Computational analysis and prediction of protein-RNA interactions

Michael Terribilini
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Cell and Developmental Biology Commons, and the Genetics and Genomics Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate College at Digital Repository @ Iowa State University. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Computational analysis and prediction of protein-RNA interactions

by

Michael Joseph Terribilini

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:
Drena Dobbs, Co-Major Professor
Vasant Honavar, Co-Major Professor
Robert L. Jernigan
Gustavo MacIntosh
Allen Miller

Iowa State University
Ames, Iowa
2008

Copyright © Michael Joseph Terribilini, 2008. All rights reserved.
DEDICATION

For Tracey, Adam, Emily, Belle, and Mary. I could not have done this without your support, encouragement, patience and love.
TABLE OF CONTENTS

ABSTRACT............................................................................................................................... Errors! Bookmark not defined.

CHAPTER 1. GENERAL INTRODUCTION ........................................................................ 1
INTRODUCTION .......................................................................................................................... 1
OVERALL GOALS ....................................................................................................................... 23
DISSERTATION ORGANIZATION ........................................................................................... 24
REFERENCES .......................................................................................................................... 27

CHAPTER 2. PREDICTION OF RNA BINDING SITES IN PROTEINS FROM AMINO
ACID SEQUENCE..................................................................................................................... 33
ABSTRACT ............................................................................................................................... 33
INTRODUCTION ....................................................................................................................... 34
RESULTS ..................................................................................................................................... 36
DISCUSSION ............................................................................................................................. 54
METHODS ................................................................................................................................ 58
ACKNOWLEDGEMENTS ............................................................................................................ 64
REFERENCES ............................................................................................................................ 64

CHAPTER 3 RNABINDR: A SERVER FOR ANALYZING AND PREDICTING RNA
BINDING SITES IN PROTEINS............................................................................................... 71
ABSTRACT ............................................................................................................................... 71
INTRODUCTION ....................................................................................................................... 71
MATERIALS AND METHODS ................................................................................................. 73
SERVER DESCRIPTION ............................................................................................................. 77
RELATED SERVERS .................................................................................................................. 82
SIGNIFICANCE AND FUTURE DIRECTIONS ......................................................................... 83
ACKNOWLEDGEMENTS ............................................................................................................ 85
REFERENCES ............................................................................................................................ 85

CHAPTER 4. PREDICTING RNA BINDING SITES IN PROTEINS: USE OF
STRUCTURAL AND EVOLUTIONARY INFORMATION LEADS TO IMPROVED
PERFORMANCE......................................................................................................................... 89
ABSTRACT ............................................................................................................................... 89
INTRODUCTION ....................................................................................................................... 90
RESULTS ..................................................................................................................................... 93
DISCUSSION ............................................................................................................................. 115
METHODS ................................................................................................................................ 120
REFERENCES ............................................................................................................................ 125

CHAPTER 5. PRIDB: PROTEIN-RNA INTERACTION DATABASE................................. 133
ABSTRACT ............................................................................................................................... 133
INTRODUCTION ....................................................................................................................... 134
METHODS ................................................................................................................................ 135
RESULTS ..................................................................................................................................... 139
DISCUSSION ............................................................................................................................. 162
REFERENCES ............................................................................................................................ 167
ABSTRACT

Protein-RNA interactions are essential for many important processes including all phases of protein production, regulation of gene expression, and replication and assembly of many viruses. This dissertation has two related goals: 1) predicting RNA-binding sites in proteins from protein sequence, structure, and conservation information, and 2) characterizing protein-RNA interactions.

We present several machine learning classifiers for predicting RNA-binding sites in proteins based on the protein sequence, protein structure, and conservation information. Our first classifier uses only amino acid sequence information as input and predicts RNA-binding sites with an area under the receiver operator characteristic curve (AUC) of 0.74. Using the neighboring amino acids in the protein structure improves prediction performance over using sequence alone. We show that using evolutionary information in the form of position specific scoring matrices provides a further significant improvement in predictions. Finally, we create an ensemble classifier that combines the predictions of the sequence, structure, and PSSM based classifiers and gives the best prediction performance, with an AUC of 0.81.

