The collection includes 33,594 distinct RPIMs extracted from 23,017 RNA-protein pairs. Figure 5.1 shows an example of an RPIM. Each RNA-protein pair is scanned for the presence of both the RNA and protein side of an RPIM; if found, the corresponding position in the feature vector is marked with 1,

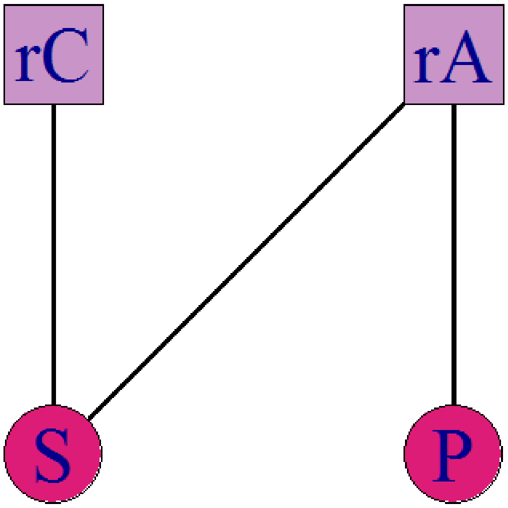
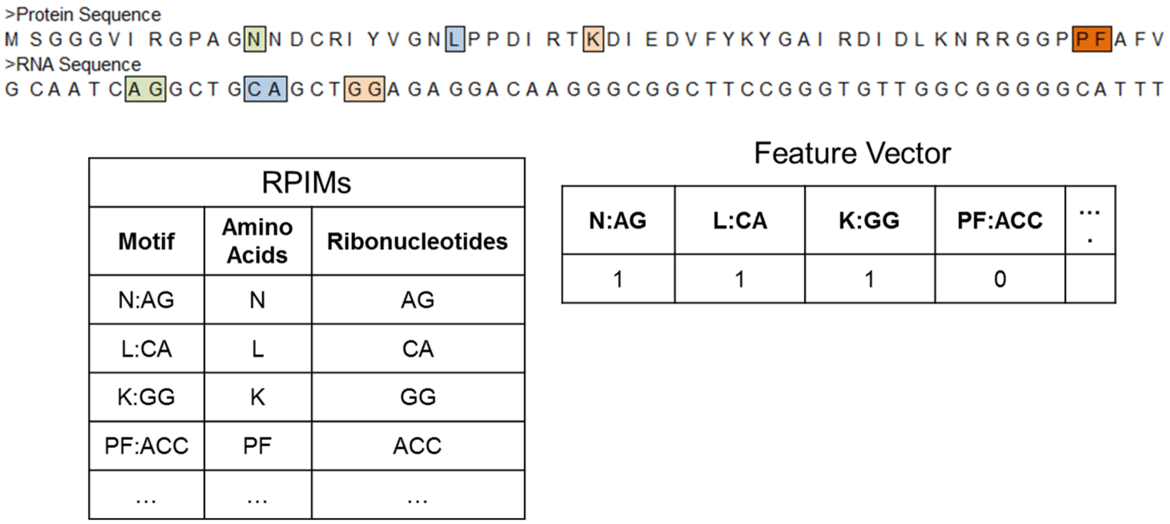


Figure 5.1 An example of an RNA-protein interaction motif (RPIM). Circles represent amino acids and squares represent ribonucleotides. The RPIM above is found in 6 PDB structures: 2NUF_AD, 2NUF_BC, 4A18_H1, 4A19_H1, 4A1B_H1, and 4A1D_H1, where the first four characters represent the PDB ID, the first character after the underscore is the protein chain id, and the next character is the RNA chain id.

else 0 (see Figure 5.2).

We also evaluated the use of conjoint triad features (ct) previously utilized in RPISeq by Muppirala et al. [Muppirala et al. (2011)]. The conjoint triad feature was originally proposed for predicting protein-protein interactions by Shen et al. [Shen et al. (2007)]. In this method, the 20 amino acids are classified into seven groups according to their dipole moments and the volume of their side chains. Protein sequences are encoded using the resulting 7-letter reduced alphabet. Each protein sequence is thus represented by a 343 (7 x 7 x 7) dimensional vector, where each element in the vector corresponds to the normalized frequency of a particular amino acid 3-mer in the sequence. RNA sequences are encoded using a 256 (4 x 4 x 4) dimensional


>Protein Sequence
M S G G G V I R G P A G N N D C R I Y V G N L P P D I R T K D I E D V F Y K Y G A I R D I D L K N R R G G P P F A F V
>RNA Sequence
G C A A T C A G G C T G C A G C T G G A G A G G A C A A G G G C G G C T T C C G G G T G T T G G C G G G G G C A T T T

RPIMs		
Motif	**Amino Acids**	**Ribonucleotides**
N:AG	N	AG
L:CA	L	CA
K:GG	K	GG
PF:ACC	PF	ACC
…	…	…

Feature Vector

N:AG	**L:CA**	**K:GG**	**PF:ACC**	**….**
1	1	1	0	

Figure 5.2 Creation of feature vectors by scanning RNA and protein sequences using RPIMs. In this example, the protein and RNA "sides" of an individual RPIM are in boxes of the same color. Because only the protein side (PF) of the fourth motif (PF ACC) was found in the RNA-protein pair, the corresponding position in the feature vector is 0.

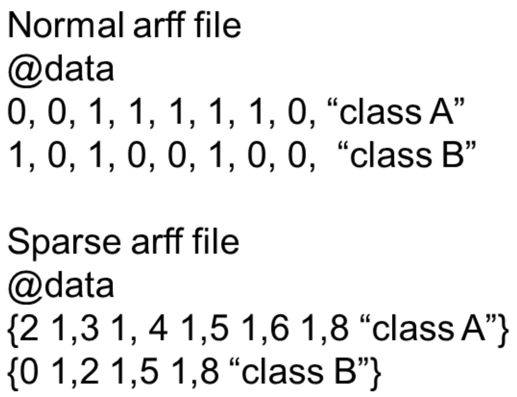
vector, in which each feature represents the normalized frequency of a particular ribonucleotide 4-mer appearing in the RNA sequence (see Muppirala et al. (2011) for more details).

We represented the feature vectors of RNA-protein pairs in the dataset using sparse arff files. These are similar to standard arff files, except that attributes with value 0 are not explicitly represented. Also, the non-zero attributes are identified by attribute number with their value stated (see Figure 5.3).

5.2.3 Classifiers

A random forest (RF) classifier [Breiman (2001)] is an ensemble learning method commonly used for classification and regression problems. RF classifiers grow many decision/classification trees, and each tree gives a vote for which class a new instance belongs to. The RF classifier then outputs the consensus class among the different trees.

In our work, we utilized the RF implementation in Weka 3-7-9 [Hall et al. (2009)]. Weka's implementation of the RF method uses $floor(log_2(m)) + 1$ as the number of randomly chosen attributes to select from at a particular node, where m is the total number of attributes in the



```
Normal arff file
@data
0, 0, 1, 1, 1, 1, 1, 0, “class A”
1, 0, 1, 0, 0, 1, 0, 0,  “class B”

Sparse arff file
@data
{2 1,3 1, 4 1,5 1,6 1,8 “class A”}
{0 1,2 1,5 1,8 “class B”}
```

Figure 5.3 Comparison of standard versus sparse arff files. In the normal arff file, each attribute is listed, even if the value is zero (top). In sparse arff files, only non-zero attributes are listed along with the index of the attribute (bottom). E.g. attribute 0 has a value of 0, attribute 1 has a value of 0, and attribute 2 has a value of 1. This is shown in the normal arff file. In the sparse arff file, we start by listing attribute 2 and its corresponding value 1, since that is the first non-zero attribute.

data, including the class label. We experimented with different tree sizes, and found that 100 trees gave the best performance (data not shown).

5.2.4 Performance Evaluation

The performance measures on the training set are computed based on 10-fold cross-validation experiments. k-fold cross-validation [Mitchell (1997)] is an evaluation scheme for estimating the generalization accuracy of a predictive algorithm (i.e. the estimated accuracy of the predictive model on the test set).

We report the following measures (described in Baldi et al. (2000)) to assess the performance of our classifiers:

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall (True Positive Rate)} = \frac{TP}{TP + FN}$$

$$\text{False Positive Rate (FPR)} = \frac{FP}{FP + TN}$$

$$\text{F-measure} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

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

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}$$

where true positives are denoted by TP, false positives by FP, true negatives by TN, and false negatives by FN. We also report the Area Under the ROC Curve (AUC) value, which is the probability that a classifier gives a higher score to a positive instance than to a negative instance. An AUC of 0.5 indicates a random discrimination between the positive and negative class, whereas an AUC of 1.0 indicates perfect discrimination. The ROCR package [Sing et al. (2005)] in R was used to generate all ROC plots.

5.3 Results and Discussion

5.3.1 Comparison of Partner-Prediction Performance using RPIMs versus Conjoint Triad Features

5.3.1.1 Performance on Training Dataset

Muppirala et al. [Muppirala et al. (2011)] predicted RNA-protein partners using conjoint triad features to train an RF classifier using 20 trees. In order to evaluate the performance of different feature representations using the training set, RPI2241, we performed 10-fold cross-validation. Table 5.1 and Figure 5.4 show these results, in comparison with the results obtained using the retrained version of RPISeq. The RPIMotif method was trained using different numbers of motifs. Specifically, if we exclude all RPIMs containing only one protein node and one RNA node (two-node RPIMs), we are left with 33,495 RPIMs (Table 5.1, row 1) (see Chapter 4 for details). If we both exclude two-node RPIMs, and consider only those RPIMs that appear more than once in the dataset of RPIMs, we are left with 17,910 RPIMs that contain ≥ 3 nodes (Table 5.1, row 2).

Combining RPIMs with conjoint triad features (+ct) increases the TPR (true positive rate), Precision, F-measure, MCC, and overall accuracy of the predictor (Table 5.1), as well as

Table 5.1 Comparison of 10-fold cross-validation results of RPISeqMod* versus RPIMotif classifiers training using different numbers of RPIMs. The best values in each column are shown in bold fold. To allow fair and direct comparison, the RPISeq algorithm was re-trained using the RPI2241 dataset used in this study; thus the RPISeqMod* model employed here is different from the original RPISeq model represented by Muppirala et al.

Method	**TPR**	**FPR**	**Precision**	**F-measure**	**MCC**	**Accuracy %**
RPIMotif 33495	0.78	**0.09**	0.89	0.84	0.70	85
RPIMotif 17910	0.79	0.10	0.89	0.84	0.70	85
RPIMotif 33495+ct	0.85	0.10	0.90	0.88	0.76	88
RPIMotif 17910+ct	0.87	**0.09**	0.90	0.88	0.77	89
RPISeqMod*	**0.88**	**0.09**	**0.91**	**0.90**	**0.79**	**90**

increasing the AUC of the ROC curve (Figure 5.4). RPIMotif appears to perform slightly better when using 17,910 motifs compared to 33,495 motifs. This suggests that RPIMs occurring only once in the 23,017 RNA-protein chains from which they were extracted are, not surprisingly, of little or no predictive value in discriminating between interacting and non-interacting RNA-protein pairs. Using conjoint triads alone (RPISeqMod) has very similar performance to using 17,910 RPIMs combined with conjoint triads.

Based on these results, we tested whether a hybrid classifier that combines the use of conjoint triad features with RPIMs that occur more than once could provide better prediction performance. To do this, we excluded all two-node RPIMs, and created different models using: (i) RPIMs that occur more than once (RPIMotif17910+ct); (ii) RPIMs that occur $\geq$ 5 times (RPIMotif8585+ct); (iii) RPIMs that occur $\geq$ 15 times (RPIMotif4252+ct); (iv) RPIMs that occur $\geq$ 25 times (RPIMotif3106+ct); (v) RPIMs that occur $\geq$ 50 times (RPIMotif1893+ct); and (vi) RPIMs that occur $\geq$ 100 times (RPIMotif767+ct).

Table 5.2 shows the 10-fold cross-validation results obtained using these different variants of RPIMotif. ROC curves are shown in Figure 5.5. As we decrease the total number of RPIMs used, every performance measure shown in Table 5.2 improves. Moreover, the ROC curve (Figure 5.5) shows that RPIMotif767 which uses conjoint triads combined with RPIMs that occur $\geq$ 100 times performs the best, although the overall AUC values for all methods

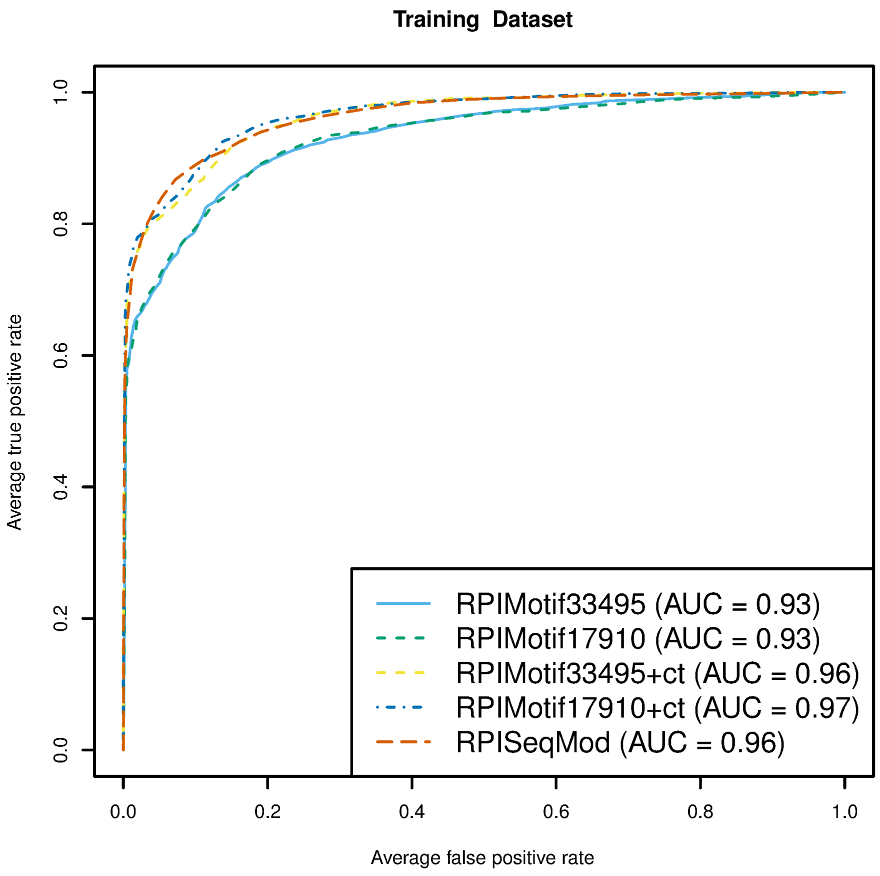


Figure 5.4 ROC curves comparing the performance on training dataset. Methods compared include RPIMotif trained using only RPIMs versus RPIMotif trained using a combination of RPIMs and conjoint triad features (+ct) versus RPISeqMod, which uses only conjoint triad features.

tested are the same, and the curves are very similar. This suggests that using RPIMs that occur more frequently in RNA-protein complexes may be able to provide better performance on independent test sets (see below). We propose that using even more frequently occurring RPIMs will further improve performance because such RPIMs capture characteristic recurring recognition interfaces between RNA-binding proteins and RNAs. RPIMs that occur only once could be an artifact, for example, due to the use of a synthetic peptide to stabilize the three-dimensional structure of the complex for X-ray crystallography.

Table 5.2 Classification performance metrics based on 10-fold cross-validation experiments for RPIMotif using conjoint triad features combined with different numbers of RPIMs. The best values in each column are shown in bold font.

Method	**Frequency of RPIMs**	**TPR**	**FPR**	**Precision**	**F-measure**	**MCC**	**Accuracy %**
RPIMotif 17910+ct	$\geq$ 2 times	0.87	0.09	0.90	0.88	0.77	89
RPIMotif 8585+ct	$\geq$ 5 times	0.87	0.09	0.91	0.89	0.78	89
RPIMotif 4252+ct	$\geq$ 15 times	0.88	0.08	0.92	0.90	0.80	90
RPIMotif 3106+ct	$\geq$ 25 times	0.88	**0.07**	0.92	0.90	0.81	90
RPIMotif 1893+ct	$\geq$ 50 times	**0.89**	**0.07**	**0.93**	**0.91**	0.82	**91**
RPIMotif 767+ct	$\geq$ 100 times	**0.89**	**0.07**	**0.93**	**0.91**	**0.83**	**91**
RPISeqMod*		0.88	0.09	0.91	0.90	0.79	90

5.3.1.2 Performance on an Independent Test Dataset

We compared the performance of different versions of RPIMotif with RPISeq [Muppirala et al. (2011)] and lncPro [Lu et al. (2013)] on an independent test set, RPI12252, which contains 11,281 positive examples and 971 negative examples (see 5.2 for more details). Because lncPro cannot make predictions on RNA sequences longer than 4095 nucleotides, we report comparisons with lncPro using a subset of RPI12252, designated RPI10199. RPI10199 consists of 9,248 positive examples, and 951 negative examples.