We construct the Protein-RNA Interaction Database, PRIDB, a comprehensive collection of all protein-RNA complexes in the PDB. PRIDB focuses on characterizing the molecular interaction at the protein-RNA interface in terms of van der Waals contacts, direct hydrogen bonds, and water-mediated hydrogen bonds. We perform an extensive analysis of the RNA-binding characteristics of a non-redundant dataset of 181 proteins to determine general characteristics of protein-RNA binding sites. We find that the overall interaction propensities for Watson-Crick paired nucleotides and non Watson-Crick paired nucleotides are very similar, with the propensities for amino acids binding to single stranded nucleotides showing more differences. We find that van der Waals contacts are more numerous than
hydrogen bonds and amino acids interact with RNA through their side chain atoms more frequently than their main chain atoms. We also find that contacts to the RNA base are not as frequent as contacts to the RNA backbone.

Together, the prediction and characterization presented in this dissertation have increased our understanding of how proteins and RNA interact.
CHAPTER 1. GENERAL INTRODUCTION

This dissertation characterizes and predicts molecular interactions between proteins and RNA. In this study, we develop several machine learning classifiers for predicting which amino acids in a protein are likely to bind RNA. The classifiers were developed to use information from a single protein sequence, from multiple related protein sequences, and from protein structure as input. This work also describes the Protein-RNA Interaction Database, PRIDB, a comprehensive database of all protein-RNA complexes with experimentally determined structures at atomic resolution. Finally, a non-redundant dataset of 181 RNA binding proteins was analyzed to determine characteristics of protein-RNA interaction sites.

INTRODUCTION

RNA is one of the most diverse biological molecules performing such varied tasks as storing genetic information, controlling the release of gene expression information in the cell, and enzymatic catalysis of biological reactions. In all known biological systems, RNA interacts with proteins to accomplish these tasks. Molecular recognition between proteins and RNA is a complex and varied process, ranging from sequence-specific interactions, to recognition through shape complementarity, to non-specific interactions. The detailed mechanisms of protein-RNA interactions are poorly understood, and concepts learned from individual complexes have not always been applicable to general studies. Much of what we know about protein-RNA recognition has come from the solved structures of protein-RNA complexes. Of the thousands of known RNA-binding proteins, there are currently only about 500 such complexes available, containing fewer than 200 unique proteins. Computational methods of identifying and characterizing protein-RNA interfaces are needed to bridge the gap between the available sequence and structure data.
In this work, we have two related goals. First, we develop methods for predicting RNA-binding sites in proteins using information from the protein sequence or structure. Second, we create a database of protein-RNA complexes and analyze a non-redundant set of proteins to determine some general features of protein-RNA interfaces.

The following is a review of published methods for predicting RNA-binding sites in proteins, computational analyses of protein-RNA interfaces, and databases of protein-RNA complexes.

**Methods for predicting RNA-binding sites in proteins**

At the time this dissertation was initiated, there were no published methods for predicting RNA-binding sites in proteins. Over the past four years, eleven papers have been published that describe machine learning approaches to this problem (Jeong et al., 2004, Jeong and Miyano, 2006, Terribilini et al, 2006, Terribilini et al., 2007, Wang and Brown, 2006a, Wang and Brown 2006b, Kim et al., 2006, Kumar et al., 2007, Tong et al., 2008, Wang et al., 2008, Towfic et al., 2008), and two more biophysical methods have also been developed (Chen and Lim, 2008, Shulman-Peleg et al., 2008). This explosion of interest reflects the increasing amounts of structural data becoming available for protein-RNA complexes and indicates the growing importance of such methods for computationally identifying RNA-binding sites. The published methods can be grouped into three major categories: i) single sequence methods, ii) multiple sequence methods, and iii) methods incorporating protein structure information. There is some overlap between the categories; for example, many multiple sequence methods also use structural information. Here, we present a summary of the published methods with an emphasis on comparing the encoding of input information, and illustrate some of the challenges in the field. The summary is organized by the research group that developed the methods and is presented in
chronological order, from the earliest methods to the most recent publications. Table 1.1 provides a brief overview of the published prediction methods.
<table>
<thead>
<tr>
<th>Group</th>
<th>Algorithm</th>
<th>Input Info</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jeong et al., 2004</td>
<td>Neural Network</td>
<td>Binary vector + predicted secondary structure</td>
<td>Single sequence</td>
</tr>
<tr>
<td>Jeong and Miyano, 2006</td>
<td></td>
<td>Weighted PSSM inputs</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td>Terribilini et al., 2006</td>
<td>Naive Bayes</td>
<td>Amino acid identities</td>
<td>Single sequence</td>
</tr>
<tr>
<td>Terribilini et al., 2007</td>
<td>Naive Bayes</td>
<td>Amino acid identities spatial neighbors</td>
<td>Single sequence + structure</td>
</tr>
<tr>
<td></td>
<td>SVM (RBF)</td>
<td>PSSM</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td></td>
<td>SVM (RBF)</td>
<td>PSSM spatial neighbors</td>
<td>Multiple sequence + structure</td>
</tr>
<tr>
<td></td>
<td>Naive Bayes</td>
<td>Ensemble – predicted probability of RNA-binding from the previous three rows</td>
<td>Multiple sequence + structure</td>
</tr>
<tr>
<td>Wang and Brown, 2006</td>
<td>SVM (RBF)</td>
<td>Hydrophobicity</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrophobicity + pKa</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrophobicity + pKa + mass</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrophobicity + pKa + mass + conservation</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hydrophobicity + pKa + mass + conservation + predicted solvent accessibility</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td>Kim et al., 2006</td>
<td>N/A</td>
<td>Single amino acid RNA-binding propensity (singlet)</td>
<td>Multiple sequence + structure</td>
</tr>
<tr>
<td></td>
<td></td>
<td>singlet + profile</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Averaged singlet – singlet propensity of all residues within 7 Angstoms</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Average singlet + profile</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Singlet + doublet propensity</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Singlet + doublet + profile</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Singlet + doublet average of all within 7 Angstroms</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Averaged singlet + doublet + profile</td>
<td></td>
</tr>
<tr>
<td>Kumar et al., 2007</td>
<td>SVM No kernel specified</td>
<td>Binary vector</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSSM</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td>Tong et al., 2008</td>
<td>SVM (RBF)</td>
<td>PSSM</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td>Wang et al., 2008</td>
<td>SVM (RBF)</td>
<td>Binary vector</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Binary vector + predicted secondary structure</td>
<td>Single sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSSM</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSSM + predicted secondary structure</td>
<td>Multiple sequence</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PSSM + actual secondary structure + actual solvent accessibility</td>
<td>Multiple sequence + structure</td>
</tr>
</tbody>
</table>
First, we present definitions for several commonly used measures of prediction performance (Baldi et al., 2001).

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

\[
\text{Specificity}^+ = \frac{TP}{TP + FN}
\]

\[
\text{Sensitivity}^+ = \text{True Positive Rate} = \frac{TP}{TP + FN}
\]

\[
\text{Correlation Coefficient} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]

\[
\text{False Positive Rate} = \frac{FP}{FP + TN}
\]

All machine learning methods have an inherent trade-off between specificity and sensitivity that is controlled through the classification threshold. A useful method of comparing classifiers across all classification thresholds is the receiver operating characteristic (ROC) curve. A ROC curve plots the false positive rate against the true positive rate. The area under the ROC curve can be used to compare the total performance of classifiers. A perfect classifier would have an AUC of 1, while a classifier that makes random guesses would have an AUC of 0.5.