Table 5.3 compares different versions of RPIMotif with RPISeq on the independent test set, RPI12252. Interestingly, every version of RPIMotif evaluated here has a significantly lower false positive rate (ranging from 0.29-0.51) than RPISeq (0.63), even though the false positive rates obtained using the training dataset, RPI2241, were very similar for RPIMotif (0.07 – 0.10) and RPISeq (0.09) (see Tables 5.1 and 5.2). One likely reason for this is that the training dataset was generated from solved RNA-protein complexes in the PDB, whereas the test dataset includes RNA-protein interactions determined using high-throughput methods such as PAR-Clip. The training and test datasets are expected to have different properties because the PDB database is known to be biased in several ways, e.g., almost 40% of RNA-protein complexes in the PDB correspond to ribosomes; many RNA-protein complexes in the PDB contain only small segments of RNA rather than the full length native RNA, and despite

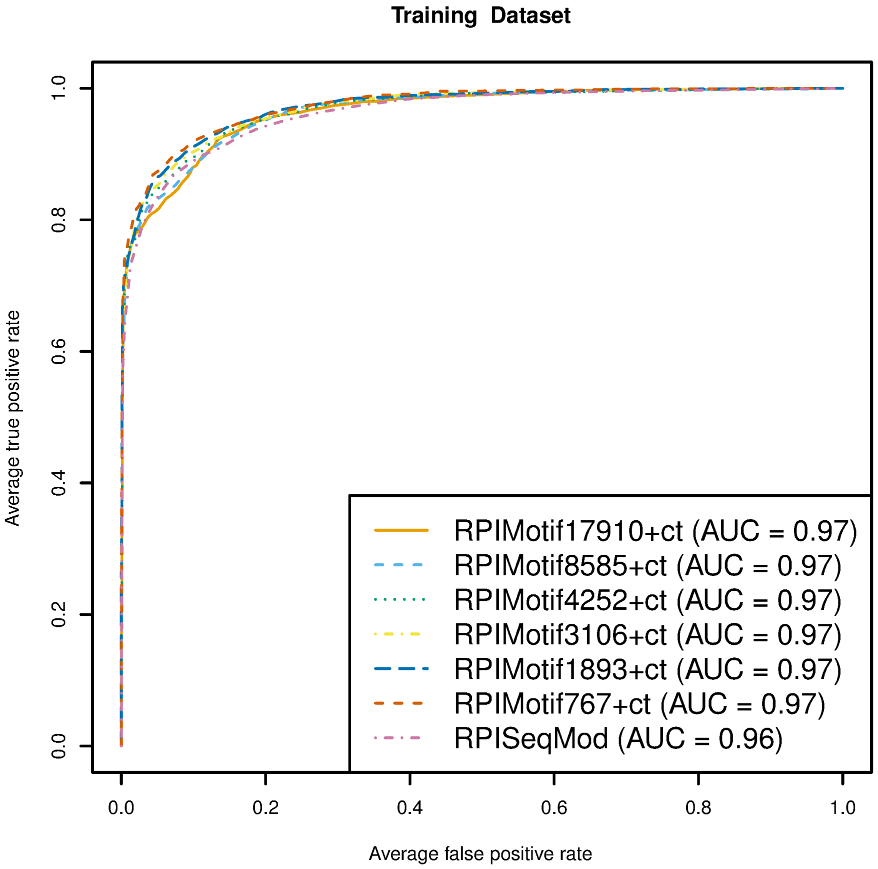


Figure 5.5 ROC curves comparing the performance on training dataset. Methods compared include RPIMotif trained using only RPIMs versus RPIMotif trained using a combination of RPIMs and conjoint triad features (+ct) versus RPISeqMod, which uses only conjoint triad features.

serious efforts, many RNPs have been recalcitrant to crystallization or analysis by NMR. This last observation strongly suggests that physicochemical properties of RNA-protein complexes represented in the PDB differ from those of many RNA-protein complexes for which it has not been possible to obtain high-resolution structures.

Based on the performance of RPIMotif on the RPI12252 test dataset, RPIMs are better features for distinguishing negative from positive examples than conjoint triads, and have better generalizability. The version of RPIMotif that uses 17,910 RPIMs combined with conjoint triads has the lowest FPR (0.29), with high precision, F-measure and accuracy values (0.97, 0.97, 0.94 respectively), as well as the highest MCC (0.64). Note that as we decrease the total number of

Table 5.3 Comparison of methods on an independent test set, RPI12252. The best values in each column are shown in bold font.

Method	TPR	FPR	Precision	F-measure	MCC	Accuracy %
RPIMotif 17910+ct	0.96	**0.29**	**0.97**	**0.97**	**0.64**	**94**
RPIMotif 8585+ct	0.96	0.31	**0.97**	**0.97**	0.62	**94**
RPIMotif 4252+ct	0.97	0.39	**0.97**	**0.97**	0.59	**94**
RPIMotif 3106+ct	0.96	0.36	**0.97**	**0.97**	0.59	**94**
RPIMotif 1893+ct	0.97	0.45	0.96	**0.97**	0.56	**94**
RPIMotif 767+ct	**0.98**	0.51	0.96	**0.97**	0.53	**94**
RPISeq	0.97	0.63	0.95	0.96	0.40	92

RPIMs used by RPIMotif, however, the false positive rate increases. The area under the ROC curve (AUC) (Figure 5.6) of all RPIMotif methods evaluated on RPI12252 is 0.92. RPISeq has a lower AUC of 0.87.

5.3.2 Comparison with Other Methods

We compared the top performing RPIMotif method, RPIMotif17910+ct (hereon referred to as simply RPIMotif-ct) with previously published lncPro and RPISeq methods on the benchmark dataset, RPI10199. lncPro [Lu et al. (2013)] uses predicted secondary structure, hydrogen bonding, and van der Waals' propensities of both RNA and protein sequences to encode feature vectors for a specific RNA-protein pair and thus relies on RNA and protein secondary structure prediction software. Because of this, a current limitation of lncPro is that it cannot make predictions on RNA sequences longer than 4095 nucleotides. RPIMotif-ct uses only RPIMs and conjoint triad features as input, both of which are easy to obtain. For RPIMotif-ct, we compared different classification thresholds, θ, varying the values from 0.1 to 1, with a step size of 0.1. We found that $\theta = 0.5$ was optimal, based on Matthews Correlation Coefficient (MCC, data not shown).

As shown in Table 5.4, on the RPI10199 dataset, RPIMotif-ct consistently showed the best performance, based on every metric evaluated. It also gave the highest AUC value of 0.91, compared with 0.85 for RPISeq, and 0.82 for lncPro (Figure 5.7). Notably, RPIMotif-ct had a much lower false positive rate (0.29) compared with RPISeq (0.63). lncPro also had a relatively low false positive rate (0.30). Both RPIMotif-ct and RPISeq had high true positive rates of

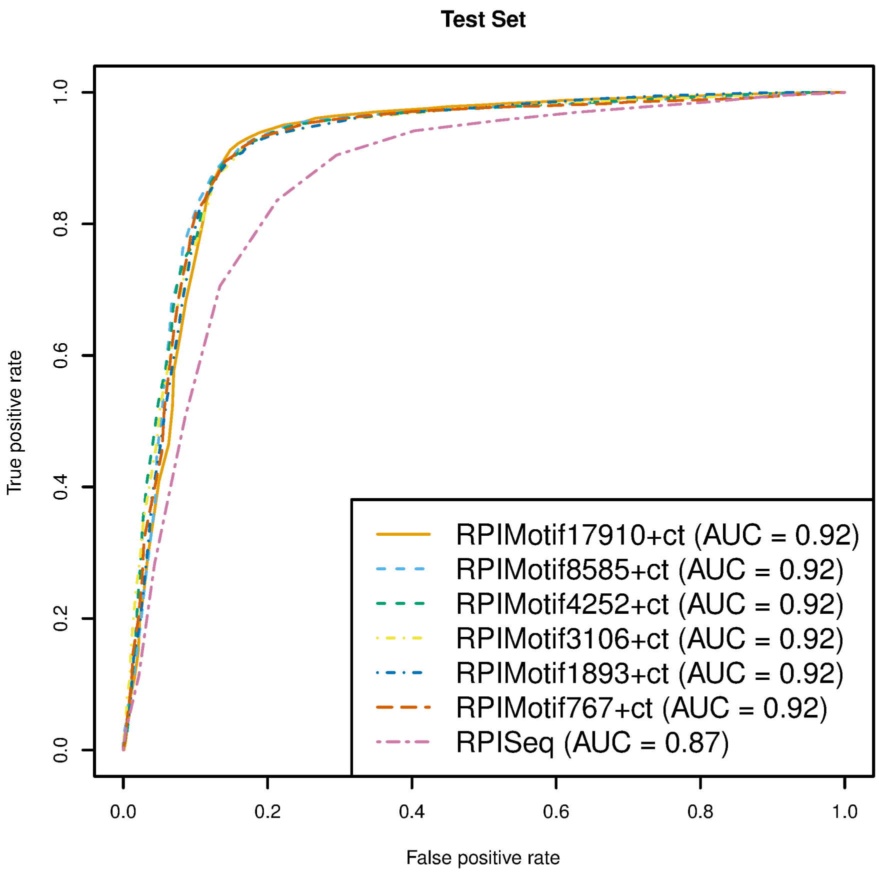


Figure 5.6 ROC curves of different RNA-protein partner predictors on RPI12252.

0.96, considerably higher than that of lncPro (0.77). These results indicate that for predicting whether or not an RNA-protein pair will interact, combining RPIMs with conjoint triads gives excellent performance, with both higher true positive rates and lower false positive rates than using conjoint triads alone. Thus, two comparatively simple sequence-based features, RPIMs and conjoint triads, are highly predictive, and can provide partner-prediction performance that is comparable to or better than the use of more complex features such as predicted secondary structures, hydrogen bonding propensities, and van der Waals' propensities.

5.4 Conclusions

We have shown that RPIMs, which are RNA-protein interaction motifs that include interfacial residues from both the RNA and protein sides of interfaces extracted from structurally

Table 5.4 Comparison of RPIMotif-ct with RPISeq (Muppirala2011) and lncPro (Lu2013) on RPI10199. The best values in each column are shown in bold font.

Method	TPR	FPR	Precision	F-measure	MCC	Accuracy %
RPIMotif-ct	**0.96**	**0.29**	**0.97**	**0.96**	**0.63**	**93**
RPISeq	**0.96**	0.63	0.94	0.95	0.38	91
lncPro	0.77	0.30	0.96	0.85	0.30	76

characterized RNA-protein complexes, can be used to reliably predict RNA-protein interaction partners. Among several methods for predicting RNA-protein interaction partners evaluated in this study, the best method, RPIMotif-ct, uses both RPIMs and conjoint triad features. Although RPISeq, which uses conjoint triads on their own, has been shown to predict RNA-protein interaction partners with high accuracy [Muppirala et al. (2011)], we show here that combining conjoint triads with RPIMs leads to improved performance, especially in terms of the false positive rate, which is substantially lower for RPIMotif-ct than RPISeq. On an independent test set, RPIMotif-ct also outperforms lncPro [Lu et al. (2013)], which uses more complex and predicted features based on sequences of potentially interacting RNAs and proteins. Future work includes using RPIMs to build a two-stage classifier, where the first stage predicts whether or not a putative RNA-protein partner interacts, and the second stage predicts the protein-binding residues on the RNA sequence and RNA-binding residues on the protein sequence (partner-specific interface residue prediction). RPIMotif-ct, along with related datasets presented in this work, is freely available at `https://github.com/Dobbs-Lab/RPIMotif`.

5.5 Acknowledgments

We thank members of the Dobbs and Honavar groups for useful discussions, and Carla Mann for her critical comments on the manuscript. We would also like to thank Qiongshi Lu and Tingting Li for helpful suggestions in running the standalone version of lncPro.

Author Contributions

RRW conceived the study (with DD and VH), carried out the experiments, implemented the methods, and prepared an initial draft of the manuscript. DD and VH contributed to the

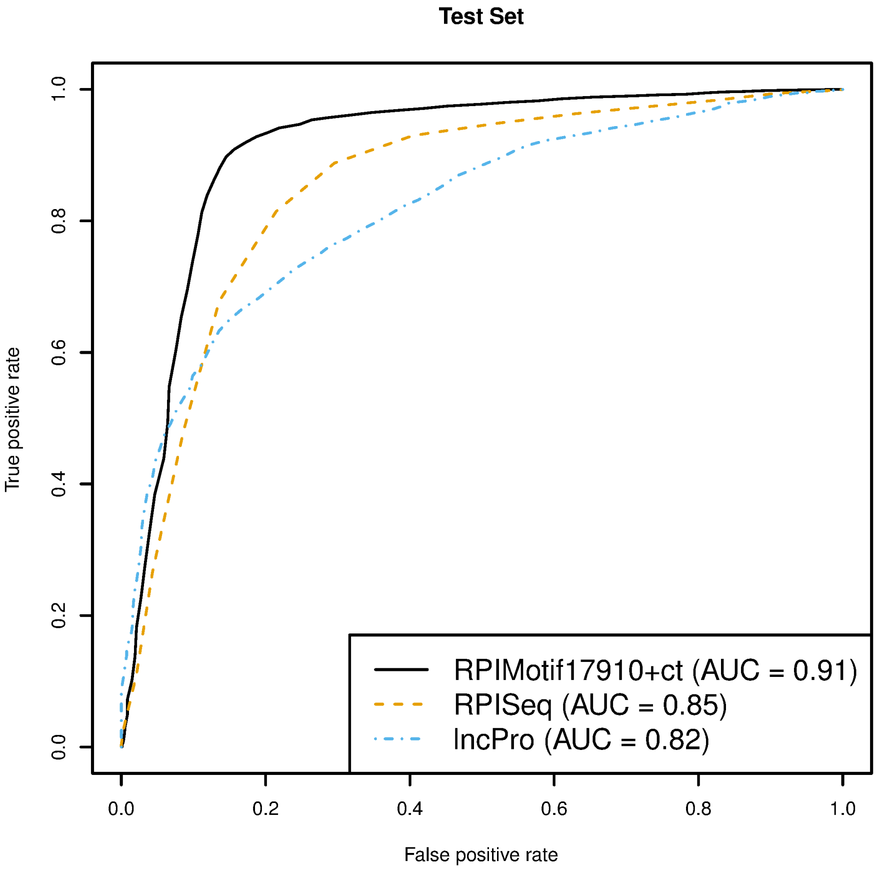


Figure 5.7 ROC curves comparing the performance of RPIMotif-ct, RPISeq, and lncPro on RPI10199.

experimental design, supervised the work, and edited the manuscript. All authors read and approved the final manuscript.

CHAPTER 6. CONCLUSIONS AND FUTURE DIRECTIONS

RNA-protein interactions play diverse and essential roles in living systems. Recently, several surprising new roles for RNA-protein interactions have emerged, following the discovery that the human genome is pervasively transcribed and produces thousands of non-coding RNAs (ncRNAs) [Charon et al. (2010); Clark et al. (2011); Kloetgen et al. (2014); Khalil and Rinn (2011); Zhao et al. (2013)]. Some of the unanticipated roles of ncRNA [Fatica and Bozzoni (2014)] include the regulation of chromatin structure by long non-coding RNA (lncRNA) [Rinn and Chang (2012); Geisler and Coller (2013)], the regulation of mircoRNA (miRNA) function by circular RNAs (circRNAs) (circular RNAs) [Jeck and Sharpless (2014)], and implications in diseases such as cancer and Alzheimer's [Khalil and Rinn (2011); Esteller (2011)]. It now appears that "dark matter" or "junk" DNA of the genome actually represents a vast resource for gene regulatory functions [Flintoft (2010); Kapranov and St. Laurent (2012)]. But, at present, the cellular roles of many ncRNAs and the RNA targets of many RNA-binding proteins are unknown. High-throughput technologies now allow for the identification of recognition sequences for RNA-binding proteins, but the current state of knowledge regarding RNA-protein complexes and RNA-protein interaction networks lags far behind our understanding of DNA-protein complexes and protein interaction networks.

Tools for analyzing RNA-protein complexes and predicting RNA-protein interactions and networks will be valuable for annotating the functional roles of ncRNAs and RNA-binding proteins. The work presented in this thesis focuses on the development of tools and methods for improving the prediction of RNA-protein interactions, both in terms of interface residue prediction and partner prediction. Integration of computational approaches with molecular, genetic, and biophysical experiments will be essential for deciphering the “recognition code” that mediates interactions between RNAs and proteins, and for comprehending the functional implications of these interactions, including their roles in disease.