The first published method for predicting RNA-binding sites in proteins was developed by Jeong, Chung, and Miyano in 2004 (Jeong et al., 2004). The authors created a non-redundant dataset of 96 proteins from protein-RNA complexes in the PDB. RNA binding residues were defined as any amino acid with any atoms within 6 Å of any RNA.
The authors developed a neural network classifier that used the identity of the target residue, plus the identities of the preceding and following \( n \) residues in the sequence as input as well as the predicted secondary structure of the target residue. The sequence window input is commonly referred to as a sliding window because the window is slid along the sequence until input instances have been created for all residues. The authors encoded each amino acid as a 21 element vector with all elements set to 0, and one element set to 1. The element set to 1 uniquely identifies one of the 20 amino acids, with the 21st element being used to indicate a position in the sliding window that is past the end of the protein sequence. This encoding is quite common in machine learning applications in bioinformatics, and we will refer to it as the binary vector encoding from this point on. The secondary structure of each residue was predicted with the PHD program (Rost and Sander, 1994, Rost and Sander, 1993). With this input encoding, we categorize this method as a single sequence method, however, the secondary structure prediction method actually utilizes multiple sequence alignment information, resulting in the implicit inclusion of evolutionary information.

Jeong et al. experimented with different window sizes ranging from 7 to 57 amino acids and reported their best prediction performance with a window of 41 amino acids. They reported an accuracy of 77.5\%, specificity\(^+\) of 46.7\%, sensitivity\(^+\) of 40.3\%, and correlation coefficient of 0.294.

In 2006, Jeong and Miyano extended their previous work by developing a multiple sequence based method. Using 87 proteins from their 2004 work, Jeong and Miyano developed a new neural network classifier using position specific scoring matrices (PSSMs) as inputs instead of a binary vector. Nine of the previously used proteins were removed due to having fewer than four RNA-binding residues or not producing valid PSSMs. The authors experimented with different methods for generating the PSSM. First, a PSI-BLAST search (Altschul et al., 1997) was performed for each sequence in the dataset to identify related
sequences. PSI-BLAST was run for three iterations with an E-value cutoff of 0.0001 for inclusion in the next iteration. The related sequences identified by PSI-BLAST were then used as input to ClustalW (Thompson et al., 1994), HMMER (Eddy, 1998), or the PSI-BLAST generated PSSM was used directly. A sliding window of 7 to 45 amino acids was used as input to the neural network, with each amino acid represented by the PSSM vector. The best prediction performance was obtained with the PSI-BLAST PSSM inputs using a window size of 15 amino acids. Jeong and Miyano reported a correlation coefficient of 0.41, with an AUC of 0.77.

In 2006, Terribilini et al. developed a single sequence method using only the amino acid identities as input (Terribilini et al., 2006, Chapter 2). The authors created a dataset of 109 non-redundant proteins and defined RNA-binding residues using the program ENTANGLE (Allers and Shamoo, 2001). ENTANGLE defines RNA-binding residues as being involved in hydrogen bonds, electrostatic interactions, stacking interactions, van der Waals interactions, and hydrophobic contacts. Using a window of 25 amino acids as input to a Naïve Bayes classifier, Terribilini et al. reported an accuracy of 84.8%, a specificity+ of 51%, a sensitivity+ of 38%, and a correlation coefficient of 0.35. Terribilini et al. presented the predictions made by their method on the telomerase reverse transcriptase protein and showed that the predicted RNA-binding residues are in good overall agreement with experimentally determined RNA-binding sites. This represented the first published application of a RNA-binding site prediction method being applied successfully to a protein without a known protein-RNA complex. Finally, Terribilini et al. obtained the dataset used by Jeong et al. in 2004 (Jeong et al., 2004) and performed a direct comparison with the neural network method. Terribilini et al. found that their Naïve Bayes classifier had comparable performance with the more complex neural network classifier.

Terribilini et al. described their web server, RNABindR, in 2007 (Terribilini et al., 2007). The prediction method was essentially the same as described above, but results were
based on an updated dataset of 147 RNA-binding proteins (RB147). They also changed their
definition of RNA-binding residues to any amino acid with any atoms within 5 Angstroms
(Å) of any RNA atoms. With the larger dataset and new definition of RNA-binding residues,
Terribilini et al reported an accuracy of 86%, specificity\(^+\) of 61%, sensitivity\(^+\) of 33%, and
correlation coefficient of 0.36.