6.1 Contributions

6.1.1 An assessment of state-of-the-art methods for predicting RNA-binding residues in proteins

Initially, we undertook a comprehensive review, systematic comparison, and critical assessment of RNA-protein interface residue predictors trained using standardized approaches on three carefully curated, non-redundant datasets. The study directly compared two widely used machine learning algorithms (Naïve Bayes and Support Vector Machine) using three different data representations in which features were encoded using either sequence or structure windows. The study showed that RNA-binding site predictors that use PSSM-based (position-specific scoring matrix) encoding of sequence windows outperform classifiers that use other encodings. Also, while structure-based methods that exploit geometric features can yield significant increases in the precision of classifiers, such increases are offset by decreases in sensitivity/recall. Our study also concluded that for rigorous benchmark comparisons of methods for predicting RNA-binding residues, the following are important factors to consider: (i) the rules used to define interface residues; (ii) the redundancy of datasets used for training; (iii) the details of the evaluation procedures (i.e., cross-validation, performance metrics used, and residue- versus protein-based evaluation). We implemented the best performing sequence-based method, PSSMSeq, as a webserver, which is freely available at http://einstein.cs.iastate.edu/RNABindR/.

6.1.2 A sequence-based classifier for predicting RNA-binding residues in proteins and a webserver, RNABindRPlus

We developed a novel sequence-based classifier for predicting RNA-binding residues in proteins. The method combines a sequence homology-based method (HomPRIP) and an optimized machine learning (SVM) approach. This new method, RNABindRPlus, outperforms several state-of-the-art predictors of RNA-binding sites that use either only sequence or only structural information, or both. The advantage of our method is that it is purely sequence-based. A web-based implementation of the method is freely accessible at http://einstein.cs.iastate.edu/RNABindRPlus/.

6.1.3 A comprehensive collection of RNA-protein interaction motifs (RPIMs)

We extracted 33,594 RNA-protein interaction motifs (RPIMs) from 1,720 characterized RNA-protein complexes in the PDB as of March 30, 2014. The RPIMs are made up of sequence contiguous interfacial residues from both the RNA and protein sides of interacting RNA-protein pairs. We showed the utility of the RPIMs in identifying similar motifs in proteins with different functions. For example, part of the motif found in the Ebola virus VP40 matrix protein is also found in the L18 protein of the human 60S ribosome.

6.1.4 A method for predicting RNA-protein interaction partners using RPIMs

We implemented a new method, RPIMotif-ct, for predicting RNA-protein interaction partners using RPIMs extracted from characterized RNA-protein complexes and conjoint triad features used in a previously published method, RPISeq [Muppirala et al. (2011)]. We compared the performance of RPIMotif-ct with RPISeq and another recently published method, lncPro [Lu et al. (2013)]. We demonstrated that RPIMotif-ct makes reliable predictions and has a high true positive rate and low false positive rate compared to RPISeq and lncPro. RPIMotif-ct and related datasets are freely available at https://github.com/Dobbs-Lab/RPIMotif.

6.2 Future Directions

6.2.1 Development of a database of RPIMs

Currently, all RPIMs and the RNA-protein pairs with which they are associated are stored in flat files. This makes it difficult to search for similar RPIMs and analyze the relationships among the RNA-protein complexes that contain them. The development of an online resource, incorporating a graph-based database such as Neo4j, would allow the broader research community to utilize and query the dataset of RPIMs to identify similar RPIMs (and consequently, similar interaction patterns) between different RNA-protein complexes and to classify novel

RNA-binding domains in uncharacterized RNA-binding proteins as well as to identify the RNA sequences recognized by them.

6.2.2 Creation of a webserver for RPIMotif-ct

The scripts and instructions necessary for running RPIMotif-ct are currently available from a github repository (https://github.com/Dobbs-Lab/RPIMotif). This is a first step in making the method available to the broader scientific community. However, a user-friendly webserver for the method would make the method of broader utility to the wider scientific community, including wet-lab scientists.

6.2.3 Extraction and analysis of non-contiguous RPIMs

In our work on extracting RPIMs, we have focused initially on motifs comprising interfacial residues that are contiguous on both RNA and protein sides of the interacting molecules. This is a first step in characterizing and creating a collection of such motifs. However, it will be important to characterize motifs composed of *non-contiguous* interacting residues in RNAs and proteins in order to understand the principles of RNA-protein recognition and build better methods for RNA-protein interface and partner prediction.

6.2.4 Improving the prediction of RNA-protein partners

RPIMotif-ct was trained on characterized RNA-protein complexes in the PDB and tested on experimentally validated RNA-protein interactions deposited in NPInter v2.0 [Yuan et al. (2014)] as well as on examples of non-RNA-binding proteins randomly paired with RNAs from fRNAdb [Kin et al. (2007)]. A major limitation for improving the prediction of RNA-protein interaction partners is the lack of "real" experimentally validated negative data. High-throughput experimental methods, such as RIP-Seq and iCLIP, are used to determine the RNA targets of RNA-binding proteins and methods such as quantitative mass spectrometry are used to determine protein partners of RNAs. Databases such as CLIPz and doRiNA store information about experimentally validated RNA-protein interactions, but there is currently no resource that also systematically catalogs the "negative" interaction data, i.e., recording which proteins do not interact with certain RNAs and vice versa. This kind of information will be highly valuable for improving the prediction of RNA-protein partners and interaction networks.

6.2.5 Partner-specific predictions of RNA- and protein-binding residues using RPIMs

Except for a few methods [Choi and Han (2011); Muppirala (2013b)], all of the predictors for RNA-binding site prediction use only features extracted from non-redundant datasets of RNA-binding proteins to generate "non-partner specific" predictions. In many cases, however, the RNA partners of RNA-binding proteins are known. While there have been quite a few methods for predicting RNA-binding sites in proteins, very few methods for predicting protein-binding sites in RNAs have been developed [Gupta (2011); Muppirala (2013b)]. Preliminary results from our group [Muppirala (2013b)] have shown that considering features extracted from the putative partners of RNA-binding proteins improves the specificity of RNA-binding site prediction. We propose to use the RPIMs we extracted to make "partner-specific" predictions of RNA-binding residues in proteins, as well as protein-binding residues in RNAs.

APPENDIX A. SUPPLEMENTARY MATERIAL FOR CHAPTER 2

Table A.1 Part I: Similarities and Differences of Methods Implemented in this Study with other Methods in the Field. The methods numbered using upper case letters are those that we implemented in this study. The methods numbered using Roman numerals are those that are similar to the ones implemented in this study. E.g. (A) IDSeq is our method and (i) RNABindR is a similar method. Abbreviations used include: NB - Naïve Bayes; SVM - Support Vector Machine; NN - Neural Network; P - Polynomial Kernel; R - Radial basis function kernel; S 5 - sequence-based five–fold; S LOO - sequence-based leave-one-out; W 5/10 - window-based five-/ten-fold.

Method	Interface Residue Definition (Å)	Dataset	Structure Resolution (Å)	Sequence Similarity (%)	Window Size (Smoothing Window Size)
(A) IDSeq	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25
(i) RNABindR	5	RB109	≥ 3.5	≤ 30	25
(B) PSSMSeq	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25
(i) Jeong and Miyano 2006	6	RB87	≥ 3.0	≤ 70	15
(ii) PPRINT	6	RB86	≥ 3.0	≤ 70	Not stated
(iii) RISP	3.5	RB147, RB71	≥ 3.5	≤ 30	7
(C) SmoPSSMSeq	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25 (3)
(i) RNAProB	5	RB86; RB109; RB107	$\geq 3.0;\geq 3.5;\geq 3.5$	$\leq 70;\leq 30;\leq 25$	25 (7)
(D) IDStr	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25
(E) PSSMStr	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25
(i) Chen et al. 2011	5	RB79	≥ 3.5	≤ 30	15
(F) SmoPSSMStr	5	RB106, RB144, RB198	≥ 3.5	≤ 30	25 (3)
(G) PSSMSeq + PA	5	RB106	≥ 3.5	≤ 30	25
(i) PiRaNhA	3.9	RB81	≥ 3.0	≤ 70	25

Table A.2 Part II: Similarities and Differences of Methods Implemented in this Study with other Methods in the Field. The methods numbered using upper case letters are those that we implemented in this study. The methods numbered using Roman numerals are those that are similar to the ones implemented in this study. E.g. (A) IDSeq is our method and (i) RNABindR is a similar method. Abbreviations used include: NB - Naïve Bayes; SVM - Support Vector Machine; NN - Neural Network; P - Polynomial Kernel; R - Radial basis function kernel; S 5 - sequence-based five–fold; S LOO - sequence-based leave-one-out; W 5/10 - window-based five-/ten-fold.

Method	E-value for PSSM; Normalization	Classifier	SVM Kernel	Parameter Tuning	Cross-validation
(A) IDSeq		NB, SVM	P R	No	S 5
(i) RNABindR		NB		Yes	S LOO
(B) PSSMSeq	10^{-3}; Logistic	NB, SVM	P R	No	S 5
(i) Jeong and Miyano 2006	10^{-4}	NN			W 10
(ii) PPRINT	10^{-3}	SVM	Not stated	Not stated	W 5
(iii) RISP	10^{-3}	SVM	R	Yes	W 5
(C) SmoPSSMSeq	10^{-3}; Logistic	NB, SVM	P R	No	S 5
(i) RNAProB	10^{-3}; Linear	SVM	R	Yes	W 5
(D) IDStr		NB, SVM	P R	No	S 5
(E) PSSMStr	10^{-3}; Logistic	NB, SVM	P R	No	S 5
(i) Chen et al. 2011	10^{-1}	SVM	R	Yes	W 5
(F) SmoPSSMStr	10^{-3}; Logistic	NB, SVM	P R	No	S 5
(G) PSSMSeq + PA	10^{-3}; Logistic	NB, SVM	P R	No	S 5
(i) PiRaNhA	10^{-3}	SVM	R	Yes	W 5

APPENDIX B. PRIDB: A PROTEIN-RNA INTERFACE DATABASE

Paper originally appeared in Nucleic Acids Research (Database Issue), Vol 39, 2011[1]

Benjamin A. Lewis*, Rasna R. Walia*, Michael Terribilini, Jeff Ferguson, Charles Zheng,
Vasant Honavar and Drena Dobbs

*Contributed equally to the manuscript

Abstract

The Protein-RNA Interface Database (PRIDB) is a comprehensive database of protein-RNA interfaces extracted from complexes in the Protein Data Bank (PDB). It is designed to facilitate detailed analyses of individual protein-RNA complexes and their interfaces, in addition to automated generation of user-defined datasets of protein-RNA interfaces for statistical analyses and machine learning applications. For any chosen PDB complex or list of complexes, PRIDB rapidly displays interfacial amino acids and ribonucleotides within the primary sequences of the interacting protein and RNA chains. PRIDB also identifies ProSite motifs in protein chains and FR3D motifs in RNA chains and provides links to these external databases, as well as to structure files in the PDB. An integrated JMol applet is provided for visualization of interacting atoms and residues in the context of the 3-dimensional complex structures. The current version of PRIDB contains structural information regarding 926 protein-RNA complexes available in the PDB (as of October 10, 2010). Atomic- and residue-level contact information for the entire dataset can be downloaded in a simple machine-readable format. Also, several non-redundant benchmark datasets of protein-RNA complexes are provided. The PRIDB database is freely available online at http://bindr.gdcb.iastate.edu/PRIDB.

[1] Copyright retained by authors

Introduction

Protein-RNA interactions play critical roles in myriad and diverse biological processes, including many recently discovered regulatory functions, in addition to well-studied roles in protein synthesis, DNA replication, regulation of gene expression, and defense against pathogens [Fabian et al. (2010); Hogan et al. (2008); Licatalosi and Darnell (2010); Lorkovic (2009); Lukong et al. (2008); Lunde et al. (2007); Mansfield and Keene (2009); Mittal et al. (2009); Mohammad et al. (2007)]. Despite their importance, structures of protein-RNA complexes have proven difficult to obtain using experimental structure determination methods; such structures constitute only ~1% of structures in the Protein Data Bank (PDB) [Berman et al. (2000)]. For this reason, several computational methods for predicting the interfaces in protein-RNA complexes have been developed [Liu et al. (2010); Murakami et al. (2010); Perez-Cano and Fernandez-Recio (2010b); Maetschke and Yuan (2009); Shazman and Mandel-Gutfreund (2008); Wang et al. (2010a); Terribilini et al. (2006b); Wang and Brown (2006a); Kumar et al. (2008); Wang et al. (2011); Towfic et al. (2010)]. Virtually all such methods require data in the form of information about structurally characterized protein-RNA complexes and their interfaces.

PRIDB is a repository of protein-RNA interface information derived from structures in the PDB. PRIDB is designed to facilitate detailed analyses of individual protein-RNA complexes of interest and rapid identification of interfacial atoms and residues in both the protein and RNA chains of a chosen complex or user-defined set of complexes. In addition, PRIDB can be used to generate datasets of protein-RNA interfaces for machine learning applications, such as the generation of classifiers for predicting interfaces in protein-RNA complexes for which high-resolution structures are not available.

Related databases/servers

To our knowledge, only one other up-to-date and comprehensive online repository of protein-RNA interfaces is currently available: Biological Interaction Database for Protein-Nucleic Acid (BIPA) [Lee and Blundell (2009)]. 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. PRIDB complements BIPA by providing atomic- and residue-level interfacial information for both the RNA and protein chains of complexes, providing previously published reduced-redundancy datasets, and allowing users to make advanced queries and compile custom datasets. Other collections of protein-RNA complexes and related resources include NDB (http://ndbserver.rutgers.edu/) [Berman et al. (1992)], PRID (http://www-bioc.rice.edu/~shamoo/prid.html) [Morozova et al. (2006)], RsiteDB (http://bioinfo3d.cs.tau.ac.il/RsiteDB/) [Shulman-Peleg et al. (2009)], w3DNA (http://w3dna.rutgers.edu/) [Zheng et al. (2009)], NPIDB (http://monkey.belozersky.msu.ru/NPIDB) [Spirin et al. (2007)], ProNIT (http://gibk26.bse.kyutech.ac.jp/jouhou/pronit/pronit.html) [Kumar et al. (2006)], and the RNP Databases (http://rnp.uthct.edu/index.html/). Several excellent databases of protein-DNA interfaces are also available, including PDIdb (http://melolab.org/pdidb/) [Norambuena and Melo (2010)] and hPDI (http://bioinfo.wilmer.jhu.edu/PDI/).

Database Contents

Data Extraction, Interface Definition and Motif Identification

Atomic coordinate information for all 926 protein-RNA complexes in the Protein Data Bank (PDB) on October 2010 was extracted using the REST API advanced search interface. To generate this comprehensive dataset (rRB926), no filters based on sequence redundancy, structure resolution or other criteria were applied (see Non-redundant Benchmark Datasets below). The complex structures in rRB926 were then scanned to identify interacting amino acids and ribonucleotides using two different definitions: 1) a simple distance-based definition in which a given amino acid residue (AA) in a protein chain is defined as interacting with a ribonucleotide (rNT) in an RNA chain if any atom in AA is within a 5 Å radius of any atom in rNT; and 2) a rule-based definition based on that of Allers and Shamoo [Allers and Shamoo (2001)], in which interactions are classified as van der Waals, hydrogen-bonding, hydrophobic or electrostatic interactions, involving specific AAs and rNTs. All such interacting AAs and rNTs are defined as “interface” residues.

ProSite patterns and profiles [Sigrist et al. (2010)] appearing in any of the protein sequences in the database were retrieved using the ScanProsite REST service [de Castro et al. (2006)]. RNA structural motifs were identified in RNA sequences using FR3D’s [Sarver et al. (2008)] pure symbolic search function; specific motif definitions used for these scans are available in the ‘Tutorial and FAQs’ section of the PRIDB online server.

Non-redundant Benchmark Datasets

Because PRIDB is intended to be a comprehensive collection of protein-RNA complexes from the PDB, the rRB926 dataset was not filtered on the basis of redundancy, structure determination method, resolution, or protein/RNA chain length. While it is possible to filter with such criteria using PRIDB’s Advanced Search function, several pre-calculated benchmark datasets, which have been filtered to limit redundancy and to exclude low-resolution structures, are also provided for the user’s convenience. These include two previously published datasets, RB109 [Terribilini et al. (2006b,a)] and RB147 [Terribilini et al. (2007)], as well as a larger, more recently extracted dataset (RB199) [B.Lewis, submitted for publication]. Complete lists of the PDB IDs for protein-RNA complexes in these datasets, in addition to the pre-calculated interface residue statistics, can be readily accessed from the ‘Datasets’ section of the PRIDB homepage.

Implementation and Availability

PRIDB runs on the Apache 2.2 web server, using MySQL 14.14 as a database backend with AJAX and PHP 5 for user interface functions. Functions not requiring use of the database (e.g., calculating interface residues for a user-submitted complex) are implemented using standalone Perl 5 scripts and the BioPerl module [Stajich et al. (2002)]. All PRIDB code is available on request under the Creative Commons Attribution Non-Commercial License. All data currently in PRIDB was obtained from databases or programs which impose no restrictions on academic use.

PRIDB Summary Statistics

As summarized in Table B.1, the current version of PRIDB contains structural information for a total of 926 protein-RNA complexes available in the PDB as of October 10, 2010. These structures contain 9,689 total protein chains, among which there are only 1,174 unique sequences. While this would seem to indicate that most sequences in the database are repeated several times, this is not the case; 395 of the 1174 (34%) sequences appear only once, and 899 (77%) appear less than the 8 times (the "expected" average redundancy). This disparity is due to the large proportion of ribosomal structures in the PDB (and, by extension, in PRIDB); 9 of the top 10 most abundant sequences, each present in more than 70 structures, are ribosomal proteins. The most abundant sequence, repeated more than 100 times, is that of the TRP-responsive attenuation protein, a protein for which numerous multimeric structures have been solved.

Table B.1 PRIDB contents: complexes and chains. *Total number in PRIDB includes redundant complexes, RNA and protein chains (i.e., chains with identical sequences.)

	Total Number in PRIDB*	**Unique**
Protein-RNA Complexes	926	926
Protein Chains	9,689	1,174
RNA Chains	2,074	746

As shown in Table B.2, PRIDB currently contains 1,475,774 amino acid residues. Based on a 5Å distance cutoff definition for interfacial residues, 397,216 of these residues interact with RNA; of 851,853 ribonucleotide residues in PRIDB, 322,858 interact with protein. On average, 38% of the amino acids in the RNA binding proteins directly interact with RNA, and 28% of the ribonucleotides in the bound RNAs directly interact with protein. As before, these averages are skewed by the prevalence of ribosome structures; ribosomal proteins account for approximately 90% of interacting amino acid residues and approximately 60% of interacting nucleotides.

Table B.2 PRIDB summary statistics

Type	Total (Interface + Non-Interface	Number in Interfaces (%)
Amino Acids	1,475,774	414,026 (38)
Ribonucleotides	851,853	326,441 (28)

User Interface

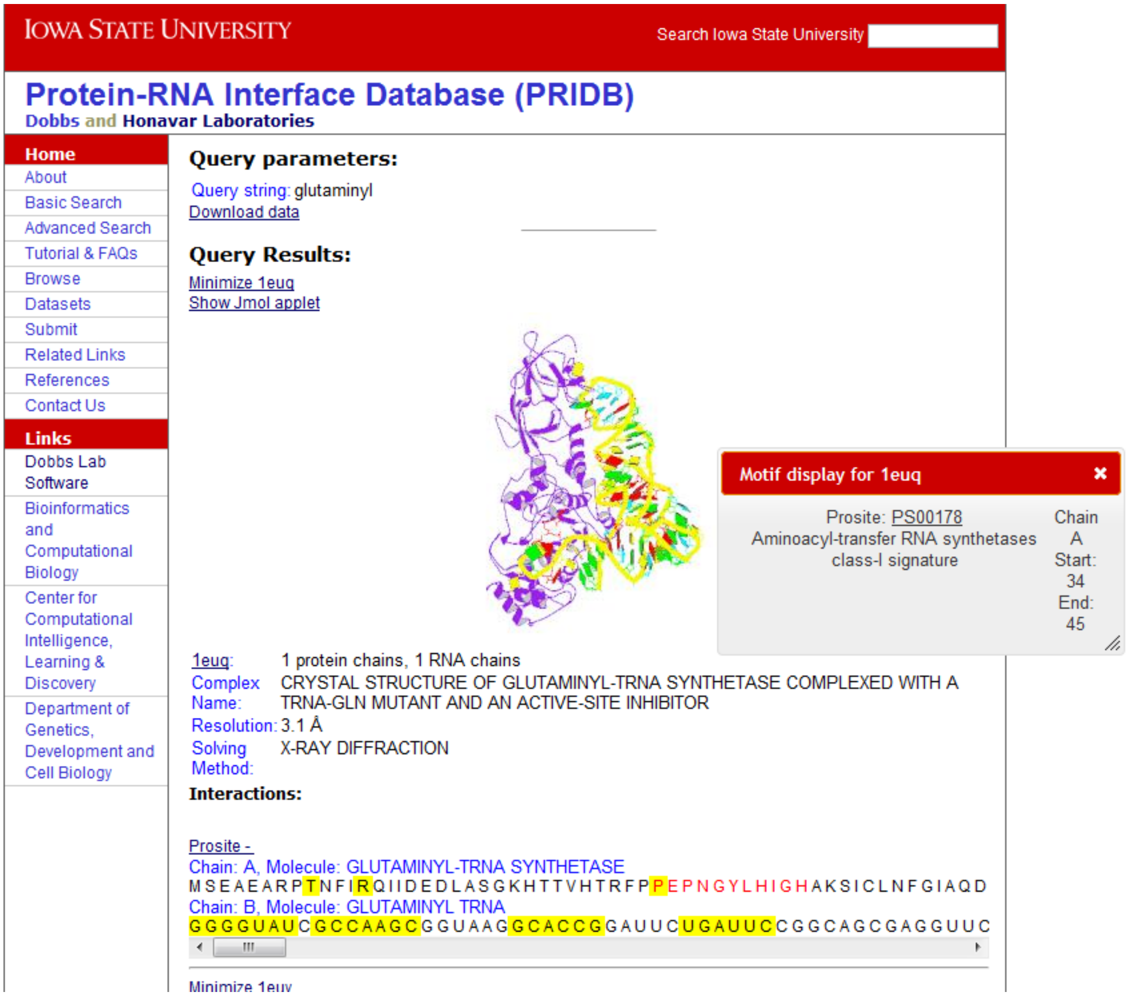
PRIDB provides a 'Tutorial and FAQs' section with detailed instructions on using PRIDB's web interface; a list and brief descriptions of key capabilities of PRIDB are provided here. Using the 'Basic Search' function, users can retrieve information about protein-RNA complexes using their PDB ID or a keyword. Using the 'Advanced Search' function, users can filter results by specifying:

- the experimental method used to determine the complex structure (e.g., x-ray diffraction, nuclear magnetic resonance);
- a resolution range or threshold (for structures determined using x-ray diffraction, electron microscopy, or fiber diffraction);
- the minimum or maximum length of protein or RNA chains within the complex;
- an amino acid or nucleotide subsequence found within the sequence of at least one of the protein or RNA chains in the complex; and
- a motif (as defined by ProSite for protein chains or FR3D for RNA chains) found within at least one chain in the complex.

The 'Advanced Search' function also allows users to either specify a different distance cutoff for the distance-based interaction definition or choose the alternative rule-based definition.

As shown in Figure B.1, when viewing search results, PRIDB provides:

- a summary of and basic information (name, resolution, and structure determination method) about each complex, as well as a link to that complex's PDB entry;
- a linear display of the amino acid and nucleotide residues in each chain of each complex, with residues in the protein-RNA interface highlighted;


IOWA STATE UNIVERSITY
Search Iowa State University
Protein-RNA Interface Database (PRIDB)
Dobbs and Honavar Laboratories
Home
About
Basic Search
Advanced Search
Tutorial & FAQs
Browse
Datasets
Submit
Related Links
References
Contact Us
Links
Dobbs Lab Software
Bioinformatics and Computational Biology
Center for Computational Intelligence, Learning & Discovery
Department of Genetics, Development and Cell Biology
Download data
Minimize 1euq
Show Jmol applet
Motif display for 1euq
✖
Prosite -
Minimize 1euv

Query parameters:

Query string: glutaminyl

Query Results:

Prosite: PS00178 Aminoacyl-transfer RNA synthetases class-I signature

Chain A Start: 34 End: 45

1euq: 1 protein chains, 1 RNA chains
Complex Name: CRYSTAL STRUCTURE OF GLUTAMINYL-TRNA SYNTHETASE COMPLEXED WITH A TRNA-GLN MUTANT AND AN ACTIVE-SITE INHIBITOR
Resolution: 3.1 Å
Solving Method: X-RAY DIFFRACTION

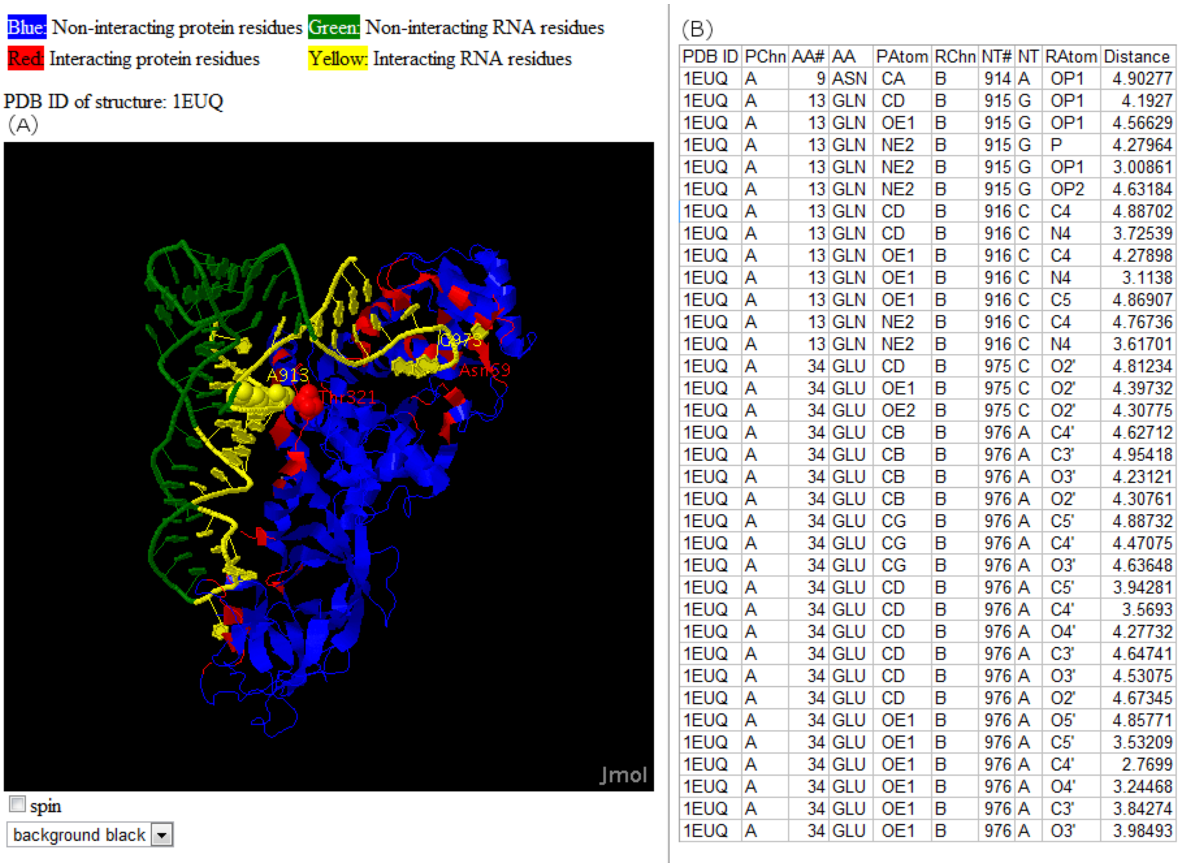
Interactions:

Chain: A, Molecule: GLUTAMINYL-TRNA SYNTHETASE
MSEAEARPTNFIRQIIDEDLASGKHTTVHTRFPPEPNGYLHIGHAKSICLNFGIAQD
Chain: B, Molecule: GLUTAMINYL TRNA
GGGGUAUCGCCAAGCGGUAAGGCACCGGAUUCUGAUUCCGGCAGCGAGGUUC

Figure B.1 Sample PRIDB output. Amino acid residues and ribonucleotides highlighted in yellow are located in the protein-RNA interface; residues in red font are part of a ProSite or FR3D motif.

- a display of residues (in red font) that are part of a protein or RNA motif, with information about that motif (and a link back to its source) provided on mouse-over;
- a JMol applet for 3-dimensional visualization of each complex, with interacting amino acid and nucleotide residues colored (Figure B.2A) ; and
- a link to a dynamically-generated file containing atomic-level interface information for each result in a machine readable format (Figure B.2B).

In addition to providing machine-readable results files for all searches, pre-computed results files for the non-redundant RB109, RB147, and RB199 datasets described above have been made



Blue: Non-interacting protein residues
Green: Non-interacting RNA residues
Red: Interacting protein residues
Yellow: Interacting RNA residues
(A)
A913
Thr321
G973
Asn69
Jmol
□ spin
background black
(B)

PDB ID of structure: 1EUQ

PDB ID	PChn	AA#	AA	PAtom	RChn	NT#	NT	RAtom	Distance
1EUQ	A	9	ASN	CA	B	914	A	OP1	4.90277
1EUQ	A	13	GLN	CD	B	915	G	OP1	4.1927
1EUQ	A	13	GLN	OE1	B	915	G	OP1	4.56629
1EUQ	A	13	GLN	NE2	B	915	G	P	4.27964
1EUQ	A	13	GLN	NE2	B	915	G	OP1	3.00861
1EUQ	A	13	GLN	NE2	B	915	G	OP2	4.63184
1EUQ	A	13	GLN	CD	B	916	C	C4	4.88702
1EUQ	A	13	GLN	CD	B	916	C	N4	3.72539
1EUQ	A	13	GLN	OE1	B	916	C	C4	4.27898
1EUQ	A	13	GLN	OE1	B	916	C	N4	3.1138
1EUQ	A	13	GLN	OE1	B	916	C	C5	4.86907
1EUQ	A	13	GLN	NE2	B	916	C	C4	4.76736
1EUQ	A	13	GLN	NE2	B	916	C	N4	3.61701
1EUQ	A	34	GLU	CD	B	975	C	O2'	4.81234
1EUQ	A	34	GLU	OE1	B	975	C	O2'	4.39732
1EUQ	A	34	GLU	OE2	B	975	C	O2'	4.30775
1EUQ	A	34	GLU	CB	B	976	A	C4'	4.62712
1EUQ	A	34	GLU	CB	B	976	A	C3'	4.95418
1EUQ	A	34	GLU	CB	B	976	A	O3'	4.23121
1EUQ	A	34	GLU	CB	B	976	A	O2'	4.30761
1EUQ	A	34	GLU	CG	B	976	A	C5'	4.88732
1EUQ	A	34	GLU	CG	B	976	A	C4'	4.47075
1EUQ	A	34	GLU	CG	B	976	A	O3'	4.63648
1EUQ	A	34	GLU	CD	B	976	A	C5'	3.94281
1EUQ	A	34	GLU	CD	B	976	A	C4'	3.5693
1EUQ	A	34	GLU	CD	B	976	A	O4'	4.27732
1EUQ	A	34	GLU	CD	B	976	A	C3'	4.64741
1EUQ	A	34	GLU	CD	B	976	A	O3'	4.53075
1EUQ	A	34	GLU	CD	B	976	A	O2'	4.67345
1EUQ	A	34	GLU	OE1	B	976	A	O5'	4.85771
1EUQ	A	34	GLU	OE1	B	976	A	C5'	3.53209
1EUQ	A	34	GLU	OE1	B	976	A	C4'	2.7699
1EUQ	A	34	GLU	OE1	B	976	A	O4'	3.24468
1EUQ	A	34	GLU	OE1	B	976	A	C3'	3.84274
1EUQ	A	34	GLU	OE1	B	976	A	O3'	3.98493

Figure B.2 (A) PRIDB provides a JMol applet for visualizing and manipulating interfaces within 3-D structures. (B) PRIDB output can be downloaded as a CSV file.

available. These files, along with the complete PRIDB database (rRB926), can be downloaded from the 'Datasets' section of the website. Users can also generate a machine-readable list of interface residues for any arbitrary collection of complexes by inputting a list of PDB IDs. Results files contain a single line for each pair of interacting atoms listing the specific interacting atoms (by chain name, residue number, and atom name) and the distance between them.

Users may also calculate interface residues for protein-RNA complexes that are not in PDB using PRIDB by submitting a structure file in PDB format. A results file containing interface residues (as calculated using PRIDB's 5Å cutoff) is returned via e-mail.