The most recent efforts of Terribilini et al. are described in Chapter 4 of this
dissertation. First, a new dataset of 181 non-redundant RNA-binding proteins (RB181) was
created. RNA-binding residues were defined using the 5 Å distance-cutoff. Five classifiers,
distinguished by the input information used, were presented. The first classifier was the
same as described above, a Naïve Bayes classifier using only a sliding window of amino acid
identities as input. The second classifier presented also used a window of amino acid
identities as input, but took a novel approach by using the nearest spatial neighbors within the
known protein \textit{structure} rather than the neighboring residues in the linear protein \textit{sequence}.
The third classifier was a multiple sequence based classifier, that used a PSSM vector
produced by PSI-BLAST to encode each amino acid in a sliding sequence window. The
fourth classifier used both the PSSM encoding for each amino acid and the nearest
neighboring amino acids in the structure to define the context of each residue. The PSSM
based classifiers were Support Vector Machine (SVM) classifiers using the radial basis
function (RBF) kernel. The fifth classifier was an Ensemble classifier, which used the
predicted probability of RNA-binding from their other classifiers as input to a Naïve Bayes
classifier. Terribilini et al. reported their best performance with the Ensemble classifier,
achieving an accuracy of 85.5%, specificity\(^+\) of 53%, sensitivity\(^+\) of 49%, correlation
coefficient of 0.43, and AUC of 0.811.

Terribilini et al. also reported several analyses of interest. First, they evaluate each of
their classifiers on each of the three datasets they generated, ranging from 109 to 181
proteins. They found that the AUC of each classifier was essentially the same on the three
datasets, indicating that prediction of RNA-binding residues has not improved as the non-redundant datasets available increased. They demonstrated that having a greater number of related sequences available to build the input PSSM made it much more likely that the PSSM based classifier will have improved prediction performance over the single sequence classifier, indicating that prediction of RNA-binding sites in proteins is likely to improve as more divergent sequence data is obtained. Finally, Terribilini et al. analyzed the types of contacts made by each amino acid in their largest dataset and grouped them based on whether the contact was to the RNA base only, the RNA backbone only, or the RNA base and backbone. They found that prediction performance was best for the amino acids that contact both the base and backbone, and worst for amino acids that contact only the RNA base.

In 2006, Wang and Brown published two similar papers describing their method for predicting RNA-binding sites in proteins (Wang and Brown, 2006, Wang and Brown, 2006). They created a dataset of 107 non-redundant proteins and defined RNA-binding residues using a 3.5 Å all-atom distance cutoff. Wang and Brown experimented with five different encodings for each amino acid. Three physico-chemical properties were used: hydrophobicity, side chain pKa value, and molecular mass of the amino acid. They also used the predicted solvent accessibility and the residue conservation score among related sequences as inputs to the SVM classifier. The final input to the classifiers was at least one of these features for a sliding window of the amino acid sequence. Wang and Brown created both neural network and SVM classifiers and found that their best prediction performance was obtained using window size of 11 amino acids, with all five features as input to a SVM classifier. This classifier achieved an accuracy of 74.3%, sensitivity of 65.8%, and an AUC of 0.754.

Kim et al. created a prediction method for RNA-binding sites based mainly on the protein structure and sequence conservation (Kim et al., 2006). They created a dataset of 86 proteins and defined an RNA-binding residue as any amino acid that had a lower solvent
accessibility in the protein-RNA complex than in the protein alone. They then calculated the interface propensity for each of the 20 amino acids as the fraction of surface residues that bind RNA divided by the fraction of total surface residues. They defined a similar propensity for pairs of amino acids on the surface, which they called the doublet propensity. To be considered a doublet, the two amino acids had to be within 7 Å of each other in the protein structure. Finally, they assign each residue a score based on how conserved it was in related sequences. Using these three values, they defined several prediction scores using values for each individual amino acid and average values over nearby residues in the protein structure. They obtained the best prediction performance using an average of the singlet and doublet scores over neighboring residues and the conservation score. We note that this method does not use any machine learning; it simply applies a threshold to the score assigned to each amino acid to determine if it is RNA-binding or not. Also, this method requires a protein structure to make a prediction. Kim et al. applied their prediction method to the structure of the nuclear mRNA export system and reported good agreement between their predicted binding residues and the available experimental data.