Conclusions and Future Directions

PRIDB provides researchers with atomic and residue-level information about structures of protein-RNA complexes and their interfaces, facilitating analyses of protein-RNA interactions by pre-computing commonly used information and by providing structural information both interactively onscreen and in a machine-readable format. It allows users to rapidly identify and visualize interfaces in protein-RNA complexes on a residue-by-residue basis and displays identified ProSite or FR3D motifs along with the amino acid or ribonucleotide sequences. PRIDB can be used to generate custom datasets of protein-RNA interfaces for statistical analyses and machine learning applications. The PRIDB server also provides pre-calculated benchmark datasets of protein-RNA complexes for evaluating the performance of interface prediction methods. PRIDB will be updated regularly as new structures are released through PDB, and is intended to be a stable resource for researchers in the field of protein-RNA interactions.

Future versions of PRIDB will include additional protein and RNA motifs from other sources, such as PRINTS [Attwood et al. (2003)], PIRSF [Wu et al. (2004)] and other InterPro [Hunter et al. (2009)] member databases. In addition, the current JMol 3-D visualization capabilities will be extended to user-submitted structures, allowing for more facile manipulation and examination of interfaces in complexes not currently in the PDB.

Acknowledgements

The authors thank members of our research groups for helpful discussions and especially Usha Muppirala for critical comments on the PRIDB server and manuscript.

Funding

This work was supported by the National Institutes of Health [GM066387 to V.H. and D.D.], the National Science Foundation [IGERT0504304 to D.D.; GK120947929 to B.L.; NIBIB-NSF0608769 to V.H., J.F., C.Z.], and by Iowa State University's Center for Integrated Animal Genomics [B.L., D.D.], and the Center for Computational Intelligence, Learning and Discovery [V.H.]

APPENDIX C. AN ANALYSIS OF CONFORMATIONAL CHANGES UPON RNA-PROTEIN BINDING

Poster paper that appeared in ACM-BCB 2014

Kannan Sankar*, Rasna R. Walia*, Carla M. Mann, Robert Jernigan, Vasant Honavar and Drena Dobbs

*Contributed equally to the manuscript

Abstract

RNA-binding proteins (RBPs) have myriad functions in transcription, translation, and post-transcriptional gene regulation, with central roles in normal development as well as in both genetic and infectious diseases. When a protein binds RNA, a conformational change often occurs. For RNA-protein complexes that have been characterized, conformational changes have been observed in the protein, the RNA, or both. These conformational changes have not been sufficiently characterized, however, in part due to the small number of structures of bound and unbound complexes of RNA-binding proteins available until recently. Here, we systematically analyze a new dataset of 90 pairs of bound and unbound proteins to evaluate the conformational changes that occur upon RNA-binding. Most of the conformational changes were observed in non-interfacial regions of the RNA-binding proteins. Detailed analyses of the modes of RNA-binding and any associated conformational changes in proteins are critical for fully understanding the mechanisms of RNA-protien recognition, for developing better RNA-protein docking methods and methods for predicting interfacial residues, and for RNA-based drug design.

Motivation

In experiments comparing the performance of sequence-based versus structure-based machine learning methods for predicting RNA-binding residues in RBPs, we have found that sequence-based methods outperform structure-based methods. These results are interesting in light of the fact that conformational changes are problematic for rigid docking and other structure-based methods for computational prediction of RNA-binding residues. Is this because many of the structures change significantly upon binding?

Ellis and Jones [Ellis and Jones (2008)] performed a quantitative evaluation of conformational changes in 12 RNA-binding proteins (RBPs) by comparing bound and unbound forms. Since 2008, the number of structurally characterized ribonucleoprotein (RNP) complexes has increased dramatically (to almost 1500 in the PDB as of November 2013), and the number of available complexes for which both bound and unbound forms are available has increased almost 8-fold. Three benchmark datasets [Barik et al. (2012); Perez-Cano et al. (2012); Huang and Zou (2013)] for protein-RNA docking have been published recently. Based on these, we generated and analyzed a dataset of 90 pairs of ’unbound’ and ’bound’ RBPs. The primary aim of this study is to characterize and quantify conformational changes in RNA-binding proteins, using the much larger dataset of RNPs now available.

Results

We systematically analyzed the conformational chnages in 90 pairs of bound versus unbound structures with regard to several criteria, including how many flexible and conformationally invariant residues mapped to buried versus exposed regions of the protein (computed using NACCESS [Hubbard and Thornton (1993)]), and how many of each residue type were located in the RNA-protein interface versus non-interfacial regions. We used ESCET [Schneider (2000, 2002, 2004)] to quantify conformational changes (Figure C.1).

Surprisingly, we found that most conformationally flexible residues in RBPs are non-interfacial surface residues. 10% of the residues that occur in both the unbound and bound structures are interfacial residues, and among these, 27% are conformationally flexible. Figure C.2 shows

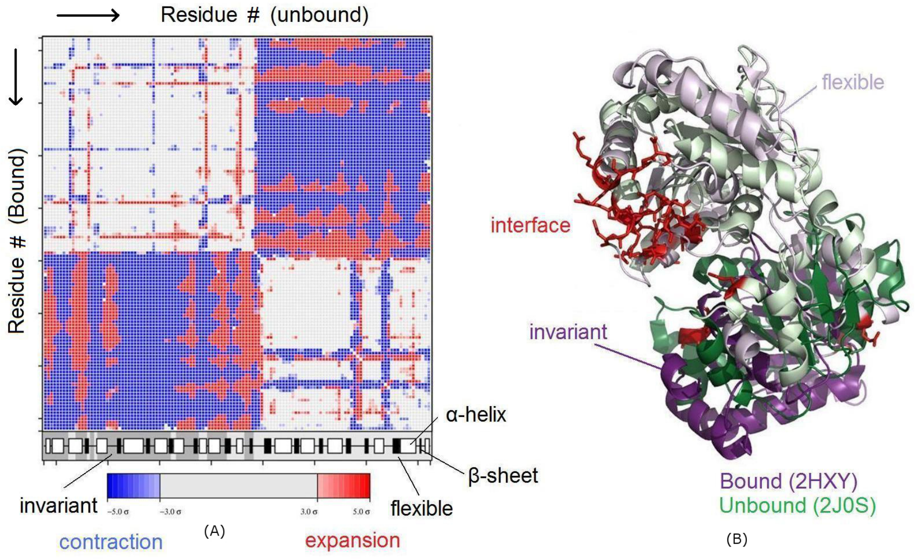


Figure C.1 ESCET analysis of the DDX48 ATP-dependent RNA helicase. (A) An EDD matrix illustrating changes in the internal distances between unbound and bound forms of the RNA helicase. Regions that undergo expansion (red) and contraction (blue) are shown, with darker shades representing larger changes. In the diagram below the matrix, black and white boxes show locations of alpha-helices (white) and beta-strands (black); assignments of conformationally invariant (dark grey) and flexible regions (light grey) are summarized by shading behind the boxes. (B) Superimposition of unbound (purple, 2HXY:A) and bound (green, 2J0S:A) PDB structures for the helicase, with the interfacial residues in the bound form shown in red. For clarity, the bound RNA is not shown.

that out of a total of 8,390 residues in the flexible regions of RBPs, 1,137 (13.6%) are interface residues and 7,253 (86.4%) are non-interface residues. Analysis of the results based on type of RNA bound illustrated that rRNA-binding proteins showed a larger fraction of interface residues in flexible regions.

Conclusions and Future Directions

Based on the 90 pairs of characterized RNA-protein complexes analyzed in this study, most RNA-binding proteins undergo a conformational change upon RNA-binding. Two important conclusions from this study are that RNA-binding residues occur primarily within conformationally invariant regions and that most flexible regions in RNA-binding proteins correspond to non-RNA-binding residues. We hypothesize that the inherently 'flexible' regions of

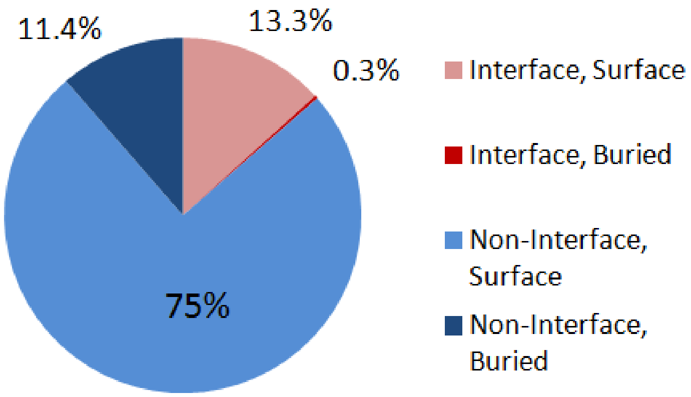


Figure C.2 Pie chart showing characteristics of 8,390 residues found in "flexible" regions (100%). Residue categories are: interface (red) vs non-interface (blue) and surface (lighter shades) vs buried (darker shades).

RNA-binding proteins play an important role in large conformational changes associated with functional motions of RNPs (e.g. in ribosomes, spliceosomes, viral capsids), rather than in RNA-protein recognition alone. Ongoing experiments are designed to directly test this hypothesis.

BIBLIOGRAPHY

Adamczak, R., Porollo, A., and Meller, J. (2005). Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, 59(3):467–475. 35

Aebersold, R. and Mann, M. (2003). Mass spectrometry-based proteomics. *Nature*, 422(6928):198–207. 6

Agostini, F., Zanzoni, A., Klus, P., Marchese, D., Cirillo, D., and Tartaglia, G. G. (2013). catRAPID omics: a web server for large-scale prediction of protein-RNA interactions. *Bioinformatics (Oxford, England)*, 29(22):2928–2930. 107

Ahmad, S. and Sarai, A. (2005). PSSM-based prediction of DNA binding sites in proteins. *BMC Bioinformatics*, 6(1):33. 20, 32

Allers, J. and Shamoo, Y. (2001). Structure-based analysis of protein-RNA interactions using the program {ENTANGLE}. *J Mol Biol*, 311(1):75–86. 20, 132

Altschul, S. F., Madden, T. L., Scheffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research*, 25(17):3389–3402. 46, 75, 80, 82

Anders, G., Mackowiak, S. D., Jens, M., Maaskola, J., Kuntzagk, A., Rajewsky, N., Landthaler, M., and Dieterich, C. (2011). doRiNA: a database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Research*, 40(D1):D180–D186. 6, 90

Andrade, M. A. (1999). Position-specific annotation of protein function based on multiple homologs. In Lengauer, T., Schneider, R., Bork, P., Brutlag, D. L., Glasgow, J. I., Mewes, H.-W., and Zimmer, R., editors, *ISMB*, pages 28–33. AAAI. 55

Anger, A. M., Armache, J.-P., Berninghausen, O., Habeck, M., Subklewe, M., Wilson, D. N., and Beckmann, R. (2013). Structures of the human and drosophila 80s ribosome. *Nature*, 497(7447):80–85. 100

Anko, M.-L. and Neugebauer, K. M. (2012). RNA-protein interactions in vivo: global gets specific. *Trends Biochem Sci*. 89, 106

Ascano, M., Gerstberger, S., and Tuschl, T. (2013). Multi-disciplinary methods to define RNA-protein interactions and regulatory networks. *Current Opinion in Genetics & Development*, 23(1):20–28. 6, 106

Ascano, M., Hafner, M., Cekan, P., Gerstberger, S., and Tuschl, T. (2011). Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdisciplinary Reviews: RNA*. 2

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A., and Zygouri, C. (2003). PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Research*, 31(1):400–402. 138

Auweter, S. D., Oberstrass, F. C., and Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, 34(17):4943–4959. 2

Bahadur, R., Zacharias, M., and Janin, J. (2008). Dissecting protein-RNA recognition sites. *Nucleic Acids Res*, 36(8):2705–2716. 7, 96

Baker, C. M. and Grant, G. H. (2007). Role of aromatic amino acids in protein-nucleic acid recognition. *Biopolymers*, 85(5-6):456–470. 97

Baldi, P., Brunak, S., Chauvin, Y., Andersen, C., and Nielsen, H. (2000). Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics*, 16(5):412–424. 49, 50, 83, 112

Barik, A., C, N., P, M., and Bahadur, R. P. (2012). A protein-RNA docking benchmark (i): nonredundant cases. *Proteins*, 80(7):1866–1871. 140

Bellucci, M., Agostini, F., Masin, M., and Tartaglia, G. G. (2011). Predicting protein associations with long noncoding rnas. *Nat Meth*, 8(6):444–445. 9, 12, 107

Berman, H., Battistuz, T., Bhat, T., Bluhm, W., Bourne, P., Burkhardt, K., Feng, Z., Gilliland, G., Iype, L., and Jain, S. (2002). The protein data bank. *Acta Crystallogr D Biol Crystallogr*, 58(Pt6No1):899–907. 16, 78, 79

Berman, H., Olson, W., Beveridge, D., Westbrook, J., Gelbin, A., Demeny, T., Hsieh, S., Srinivasan, A., and Schneider, B. (1992). The nucleic acid database. a comprehensive relational database of three-dimensional structures of nucleic acids. *Biophys J*, 63(3):751–759. 132

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, 28:235—242. 6, 57, 78, 79, 131

Bharat, T. A. M., Noda, T., Riches, J. D., Kraehling, V., Kolesnikova, L., Becker, S., Kawaoka, Y., and Briggs, J. A. G. (2012). Structural dissection of Ebola virus and its assembly determinants using cryo-electron tomography. *Proceedings of the National Academy of Sciences*, 109(11):4275–4280. 99

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigo, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., Kuehn, M. S., Taylor, C. M., Neph, S., Koch, C. M., Asthana, S., Malhotra, A., Adzhubei, I., Greenbaum, J. A., Andrews, R. M., Flicek, P., Boyle, P. J., Cao, H., Carter, N. P., Clelland, G. K., Davis, S., Day, N., Dhami, P., Dillon, S. C., Dorschner, M. O., Fiegler, H., Giresi, P. G., Goldy, J., Hawrylycz, M., Haydock, A., Humbert, R., James, K. D., Johnson, B. E., Johnson, E. M., Frum, T. T., Rosenzweig, E. R., Karnani, N., Lee, K., Lefebvre, G. C., Navas, P. A., Neri, F., Parker, S. C. J., Sabo, P. J., Sandstrom, R., Shafer, A., Vetrie, D., Weaver, M., Wilcox, S., Yu1, M., Collins, F. S., Dekker, J., Lieb, J. D., Tullius, T. D., Crawford, G. E., Sunyaev, S., Noble, W. S., Dunham, I., Denoeud, F., Reymond, A., Kapranov, P., Rozowsky, J., Zheng, D., Castelo, R., Frankish, A., Harrow, J., Ghosh, S., Sandelin, A., Hofacker, I. L., Baertsch, R., Keefe, D., Dike, S., Cheng, J., Hirsch, H. A., Sekinger, E. A., Lagarde, J., Abril, J. F., Shahab, A., Flamm, C., Fried, C., Hackermuller, J., Hertel, J., Lindemeyer, M., Missal, K., Tanzer, A., Washietl, S., Korbel, J., Emanuelsson, O., Pedersen, J. S., Holroyd, N., Taylor, R., Swarbreck, D., Matthews, N., Dickson, M. C., Thomas, D. J., Weirauch, M. T., Gilbert, J., Drenkow, J., Bell, I., Zhao, X., Srinivasan, K. G., Sung, W.-K., Ooi, H. S., Chiu, K. P., Foissac, S., Alioto, T., Brent, M., Pachter, L., Tress, M. L., Valencia, A., Choo, S. W., Choo, C. Y., Ucla, C., Manzano, C., Wyss, C., Cheung, E., Clark, T. G., Brown, J. B., Ganesh, M., Patel, S., Tammana, H., Chrast, J., Henrichsen, C. N., Kai, C., Kawai, J., Nagalakshmi, U., Wu, J., Lian, Z., Lian, J., Newburger, P., Zhang, X., Bickel, P., Mattick, J. S., Carninci, P., Hayashizaki, Y., Weissman, S., Hubbard, T., Myers, R. M., Rogers, J., Stadler, P. F., Lowe, T. M., Wei, C.-L., Ruan, Y., Struhl, K., Gerstein, M., Antonarakis, S. E., Fu, Y., Green, E. D., KaraÃűz, U., Siepel, A., Taylor, J., Liefer, L. A., Wetterstrand, K. A., Good, P. J., Feingold, E. A., Guyer, M. S., Cooper, G. M., Asimenos, G., Dewey, C. N., Hou, M., Nikolaev, S., Montoya-Burgos, J. I., LÃűytynoja, A., Whelan, S., Pardi, F., Massingham, T., Huang, H., Zhang, N. R., Holmes, I., Mullikin, J. C., Ureta-Vidal, A., Paten, B., Seringhaus, M., Church, D., Rosenbloom, K., Kent, W. J., Stone, E. A., Batzoglou, S., Goldman, N., Hardison, R. C., Haussler, D., Miller, W., Sidow, A., Trinklein, N. D., Zhang, Z. D., Barrera, L., Stuart, R., King, D. C., Ameur, A., Enroth, S., Bieda, M. C., Kim, J., Bhinge, A. A., Jiang, N., Liu, J., Yao, F., Vega, V. B., Lee, C. W. H., Ng, P., Yang, A., Moqtaderi, Z., Zhu, Z., Xu, X., Squazzo, S., Oberley, M. J., Inman, D., Singer, M. A., Richmond, T. A., Munn, K. J., Rada-Iglesias, A., Wallerman, O., Komorowski, J., Fowler, J. C., Couttet, P., Bruce, A. W., Dovey, O. M., Ellis, P. D., Langford, C. F., Nix, D. A., Euskirchen, G., Hartman, S., Urban, A. E., Kraus, P., Calcar, S. V., Heintzman, N., Kim, T. H., Wang, K., Qu, C., Hon, G., Luna, R., Glass, C. K., Rosenfeld, M. G., Aldred, S. F., Cooper, S. J., Halees, A., Lin, J. M., Shulha, H. P., Zhang, X., Xu, M., Haidar, J. N. S., Yu, Y., Birney*, E., Iyer, V. R., Green, R. D., Wadelius, C., Farnham, P. J., Ren, B., Harte, R. A., Hinrichs, A. S., Trumbower, H., Clawson, H., Hillman-Jackson, J., Zweig, A. S., Smith, K., Thakkapallayil, A., Barber, G., Kuhn, R. M., Karolchik, D., Armengol, L., Bird, C. P., Bakker, P. I. W. d., Kern, A. D., Lopez-Bigas, N., Martin, J. D., Stranger, B. E., Woodroffe, A., Davydov, E., Dimas, A., Eyras, E., HallgrÃŋmsdÃşttir, I. B., Huppert, J., Zody, M. C., Abecasis, G. R., Estivill, X., Bouffard, G. G., Guan, X., Hansen, N. F., Idol, J. R., Maduro, V. V. B., Maskeri, B., McDowell, J. C., Park, M., Thomas, P. J., Young, A. C., Blakesley, R. W., Muzny, D. M., Sodergren, E., Wheeler, D. A., Worley, K. C., Jiang, H., Weinstock, G. M., Gibbs, R. A., Graves, T., Fulton, R., Mardis, E. R., Wilson, R. K., Clamp, M., Cuff, J., Gnerre, S., Jaffe, D. B., Chang, J. L., Lindblad-Toh, K., Lander, E. S., Koriabine, M., Nefedov, M., Osoegawa, K., Yoshinaga, Y., Zhu, B., and Jong, P. J. d. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816. 2