Kumar et al. developed a multiple sequence based classifier (Kumar et al., 2008) using the dataset of 86 proteins from Jeong et al. (Jeong et al., 2006). They also used the dataset of 107 proteins from Wang and Brown (Wang and Brown, 2006, Wang and Brown, 2006). They encoded each amino acid as either a binary vector or a PSSM vector obtained from a PSI-BLAST search against the NCBI nr database. A sliding window was used as input to a SVM classifier. Their best prediction performance was obtained using the PSSM encoding, which gave an accuracy of 81.2% and a correlation coefficient of 0.45. Kumar et al. performed an interesting comparison between SVM, neural network, and Naïve Bayes classifiers. They found that when evaluated on the same dataset and using the same input information (i.e., only the amino acid identities or binary vector encoding) the three algorithms obtained virtually equal performance.
Tong et al. (Tong et al., 2008) independently developed essentially the same method as described by Kumar et al.

Wang et al. (Wang et al., 2008) also created a very similar PSSM-based SVM classifier; however they experimented with some additional types of input information. They developed a classifier that used both the PSSM encoding and the predicted secondary structure for the target residue. They found that the predicted secondary structure input gave no improvement in prediction performance. They also used the actual secondary structure and the actual solvent accessibility of the target residue in addition to the PSSM encoding. This combination of input information gave a slight improvement over the PSSM encoding alone. The correlation coefficient was increased from 0.432 to 0.457 and the AUC increased from 0.82 to 0.83.

Towfic et al. (Towfic et al., 2008) experimented with several structural features to identify ways of using protein structure to improve prediction of RNA-binding residues. Using the RB147 dataset of Terribilini et al. (Terribilini et al., 2007), they computed the CX value and the surface roughness for each residue. The CX value (Pinatar et al., 2002) is a measure of the protrusion of the residue and the surface roughness (Lewis and Rees, 1985) measures if the surface is smooth or irregular. Towfic et al. created a number of Naïve Bayes classifiers using a sliding window approach and adding as input the CX value or the surface roughness value, or both. Their best performance was achieved with an ensemble of Naïve Bayes classifiers that used both the CX values and the surface roughness values. The AUC of this classifier was 0.752 compared to 0.736 for a simple sequence based Naïve Bayes classifier.

Chen and Lim developed a novel method to predict RNA-binding sites in a protein structure (Chen and Lim, 2008). They used a dataset of 69 non-redundant proteins and defined RNA-binding residues as any amino acid in van der Waals contact with RNA, or any amino acid involved in either a direct or water-mediated hydrogen bond with RNA. They
gave each residue an electrostatic rank based on whether a substitution of the negatively charged residue aspartate or glutamate would stabilize the structure. This was based on the assumption that stabilization upon mutation to a negatively charged amino acid indicates that there is a positive charge in the local region of the protein structure, and may therefore be important in RNA-binding. They also assigned each residue a conservation score. They identified surface patches and clefts that had a high ranking for both the electrostatic score and the conservation score. These patches and clefts represented the most conserved positively charged regions of the protein. The highest scoring surface patch and cleft were predicted to be RNA binding sites. They reported some overlap between their predicted binding sites and the actual binding sites in most proteins in their dataset.

The method of Chen and Lim provides an interesting alternative to the machine learning based methods. It is based on the properties of protein-RNA interfaces observed in many structures and attempts to apply these observations to prediction of new binding sites. The drawback of this method is that it predicts only a single surface patch and a single cleft for each protein. Proteins that do not have a single cleft or surface patch will not be suitable for this method.