Blencowe, B. J. (2006). Alternative splicing: New insights from global analyses. *Cell*, 126(1):37–47. 54

Bohnsack, M. T., Tollervey, D., and Granneman, S. (2012). Identification of RNA helicase target sites by UV cross-linking and analysis of cDNA. *Methods in Enzymology*, 511:275–288. 5

Bornholdt, Z. A., Noda, T., Abelson, D. M., Halfmann, P., Wood, M., Kawaoka, Y., and Saphire, E. O. (2013). Structural basis for ebolavirus matrix assembly and budding; protein plasticity allows multiple functions. *Cell*, 154(4):763–774. 99

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32. 25, 111

Butter, F., Scheibe, M., Mral, M., and Mann, M. (2009). Unbiased RNA-protein interaction screen by quantitative proteomics. *Proceedings of the National Academy of Sciences*, 106(26):10626–10631. 6

Capra, J. A. and Singh, M. (2007). Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23(15):1875–1882. 32

Caragea, C., Sinapov, J., Honavar, V., and Dobbs, D. (2007a). Assessing the performance of macromolecular sequence classifiers. *Bioinformatics and Bioengineering, 2007*, pages 320–326. 19, 50, 82

Caragea, C., Sinapov, J., Silvescu, A., Dobbs, D., and Honavar, V. (2007b). Glycosylation site prediction using ensembles of support vector machine classifiers. *BMC Bioinformatics*, 8(1):438. 108

Carson, M. B., Langlois, R., and Lu, H. (2010). Naps: a residue-level nucleic acid-binding prediction server. *Nucleic Acids Research*, 38(suppl 2):W431–W435. 22, 55, 65, 71, 72, 76

Charon, C., Moreno, A. B., Bardou, F., and Crespi, M. (2010). Non-protein-coding RNAs and their interacting RNA-binding proteins in the plant cell nucleus. *Molecular plant*, 3(4):729–739. 122

Cheetham, S. W., Gruhl, F., Mattick, J. S., and Dinger, M. E. (2013). Long noncoding RNAs and the genetics of cancer. *British Journal of Cancer*, 108(12):2419–2425. 54

Chen, W., Zhang, S.-W., Cheng, Y.-M., and Pan, Q. (2011). Identification of protein-RNA interaction sites using the information of spatial adjacent residues. *Proteome Science*, 9(Suppl 1):S16. 17, 24

Chen, Y. and Lim, C. (2008). Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. *Nucleic Acids Res*, 36(5):e29. 17, 23, 26

Chen, Y. and Varani, G. (2013). Engineering RNA-binding proteins for biology. *The FEBS journal*, 280(16):3734–3754. 89, 101

Cheng, C., Su, E., Hwang, J., Sung, T., and Hsu, W. (2008). Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. *BMC Bioinformatics*, 9(Suppl 12):S6. 17, 21, 46, 47, 55, 62

Choi, S. and Han, K. (2011). Prediction of rna-binding amino acids from protein and rna sequences. *BMC Bioinformatics*, 12(Suppl 13):S7. 126

Choi, S. and Han, K. (2013). Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Computers in Biology and Medicine*, 43(11):1687–1697. 9

Chojnowski, G., Walen, T., and Bujnicki, J. M. (2014). RNA bricks–a database of RNA 3d motifs and their interactions. *Nucleic Acids Research*, 42(Database issue):D123–131. 11

Cirillo, D., Livi, C. M., Agostini, F., and Tartaglia, G. G. (2014). Discovery of protein-RNA networks. *Molecular BioSystems*, 10(7):1632–1642. 11, 106, 107

Clark, M. B., Amaral, P. P., Schlesinger, F. J., Dinger, M. E., Taft, R. J., Rinn, J. L., Ponting, C. P., Stadler, P. F., Morris, K. V., Morillon, A., Rozowsky, J. S., Gerstein, M. B., Wahlestedt, C., Hayashizaki, Y., Carninci, P., Gingeras, T. R., and Mattick, J. S. (2011). The reality of pervasive transcription. *PLoS Biol*, 9(7):e1000625. 122

Clery, A., Blatter, M., and Allain, F. H.-T. (2008). RNA recognition motifs: boring? not quite. *Current Opinion in Structural Biology*, 18(3):290–298. 89, 101

Coimbatore Narayanan, B., Westbrook, J., Ghosh, S., Petrov, A. I., Sweeney, B., Zirbel, C. L., Leontis, N. B., and Berman, H. M. (2014). The nucleic acid database: new features and capabilities. *Nucleic Acids Research*, 42(Database issue):D114–122. 6, 104

Conrad, N. K. (2008). Chapter 15. co-immunoprecipitation techniques for assessing RNA-protein interactions in vivo. *Methods in Enzymology*, 449:317–342. 6

Consortium, T. E. P. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414):57–74. 2

Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., and Hughes, T. R. (2011). RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research*, 39(Database issue):D301–D308. 6, 104

Darnell, R. B. (2010). Hits-clip: panoramic views of proteinâĂŞrna regulation in living cells. *Wiley Interdisciplinary Reviews - RNA*, 1(2):266–286. 2, 5

Das, P., Basu, A., Biswas, A., Poddar, D., Andrews, J., Barik, S., Komar, A. A., and Mazumder, B. (2013). Insights into the mechanism of ribosomal incorporation of mammalian l13a protein during ribosome biogenesis. *Molecular and Cellular Biology*, 33(15):2829–2842. 9

Daubner, G. M., ClÃĺry, A., and Allain, F. H.-T. (2013). RRM-RNA recognition: NMR or crystallographyâĂęand new findings. *Current Opinion in Structural Biology*, 23(1):100–108. 89

Davis, J. and Goadrich, M. (2006). The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd international conference on Machine learning*, ICML '06, pages 233–240, New York, NY, USA. ACM. 50

de Castro, E., Sigrist, C. J. A., Gattiker, A., Bulliard, V., Langendijk-Genevaux, P. S., Gasteiger, E., Bairoch, A., and Hulo, N. (2006). ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. *Nucleic Acids Research*, 34(suppl 2):W362–W365. 133

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, 7:1–30. 28, 51

Denison, M. R. (2008). Seeking membranes: Positive-Strand RNA virus replication complexes. *PLoS Biology*, 6(10):e270. 54

Dominguez, C., Schubert, M., Duss, O., Ravindranathan, S., and Allain, F. H.-T. (2011). Structure determination and dynamics of protein-RNA complexes by NMR spectroscopy. *Progress in Nuclear Magnetic Resonance Spectroscopy*, 58(1-2):1–61. 4

EL-Manzalawy, Y. and Honavar, V. (2010). Recent advances in b-cell epitope prediction methods. *Immunome Research*, 6(Suppl 2):S2. 108

Ellis, J., Broom, M., and Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins*, 66(4):903–911. 7, 17

Ellis, J. J. and Jones, S. (2008). Evaluating conformational changes in protein structures binding RNA. *Proteins*, 70(4):1518–1526. 140

Esteller, M. (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics*, 12(12):861–874. 54, 122

Fabian, M. R., Sonenberg, N., and Filipowicz, W. (2010). Regulation of mRNA Translation and Stability by microRNAs. volume 79 of *Annual Review of Biochemistry*, pages 351–379. 16, 131

Faoro, C. and Ataide, S. F. (2014). Ribonomic approaches to study the RNA-binding proteome. *FEBS letters*, 588(20):3649–3664. 2, 11

Fatica, A. and Bozzoni, I. (2014). Long non-coding RNAs: new players in cell differentiation and development. *Nature Reviews Genetics*, 15(1):7–21. 2, 122

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., and Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230. 10

Flintoft, L. (2010). Transcriptomics: Throwing light on dark matter. *Nature Reviews Genetics*, 11(7):455–455. 122

Fodor, A. A. and Aldrich, R. W. (2004). Influence of conservation on calculations of amino acid covariance in multiple sequence alignments. *Proteins: Structure, Function, and Bioinformatics*, 56(2):211–221. 32

Font, J. and Mackay, J. P. (2010). Beyond DNA: zinc finger domains as RNA-binding modules. *Methods in Molecular Biology (Clifton, N.J.)*, 649:479–491. 89

Friedberg, I. and Margalit, H. (2002). Persistently conserved positions in structurally similar, sequence dissimilar proteins: Roles in preserving protein fold and function. *Protein Science*, 11(2):350–360. 32

Friedman, M. (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11:86–92. 51

Fritsch, V. and Westhof, E. (2010). The architectural motifs of folded RNAs. In Gunteryer, editor, *The Chemical Biology of Nucleic Acids*, pages 141–174. John Wiley & Sons, Ltd. 88

Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152. 79

Fu, X.-D. and Ares Jr, M. (2014). Context-dependent control of alternative splicing by rna-binding proteins. *Nat Rev Genet*, 15(10):689–701. 1

Gagnon, K. T. and Maxwell, E. S. (2011). Electrophoretic mobility shift assay for characterizing RNA-protein interaction. *Methods in Molecular Biology (Clifton, N.J.)*, 703:275–291. 6

Galicia-Vazquez, G., Lindqvist, L., Wang, X., Harvey, I., Liu, J., and Pelletier, J. (2009). High-throughput assays probing protein-RNA interactions of eukaryotic translation initiation factors. *Analytical Biochemistry*, 384(1):180–188. 54

Garg, A., Kaur, H., and Raghava, G. P. S. (2005). Real value prediction of solvent accessibility in proteins using multiple sequence alignment and secondary structure. *Proteins*, 61(2):318–324. 20

Geisler, S. and Coller, J. (2013). RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts. *Nature Reviews Molecular Cell Biology*, 14(11):699–712. 2, 89, 122

Glisovic, T., Bachorik, J. L., Yong, J., and Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS letters*, 582(14):1977–1986. 2, 5

Gomis-Ruth, F. X., Dessen, A., Timmins, J., Bracher, A., Kolesnikowa, L., Becker, S., Klenk, H. D., and Weissenhorn, W. (2003). The matrix protein VP40 from ebola virus octamerizes into pore-like structures with specific RNA binding properties. *Structure (London, England: 1993)*, 11(4):423–433. 99

Granneman, S., Kudla, G., Petfalski, E., and Tollervey, D. (2009). Identification of protein binding sites on u3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs. *Proceedings of the National Academy of Sciences*, 106(24):9613–9618. 11

Grigull, J., Mnaimneh, S., Pootoolal, J., Robinson, M. D., and Hughes, T. R. (2004). Genome-Wide analysis of mRNA stability using transcription inhibitors and microarrays reveals posttranscriptional control of ribosome biogenesis factors. *Molecular and Cellular Biology*, 24(12):5534–5547. 54

Gupta, A. (2011). *RNA-protein interactions: Analysis of binding interfaces and prediction of protein binding sites in RNA*. PhD thesis, Purdue University. 9, 97, 126

Gupta, A. and Gribskov, M. (2011). The Role of RNA Sequence and Structure in RNA-Protein Interactions. *Journal of Molecular Biology*, 409(4):574 – 587. 7, 8, 96, 97

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, M., Jungkamp, A., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010a). PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *Journal of Visualized Experiments: JoVE*, (41). 5, 11, 90

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., Ascano, Jr, M., Jungkamp, A.-C., Munschauer, M., Ulrich, A., Wardle, G. S., Dewell, S., Zavolan, M., and Tuschl, T. (2010b). Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141. 90

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., and Witten, I. H. (2009). The WEKA data mining software: An update. *SIGKDD Explor. Newsl.*, 11(1):10–18. 111

Haralick, R. M. and Shapiro, L. G. (1991). *Computer and Robot Vision, Vol. 1*. Addison-Wesley, Reading, Mass. 91, 93, 104

Hellman, L. M. and Fried, M. G. (2007). Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat. Protocols*, 2(8):1849–1861. 16, 54

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., and Brown, P. O. (2008). Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLoS Biol*, 6(10):e255. 10, 16, 89, 131

Hook, B., Bernstein, D., Zhang, B., and Wickens, M. (2005). RNA-protein interactions in the yeast three-hybrid system: affinity, sensitivity, and enhanced library screening. *RNA (New York, N.Y.)*, 11(2):227–233. 6

Huang, L. and Lilley, D. M. J. (2014). Structure of a rare non-standard sequence k-turn bound by l7ae protein. *Nucleic Acids Research*, 42(7):4734–4740. 103

Huang, S.-Y. and Zou, X. (2013). A nonredundant structure dataset for benchmarking protein-RNA computational docking. *Journal of Computational Chemistry*, 34(4):311–318. 140

Huang, Y., Liu, S., Guo, D., Li, L., and Xiao, Y. (2013). A novel protocol for three-dimensional structure prediction of RNA-protein complexes. *Scientific reports*, 3:1887. 8

Huang, Y.-F., Chiu, L.-Y., Huang, C.-C., and Huang, C.-K. (2010). Predicting RNA-binding residues from evolutionary information and sequence conservation. *BMC Genomics*, 11(Suppl 4):S2. 17, 21

Huarte, M. and Rinn, J. L. (2010). Large non-coding RNAs: missing links in cancer? *Human Molecular Genetics*, 19(R2):R152–R161. 54

Hubbard, S. J. and Thornton, J. M. (1993). Naccess: A computer program. Department of Biochemistry and Molecular Biology, University College London. 40, 140

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J. A., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009). InterPro: the integrative protein signature database. *Nucleic Acids Research*, 37(suppl 1):D211–D215. 138

Huntzinger, E. and Izaurralde, E. (2011). Gene silencing by microRNAs: contributions of translational repression and mRNA decay. *Nat Rev Genet*, 12(2):99–110. 16

Hutvagner, G. and Simard, M. J. (2008). Argonaute proteins: key players in rna silencing. *Nat Rev Mol Cell Biol*, 9(1):22–32. 1

Imig, J., Kanitz, A., and Gerber, A. (2012). RNA regulons and the RNA-protein interaction network. *BioMol. Concepts*, volume. The final publication is available at www.degruyter.com. 2

Iwakiri, J., Tateishi, H., Chakraborty, A., Patil, P., and Kenmochi, N. (2012). Dissecting the protein-RNA interface: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. *Nucleic Acids Research*, 40(8):3299–3306. 7, 97

Iwasaki, A. (2012). A virological view of innate immune recognition. *Annual Review of Microbiology*, 66(1):177–196. PMID: 22994491. 68

Jeck, W. R. and Sharpless, N. E. (2014). Detecting and characterizing circular RNAs. *Nature Biotechnology*, 32(5):453–461. 122

Jedamzik, B. and Eckmann, C. R. (2009). Analysis of RNA-protein complexes by RNA coimmunoprecipitation and RT-PCR analysis from caenorhabditis elegans. *Cold Spring Harbor Protocols*, 2009(10):pdb.prot5300. 6

Jeong, E., Chung, I., and Miyano, S. (2004a). A neural network method for identification of RNA-interacting residues in protein. *Genome Inform*, 15(1):105–116. 17, 20

Jeong, E., Chung, I., and Miyano, S. (2004b). A neural network method for identification of RNA-interacting residues in protein. *Genome Informatics. International Conference on Genome Informatics*, 15(1):105–116. 17, 55

Jeong, E. and Miyano, S. (2006a). A weighted profile based method for protein-RNA interacting residue prediction. *Trans on Comput Syst Biol IV*, 3939:123–139. 17

Jeong, E. and Miyano, S. (2006b). A weighted profile based method for Protein-RNA interacting residue prediction. In Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., Priami, C., Cardelli, L., and Emmott, S., editors, *Transactions on Computational Systems Biology IV*, volume 3939, pages 123–139. Springer Berlin Heidelberg, Berlin, Heidelberg. 17, 20, 32, 55

Jeong, J. C., Lin, X., and Chen, X. (2011). On Position-Specific scoring matrix for protein function prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 8(2):308–315. 20

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202. 20, 32, 34

Jones, D. T. and Ward, J. J. (2003). Prediction of disordered regions in proteins from position specific score matrices. *Proteins: Structure, Function, and Bioinformatics*, 53(S6):573–578. 20

Jones, S., Daley, D., Luscombe, N., Berman, H., and Thornton, J. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Res*, 29(4):943–954. 7, 8, 17, 40, 96

Jordan, R. A., EL-Manzalawy, Y., Dobbs, D., and Honavar, V. (2012). Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinformatics*, 13(1):41. 56

Kakuta, M., Nakamura, S., and Shimizu, K. (2008). Prediction of protein-protein interaction sites using only sequence information and using both sequence and structural information. *IPSJ Digital Courier*, 4:217–227. 32

Kapranov, P. and St. Laurent, G. (2012). Dark matter RNA: Existence, function, and controversy. *Frontiers in Genetics*, 3. 122

Kauffman, C. and Karypis, G. (2009). LIBRUS: combined machine learning and homology information for sequence-based ligand-binding residue prediction. *Bioinformatics*, 25(23):3099–3107. 56

Kazan, H., Ray, D., Chan, E. T., Hughes, T. R., and Morris, Q. (2010). RNAcontext: A new method for learning the sequence and structure binding preferences of RNA-binding proteins. *PLoS Comput Biol*, 6(7):e1000832. 11, 90

Ke, A. and Doudna, J. A. (2004). Crystallization of RNA and RNA-protein complexes. *Methods*, 34(3):408–414. 4, 54

Kechavarzi, B. and Janga, S. (2014). Dissecting the expression landscape of RNA-binding proteins in human cancers. *Genome Biology*, 15(1):R14. 54

Keene, J. D. (2010). Minireview: Global regulation and dynamics of ribonucleic acid. *Endocrinology*, 151(4):1391–1397. PMID: 20332203. 1

Keene, J. D., Komisarow, J. M., and Friedersdorf, M. B. (2006). Rip-chip: the isolation and identification of mrnas, micrornas and protein components of ribonucleoprotein complexes from cell extracts. *Nat. Protocols*, 1(1):302–307. 2, 5, 90

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murthy, K. R. K. (2001). Improvements to Platt's SMO Algorithm for SVM Classifier Design. *Neural Comput.*, 13:637–649. 48

Khalil, A. M. and Rinn, J. L. (2011). RNA-protein interactions in human health and disease. *Seminars in Cell and Developmental Biology*, 22(4):359–365. 54, 89, 122

Khorshid, M., Rodak, C., and Zavolan, M. (2010). CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins. *Nucleic Acids Research*, 39(Database):D245–D252. 6, 90

Kim, H., Jeong, E., Lee, S., and Han, K. (2003). Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. *FEBS Lett*, 552(2-3):231–239. 17

Kim, O., Yura, K., and Go, N. (2006). Amino acid residue doublet propensity in the protein-RNA interface and its application to RNA interface prediction. *Nucleic Acids Res*, 34:6450–6460. 17, 22, 26, 32, 38, 44, 55, 65, 69, 71, 73, 74, 76, 77

Kin, T., Yamada, K., Terai, G., Okida, H., Yoshinari, Y., Ono, Y., Kojima, A., Kimura, Y., Komori, T., and Asai, K. (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research*, 35(Database issue):D145–148. 109, 125

Kishore, S., Luber, S., and Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in functional genomics*, 9(5-6):391–404. 89

Klass, D. M., Scheibe, M., Butter, F., Hogan, G. J., Mann, M., and Brown, P. O. (2013). Quantitative proteomic analysis reveals concurrent RNA-protein interactions and identifies new RNA-binding proteins in saccharomyces cerevisiae. *Genome research*, 23(6):1028–1038. 6

Kloczkowski, A., Ting, K., Jernigan, R. L., and Garnier, J. (2002). Combining the GOR v algorithm with evolutionary information for protein secondary structure prediction from amino acid sequence. *Proteins*, 49(2):154–166. 34

Kloetgen, A., MÃijnch, P. C., Borkhardt, A., Hoell, J. I., and McHardy, A. C. (2014). Biochemical and bioinformatic methods for elucidating the role of RNAâĂŞprotein interactions in posttranscriptional regulation. *Briefings in Functional Genomics*, page elu020. 122

Konc, J. and Janezic, D. (2010). ProBiS algorithm for detection of structurally similar protein binding sites by local structural alignment. *Bioinformatics*, 26(9):1160–1168. 56

Konig, J., Zarnack, K., Luscombe, N. M., and Ule, J. (2012). Protein-RNA interactions: new genomic technologies and perspectives. *Nat Rev Genet*, 13(2):77–83. 11, 106

Konig, J., Zarnack, K., Rot, G., Curk, T., Kayikci, M., Zupan, B., Turner, D. J., Luscombe, N. M., and Ule, J. (2010). iclip reveals the function of hnRNP particles in splicing at individual nucleotide resolution. *Nat Struct Mol Biol*, 17(7):909–915. 5, 11

Kramer, K., Sachsenberg, T., Beckmann, B. M., Qamar, S., Boon, K.-L., Hentze, M. W., Kohlbacher, O., and Urlaub, H. (2014). Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins. *Nature Methods*, 11(10):1064–1070. 6

Kuersten, S., Radek, A., Vogel, C., and Penalva, L. O. F. (2013). Translation regulation gets its 'omics' moment. *Wiley interdisciplinary reviews. RNA*, 4(6):617–630. 89

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2008). Prediction of RNA binding sites in a protein using SVM and PSSM profile. *Proteins*, 71(1):189–194. 8, 17, 20, 21, 32, 44, 55, 65, 69, 71, 72, 73, 76, 77, 131

Kumar, M., Gromiha, M. M., and Raghava, G. P. S. (2011). SVM-based prediction of RNA-binding proteins using binding residues and evolutionary information. *Journal of Molecular Recognition*, 24(2):303–313. 109

Kumar, M. D. S., Bava, K. A., Gromiha, M. M., Prabakaran, P., Kitajima, K., Uedaira, H., and Sarai, A. (2006). Protherm and pronit: thermodynamic databases for proteins and protein-nucleic acid interactions. *Nucleic Acids Research*, 34(suppl 1):D204–D206. 132

le Cessie, S. and van Houwelingen, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 41(1):191–201. 82

Lee, S. and Blundell, T. L. (2009). BIPA: a database for proteinŰnucleic acid interaction in 3D structures. *Bioinformatics*, 25(12):1559–1560. 131

Leung, A. K. L., Young, A. G., Bhutkar, A., Zheng, G. X., Bosson, A. D., Nielsen, C. B., and Sharp, P. A. (2011). Genome-wide identification of Ago2 binding sites from mouse embryonic stem cells with and without mature microRNAs. *Nature Structural & Molecular Biology*, 18(2):237–244. 11

Leung, D. W., Basler, C. F., and Amarasinghe, G. K. (2012). Molecular mechanisms of viral inhibitors of RIG-I-like receptors. *Trends in Microbiology*, 20(3):139 – 146. 68

Lewis, B. A., Walia, R. R., Terribilini, M., Ferguson, J., Zheng, C., Honavar, V., and Dobbs, D. (2010). PRIDB: a Protein-RNA interface database. *Nucleic Acids Research*, 39(suppl 1):D277–D282. 45, 60, 61, 75, 78

Li, H. and Li, J. (2005). Discovery of stable and significant binding motif pairs from PDB complexes and protein interaction datasets. *Bioinformatics*, 21(3):314–324. 89, 90

Li, H., Li, J., Tan, S. H., and Ng, S. K. (2004). Discovery of binding motif pairs from protein complex structural data and protein interaction sequence data. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pages 312–323. 90

Li, Q., Cao, Z., and Liu, H. (2010). Improve the prediction of RNA-Binding residues using structural neighbours. *Protein and Peptide Letters*, 17(3):287–296. 17, 24

Li, W. and Godzik, A. (2006). CD-Hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659. 79

Li, X., Kazan, H., Lipshitz, H. D., and Morris, Q. D. (2014). Finding the target sites of RNA-binding proteins. *Wiley Interdisciplinary Reviews: RNA*, 5(1):111–130. 10

Licatalosi, D. D. and Darnell, R. B. (2010). RNA processing and its regulation: global insights into biological networks. *Nat Rev Genet*, 11(1):75–87. 1, 16, 89, 131

Lichtarge, O. and Sowa, M. E. (2002). Evolutionary predictions of binding surfaces and interactions. *Current Opinion in Structural Biology*, 12(1):21 – 27. 34

Liu, G., Mattick, J. S., and Taft, R. J. (2013). A meta-analysis of the genomic and transcriptomic composition of complex life. *cc*, 12(1538-4101):2061–2072. 2

Liu, Z.-P., Wu, L.-Y., Wang, Y., Zhang, X.-S., and Chen, L. (2010). Prediction of protein-RNA binding sites by a random forest method with combined features. *Bioinformatics*, 26(13):1616–1622. 17, 25, 44, 131

Livi, C. M. and Blanzieri, E. (2014). Protein-specific prediction of mRNA binding using RNA sequences, binding motifs and predicted secondary structures. *BMC Bioinformatics*, 15(1):123. 12, 107

Lorkovic, Z. J. (2009). Role of plant RNA-binding proteins in development, stress response and genome organization. *Trends in Plant Science*, 14(4):229 – 236. 131

Lu, Q., Ren, S., Lu, M., Zhang, Y., Zhu, D., Zhang, X., and Li, T. (2013). Computational prediction of associations between long non-coding RNAs and proteins. *BMC Genomics*, 14(1):651. 12, 107, 108, 116, 118, 120, 124

Lukong, K. E., Chang, K.-w., Khandjian, E. W., and Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends in Genetics*, 24(8):416–425. 131

Lunde, B., Moore, C., and Varani, G. (2007). RNA-binding proteins: modular design for efficient function. *Nat Rev Mol Cell Biol*, 8(6):479–490. 2, 10, 131

Ma, X., Guo, J., Wu, J., Liu, H., Yu, J., Xie, J., and Sun, X. (2011). Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. *Proteins: Structure, Function, and Bioinformatics*. 17, 24, 44, 55, 65, 72, 76

Maetschke, S. and Yuan, Z. (2009). Exploiting structural and topological information to improve prediction of RNA-protein binding sites. *BMC Bioinformatics*, 10(1):341. 9, 17, 23, 32, 45, 55, 65, 69, 71, 74, 76, 77, 131

Mansfield, K. D. and Keene, J. D. (2009). The ribonome: a dominant force in co-ordinating gene expression. *Biology of the Cell*, 101(3):169–181. 131

Maris, C., Dominguez, C., and Allain, F. H. (2005). The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression. *FEBS Journal*, 272(9):2118–2131. 89, 101

Martiŋ-Renom, M. A., Stuart, A. C., Fiser, A., Sanchez, R., Melo, F., and Sali, A. (2000). Comparative protein structure modeling of genes and genomes. *Annual Review of Biophysics and Biomolecular Structure*, 29:291–325. 55

Masliah, G., Barraud, P., and Allain, F. H.-T. (2013). RNA recognition by double-stranded RNA binding domains: a matter of shape and sequence. *Cellular and Molecular Life Sciences*, 70(11):1875–1895. 89

Matthews, L. R., Vaglio, P., Reboul, J., Ge, H., Davis, B. P., Garrels, J., Vincent, S., and Vidal, M. (2001). Identification of potential interaction networks using sequence-based searches for conserved protein-protein interactions or "interologs". *Genome Research*, 11(12):2120–2126. 55

McHugh, C. A., Russell, P., and Guttman, M. (2014). Methods for comprehensive experimental identification of RNA-protein interactions. *Genome Biology*, 15(1):203. 5, 90

Mills, N. L., Shelat, A. A., and Guy, R. K. (2007). Assay Optimization and Screening of RNA-Protein Interactions by AlphaScreen. *Journal of Biomolecular Screening*, 12(7):946–955. 16

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill, New York. 18, 48, 50, 112

Mitra, S. A., Mitra, A. P., and Triche, T. J. (2012). A Central Role for Long Non-coding RNA in Cancer. *Frontiers in Genetics*, 3(17). 54

Mittal, N., Roy, N., Babu, M. M., and Janga, S. C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences*, 106(48):20300–20305. 131

Mohammad, M. M., Donti, T. R., Yakisich, J. S., Smith, A. G., and Kapler, G. M. (2007). Tetrahymena ORC contains a ribosomal RNA fragment that participates in rDNA origin recognition. *The EMBO Journal*, 26(24):5048–5060. 131

Morozova, N., Allers, J., Myers, J., and Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, 22(22):2746–2752. 132

Muers, M. (2008). RNA splicing: Counting, coordinating and controlling the alternatives. *Nature Reviews Genetics*, 9(12):894–895. 54

Mukherjee, S. and Zhang, Y. (2011). Protein-protein complex structure predictions by multimeric threading and template recombination. *Structure*, 19(7):955 – 966. 55

Muppirala, Usha, L.-B. D. D. (2013a). Computational tools for investigating RNA-protein interaction partners. *Journal of Computer Science & Systems Biology*, 06(04). 11

Muppirala, U. (2013b). *Computational prediction of RNA-protein interaction partners and interfaces*. PhD thesis, Iowa State University. 9, 10, 90, 107, 126

Muppirala, U., Honavar, V., and Dobbs, D. (2011). Predicting RNA-Protein Interactions Using Only Sequence Information. *BMC Bioinformatics*, 12(1):489. 12, 107, 108, 109, 110, 111, 113, 116, 120, 124

Murakami, Y., Spriggs, R. V., Nakamura, H., and Jones, S. (2010). PiRaNhA: a server for the computational prediction of RNA-Binding residues in protein sequences. *Nucleic Acids Research*, 38(suppl 2):W412–W416. 65, 71, 72, 73, 76, 131

Nagy, P. D. and Pogany, J. (2011). The dependence of viral RNA replication on co-opted host factors. *Nature Reviews Microbiology*, 10(2):137–149. 54

Nguyen, M. N. and Rajapakse, J. C. (2006). Two-stage support vector regression approach for predicting accessible surface areas of amino acids. *Proteins*, 63(3):542–550. 20

Norambuena, T. and Melo, F. (2010). The Protein-DNA Interface database. *BMC Bioinformatics*, 11(1):262. 132

Ofran, Y., Mysore, V., and Rost, B. (2007). Prediction of DNA-binding residues from sequence. *Bioinformatics*, 23(13):i347–i353. 32, 108

Pancaldi, V. and Bahler, J. (2011). In silico characterization and prediction of global protein-mRNA interactions in yeast. *Nucleic Acids Research*, 39(14):5826–5836. 106

Paz, I., Kosti, I., Ares, M., Cline, M., and Mandel-Gutfreund, Y. (2014). RBPmap: a web server for mapping binding sites of RNA-binding proteins. *Nucleic Acids Research*, 42(W1):W361–W367. 11, 90

Perez-Cano, L. and Fernandez-Recio, J. (2010a). Dissection and prediction of RNA-binding sites on proteins. *BioMol Concepts*, 1:345–355. 8, 9, 16, 17, 18

Perez-Cano, L. and Fernandez-Recio, J. (2010b). Optimal protein-RNA area, OPRA: A propensity-based method to identify rna-binding sites on proteins. *Proteins: Structure, Function, and Bioinformatics*, 78(1):25–35. 23, 26, 38, 44, 55, 65, 74, 76, 97, 131

Perez-Cano, L., JimenezeGarcia, B., and Fernandez-Recio, J. (2012). A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. *Proteins: Structure, Function, and Bioinformatics*. 140

Petrov, A. I., Zirbel, C. L., and Leontis, N. B. (2013). Automated classification of RNA 3D motifs and the RNA 3D Motif Atlas. *RNA*, 19(10):1327–1340. 11

Platt, J. C. (1999). *Fast training of support vector machines using sequential minimal optimization*, pages 185–208. MIT Press, Cambridge, MA, USA. 48

Pollastri, G., Przybylski, D., Rost, B., and Baldi, P. (2002). Improving the prediction of protein secondary structure in three and eight classes using recurrent neural networks and profiles. *Proteins: Structure, Function, and Bioinformatics*, 47(2):228–235. 20, 32

Prasanth, K. and Spector, D. (2007). Eukaryotic regulatory RNAs: an answer to the "genome complexity" conundrum. *Genes & Development*, 21(1):11–42. 2

Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J. M. (2012). Computational methods for prediction of protein-RNA interactions. *Journal of Structural Biology*, 179(3):261 – 268. 8, 9, 16, 17, 18, 26, 38, 40, 46, 54, 55, 63, 64, 65, 70, 71, 72, 73, 76, 78, 108

Qu, J., Kang, S. G., Wang, W., Musier-Forsyth, K., and Jang, J.-C. (2014). The arabidopsis thaliana tandem zinc finger 1 (AtTZF1) protein in RNA binding and decay. *The Plant Journal*, 78(3):452–467. 9

Qu, Z. and Adelson, D. L. (2012). Evolutionary Conservation and Functional Roles of ncRNA. *Frontiers in Genetics*, 3(205). 2

Rambo, R. P. and Tainer, J. A. (2010). Bridging the solution divide: comprehensive structural analyses of dynamic RNA, DNA, and protein assemblies by small-angle x-ray scattering. *Current Opinion in Structural Biology*, 20(1):128–137. 4

Ray, D., Kazan, H., Chan, E. T., Pena Castillo, L., Chaudhry, S., Talukder, S., Blencowe, B. J., Morris, Q., and Hughes, T. R. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature biotechnology*, 27(7):667–670. 5, 11, 90

Ray, D., Kazan, H., Cook, K. B., Weirauch, M. T., Najafabadi, H. S., Li, X., Gueroussov, S., Albu, M., Zheng, H., Yang, A., Na, H., Irimia, M., Matzat, L. H., Dale, R. K., Smith, S. A., Yarosh, C. A., Kelly, S. M., Nabet, B., Mecenas, D., Li, W., Laishram, R. S., Qiao, M., Lipshitz, H. D., Piano, F., Corbett, A. H., Carstens, R. P., Frey, B. J., Anderson, R. A., Lynch, K. W., Penalva, L. O. F., Lei, E. P., Fraser, A. G., Blencowe, B. J., Morris, Q. D., and Hughes, T. R. (2013). A compendium of RNA-binding motifs for decoding gene regulation. *Nature*, 499(7457):172–177. 11, 90

Riley, K. J. and Steitz, J. A. (2013). The "observer effect" in genome-wide surveys of protein-RNA interactions. *Molecular Cell*, 49(4):601–604. 6, 11, 89

Rinn, J. L. and Chang, H. Y. (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry*, 81. 2, 106, 122

Sarver, M., Zirbel, C., Stombaugh, J., Mokdad, A., and Leontis, N. (2008). FR3D: finding local and composite recurrent structural motifs in RNA 3D structures. *Journal of Mathematical Biology*, 56(1-2):215–252. 133

Sathyapriya, R., Vijayabaskar, M. S., and Vishveshwara, S. (2008). Insights into proteinâĂŞDNA interactions through structure network analysis. *PLoS Comput Biol*, 4(9):e1000170. 90

Scheibe, M., Butter, F., Hafner, M., Tuschl, T., and Mann, M. (2012). Quantitative mass spectrometry and PAR-CLIP to identify RNA-protein interactions. *Nucleic Acids Research*, 40(19):9897–9902. 6

Schmidt, C., Kramer, K., and Urlaub, H. (2012). Investigation of protein-RNA interactions by mass spectrometry–techniques and applications. *Journal of proteomics*, 75(12):3478–3494. 6

Schneider, T. R. (2000). Objective comparison of protein structures: error-scaled difference distance matrices. *Acta crystallographica. Section D, Biological crystallography*, 56(Pt 6):714–721. 140

Schneider, T. R. (2002). A genetic algorithm for the identification of conformationally invariant regions in protein molecules. *Acta Crystallographica. Section D, Biological Crystallography*, 58(Pt 2):195–208. 140

Schneider, T. R. (2004). Domain identification by iterative analysis of error-scaled difference distance matrices. *Acta Crystallographica. Section D, Biological Crystallography*, 60(Pt 12 Pt 1):2269–2275. 140

Schonrock, N. and Götz, J. (2012). Decoding the non-coding RNAs in Alzheimer's disease. *Cellular and Molecular Life Sciences*, 69(21):3543–3559. 54

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., and Selbach, M. (2011). Global quantification of mammalian gene expression control. *Nature*, 473(7347):337–342. 1

Shazman, S., Elber, G., and Mandel-Gutfreund, Y. (2011). From face to interface recognition: a differential geometric approach to distinguish DNA from RNA binding surfaces. *Nucleic Acids Research*. 8, 23

Shazman, S. and Mandel-Gutfreund, Y. (2008). Classifying RNA-binding proteins based on electrostatic properties. *PLoS Comput Biol*, 4(5):e1000146. 8, 23, 131

Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., Li, Y., and Jiang, H. (2007). Predicting proteinâĂŞprotein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341. 107, 110

Shulman-Peleg, A., Nussinov, R., and Wolfson, H. J. (2009). RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Research*, 37(suppl 1):D369–D373. 132

Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A., and Hulo, N. (2010). PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38(suppl 1):D161–D166. 10, 133

Sigrist, C. J. A., de Castro, E., Cerutti, L., Cuche, B. A., Hulo, N., Bridge, A., Bougueleret, L., and Xenarios, I. (2013). New and continuing developments at PROSITE. *Nucleic Acids Research*, 41(Database issue):D344–347. 10, 104

Sing, T., Sander, O., Beerenwinkel, N., and Lengauer, T. (2005). ROCR: visualizing classifier performance in R. *Bioinformatics*, 21(20):3940–3941. 50, 83, 113

Singh, G., Ricci, E. P., and Moore, M. J. (2014). RIPiT-seq: a high-throughput approach for footprinting RNA:protein complexes. *Methods (San Diego, Calif.)*, 65(3):320–332. 6

Spirin, S., Titov, M., Karyagina, A., and Alexeevski, A. (2007). NPIDB: a Database of Nucleic Acids-Protein Interactions. *Bioinformatics*, 23(23):3247–3248. 132

Spriggs, R. and Jones, S. (2009). RNA-binding residues in sequence space: Conservation and interaction patterns. *Computational Biology and Chemistry*, 33(5):397–403. 34, 56, 64

Spriggs, R., Murakami, Y., Nakamura, H., and Jones, S. (2009). Protein function annotation from sequence: prediction of residues interacting with RNA. *Bioinformatics*, 25(12):1492–1497. 17, 22, 35, 44, 55

Stajich, J. E., Block, D., Boulez, K., Brenner, S. E., Chervitz, S. A., Dagdigian, C., Fuellen, G., Gilbert, J. G., Korf, I., Lapp, H., Lehväslaiho, H., Matsalla, C., Mungall, C. J., Osborne, B. I., Pocock, M. R., Schattner, P., Senger, M., Stein, L. D., Stupka, E., Wilkinson, M. D., and Birney, E. (2002). The Bioperl Toolkit: Perl Modules for the Life Sciences. *Genome Research*, 12(10):1611–1618. 133

Standart, N. and Jackson, R. (1994). Regulation of translation by specific protein/mRNA interactions. *Biochimie*, 76(9):867–879. 54

Stefl, R., Skrisovska, L., and Allain, F. H.-T. (2005). RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle. *EMBO Reports*, 6(1):33–38. 89

Tadros, W., Goldman, A. L., Babak, T., Menzies, F., Vardy, L., Orr-Weaver, T., Hughes, T. R., Westwood, J. T., Smibert, C. A., and Lipshitz, H. D. (2007). SMAUG is a major regulator of maternal mRNA destabilization in drosophila and its translation is activated by the PAN GU kinase. *Developmental Cell*, 12(1):143–155. 54

Tan, L., Yu, J.-T., Hu, N., and Tan, L. (2013). Non-coding RNAs in Alzheimer's Disease. *Molecular Neurobiology*, 47(1):382–393. 54

Terribilini, M., Lee, J., Yan, C., Jernigan, R., Carpenter, S., Honavar, V., and Dobbs, D. (2006a). Identifying Interaction Sites in "Recalcitrant" Proteins: Predicted Protein and RNA Binding Sites in Rev Proteins of HIV-1 and EIAV Agree with Experimental Data. *Pac Symp Biocomput*, pages 415–426. 133

Terribilini, M., Lee, J., Yan, C., Jernigan, R. L., Honavar, V., and Dobbs, D. (2006b). Prediction of RNA binding sites in proteins from amino acid sequence. *RNA*, 12(8):1450–1462. 7, 8, 17, 20, 21, 24, 45, 55, 131, 133

Terribilini, M., Sander, J., Lee, J., Zaback, P., Jernigan, R., Honavar, V., and Dobbs, D. (2007). RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. *Nucleic Acids Res*, 35(WebServerissue):W578–584. 8, 20, 21, 24, 45, 65, 72, 76, 133

Tome, J. M., Ozer, A., Pagano, J. M., Gheba, D., Schroth, G. P., and Lis, J. T. (2014). Comprehensive analysis of RNA-protein interactions by high-throughput sequencing-RNA affinity profiling. *Nature Methods*, 11(6):683–688. 6

Tong, J., Jiang, P., and Lu, Z. (2008). RISP: a web-based server for prediction of RNA-binding sites in proteins. *Comput Methods Programs Biomed*, 90(2):148–153. 17, 21, 33, 44

Towfic, F., Caragea, C., Gemperline, D. C., Dobbs, D., and Honavar, V. (2010). Struct-NB: predicting protein-RNA binding sites using structural features. *International Journal of Data Mining and Bioinformatics*, 4(1):21–43. 7, 8, 17, 24, 44, 55, 131

Treger, M. and Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *J Mol Recognit*, 14(4):199–214. 7, 17, 96

Tsai, M., Spitale, R. C., and Chang, H. Y. (2011). Long intergenic noncoding RNAs: new links in cancer progression. *Cancer Research*, 71(1):3–7. 54

Tuerk, C. and Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, 249(4968):505–510. 5

Tuukkanen, A. T. and Svergun, D. I. (2014). Weak protein-ligand interactions studied by small-angle x-ray scattering. *The FEBS journal*, 281(8):1974–1987. 4

Ule, J., Jensen, K., Mele, A., and Darnell, R. B. (2005). CLIP: A method for identifying protein-RNA interaction sites in living cells. *Methods*, 37(4):376 – 386. 5, 16, 54

Valverde, R., Edwards, L., and Regan, L. (2008). Structure and function of KH domains. *FEBS Journal*, 275(11):2712–2726. 89

Van Roosbroeck, K., Pollet, J., and Calin, G. A. (2013). miRNAs and long noncoding RNAs as biomarkers in human diseases. *Expert Review of Molecular Diagnostics*, 13(2):183–204. 54

Walia, R., Caragea, C., Lewis, B., Towfic, F., Terribilini, M., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (2012). Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art. *BMC Bioinformatics*, 13(1):89. 8, 9, 55, 61, 62, 67, 69, 71, 72, 73, 77, 81, 108

Walia, R. R., Xue, L. C., Wilkins, K., El-Manzalawy, Y., Dobbs, D., and Honavar, V. (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. 108

Walker, S. C., Scott, F. H., Srisawat, C., and Engelke, D. R. (2008). RNA affinity tags for the rapid purification and investigation of RNAs and RNA-protein complexes. *Methods in Molecular Biology (Clifton, N.J.)*, 488:23–40. 6

Wang, C.-C., Fang, Y., Xiao, J., and Li, M. (2011). Identification of RNA-binding sites in proteins by integrating various sequence information. *Amino Acids*, 40:239–248. 17, 21, 55, 131

Wang, G. and Dunbrack, Roland L, J. (2003). PISCES: a protein sequence culling server. *Bioinformatics*, 19(12):1589–1591. 78, 79

Wang, L. and Brown, S. (2006a). Prediction of RNA-binding residues in protein sequences using support vector machines. *Proc of the 26th IEEE EMBS Ann Int Conf*, pages 5830–5832. 17, 21, 55, 65, 69, 71, 72, 76, 77, 131

Wang, L. and Brown, S. J. (2006b). BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences. *Nucleic Acids Research*, 34(Web Server issue):W243–248. 21, 44, 55, 108

Wang, L., Huang, C., Yang, M., and Yang, J. (2010a). BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. *BMC Systems Biology*, 4(Suppl 1):S3. 21, 55, 65, 69, 71, 72, 73, 76, 77, 108, 131

Wang, Y., Ludwig, J., Schuberth, C., Goldeck, M., Schlee, M., Li, H., Juranek, S., Sheng, G., Micura, R., Tuschl, T., Hartmann, G., and Patel, D. J. (2010b). Structural and functional insights into 5'-ppp RNA pattern recognition by the innate immune receptor RIG-I. *Nature Structural and Molecular Biology*, 17(7):781–787. 68

Wang, Y., Xue, Z., Shen, G., and Xu, J. (2008). PRINTR: prediction of RNA binding sites in proteins using SVM and profiles. *Amino Acids*, 35(2):295–302. 8, 17, 24, 33, 44

Witten, I. and Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques.* San Francisco: Morgan Kaufmann, 2 edition. 47, 48

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L.-S. L., Natale, D. A., Vinayaka, C. R., Hu, Z.-Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R. S., Suzek, B. E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J. L., Chung, S., Castro-Alvear, J., Dinkov, G., and Barker, W. C. (2004). PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research*, 32(suppl 1):D112–D114. 138

Wu, H., Finger, L. D., and Feigon, J. (2005). Structure determination of protein/RNA complexes by NMR. *Methods in Enzymology*, 394:525–545. 4, 54

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., Zhang, Z., SkogerbÃÿ, G., Chen, L., Lu, H., Zhao, Y., and Chen, R. (2006). NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Research*, 34(suppl 1):D150–D152. 6, 12, 109

Xue, L. C., Dobbs, D., and Honavar, V. (2011). HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformatics*, 12(1):244. 56, 57, 108

Xue, L. C., Jordan, R. A., Yasser, E.-M., Dobbs, D., and Honavar, V. (2014). DockRank: Ranking docked conformations using partner-specific sequence homology-based protein interface prediction. *Proteins: Structure, Function, and Bioinformatics*, 82(2):250–267. 55

Xue, S. and Barna, M. (2012). Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat Rev Mol Cell Biol*, 13(6):355–369. 1

Yuan, J., Wu, W., Xie, C., Zhao, G., Zhao, Y., and Chen, R. (2014). NPInter v2.0: an updated database of ncRNA interactions. *Nucleic Acids Research*, 42(Database issue):D104–108. 6, 12, 109, 125

Zehetner, G. (2003). OntoBlast function: From sequence similarities directly to potential functional annotations by ontology terms. *Nucleic Acids Research*, 31(13):3799–3803. 55

Zhang, Q. C., Deng, L., Fisher, M., Guan, J., Honig, B., and Petrey, D. (2011). PredUs: a web server for predicting protein interfaces using structural neighbors. *Nucleic Acids Research*, 39(suppl 2):W283–W287. 56

Zhang, Q. C., Petrey, D., Norel, R., and Honig, B. H. (2010a). Protein interface conservation across structure space. *Proceedings of the National Academy of Sciences USA*, 107:10896–10901. 56

Zhang, T., Zhang, H., Chen, K., Ruan, J., Shen, S., and Kurgan, L. (2010b). Analysis and Prediction of RNA-Binding Residues Using Sequence, Evolutionary Conservation, and Predicted Secondary Structure and Solvent Accessibility. *Current Protein and Peptide Science*, 11(7):609–628. 17, 22

Zhao, H., Yang, Y., and Zhou, Y. (2010). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Research*. 23, 26, 38, 39, 40, 44, 71

Zhao, H., Yang, Y., and Zhou, Y. (2013). Prediction of RNA binding proteins comes of age from low resolution to high resolution. *Molecular BioSystems*. 122

Zheng, G., Lu, X.-J., and Olson, W. K. (2009). Web 3DNA-a web server for the analysis, reconstruction, and visualization of three-dimensional nucleic-acid structures. *Nucleic Acids Research*, 37(suppl 2):W240–W246. 132