Shulman-Peleg et al. have recently created a specialized method for predicting binding pockets for single-stranded nucleotides (Shulman-Peleg et al., 2008). Their approach was to extract general features of binding pockets for single-stranded nucleotides from known protein-RNA complexes, then search for similar pockets in structures of free proteins. After identifying potential binding pockets, they attempted to fit a modeled RNA structure into the binding pocket. This fascinating approach not only aims to predict potential RNA binding sites, but also the structure of the protein-RNA complex. By limiting themselves to single-stranded nucleotide binding pockets accommodating only one or two nucleotides, the authors illustrate that one way to improve prediction of RNA-binding site prediction is to build several highly specialized prediction methods.
Conclusions on methods for predicting RNA-binding sites in proteins

Much progress has been made in only a few years on predicting RNA-binding sites in proteins. Simple single sequence methods have been replaced with multiple sequence methods that achieve much better prediction performance. Kumar et al. (Kumar et al., 2007) showed that all current methods using PSSM-based inputs achieve roughly equal performance. When available, structural information can increase prediction performance slightly over multiple sequence information alone (Chapter 4, Wang et al., 2008).

Despite the progress that has been made, there are some significant challenges facing the field. First, a detailed comparison of the different prediction methods has been difficult, due to disagreements regarding the definition of RNA-binding residues and differences in datasets used in each study. Several authors have attempted comparisons of different ways of defining RNA-binding residues, but results were based on different datasets, which may lead to false conclusions. We have provided our datasets on the RNABindR website so that other groups can use exactly the same data we have used in comparing methods. Second, it has been difficult to identify features of the protein structure that significantly enhance prediction of RNA-binding sites. One final challenge for the field is identifying reasons for failures. Prediction performance varies widely on different proteins, ranging from excellent predictions to useless predictions that are worse than random guessing. To date, no detailed analysis has been published explaining how or why RNA binding sites in some proteins are predicted so poorly. Understanding why the prediction methods fail on some proteins will indicate important directions for future improvements.

Computational analyses of protein-RNA interfaces

Computational studies of protein-RNA interactions began in earnest in 2001. Jones et al. (Jones et al., 2001) created a set of protein-RNA complexes by extracting all complexes with at least 3 Å resolution from the Nucleic Acid Database, NDB (Berman et al., 1992) and
removing redundant sequences. The final dataset included 32 protein-RNA complexes. They defined an RNA-binding residue as any amino acid that lost at least 1 square Å of solvent accessible surface area after binding RNA. Direct hydrogen bonds were also considered in the analysis. Jones et al. defined the interface propensity of an amino acid as the fraction of accessible surface area of the RNA-binding site divided by the fraction of accessible surface area of the entire protein.

Jones et al. (Jones et al., 2001) found that the positively charged and aromatic amino acids had the highest propensities for binding RNA. Interestingly, they also observed that isoleucine had a high RNA-binding propensity. They observed that van der Waals contacts were much more common than hydrogen bonds, with van der Waals contacts comprising 92% of the total interactions observed. Both van der Waals and hydrogen bond contacts were more often found to involve the base atoms of the RNA rather than the backbone atoms. The base guanine was found to be favored for contact by amino acids.

This early study identified the importance of positively charged amino acids and aromatic amino acids in protein-RNA interfaces. However, some of the Jones et al. (Jones et al., 2001) results have not been observed in later studies with larger datasets. For example, all later studies found that contacts to the RNA backbone are more prevalent than contacts to the RNA base, and isoleucine has not been a favored RNA-binding residue.

Treger and Westhof performed an analysis of protein-RNA interfaces (Treger and Westhof, 2001). Their dataset was created by extracting all non-homologous protein-RNA complexes available in the PDB at the time. The dataset contained 45 complexes. RNA-binding residues were defined as those amino acids that form ionic bonds, hydrogen bonds, or van der Waals contacts. Each type of contact was defined by the distance between atoms and ionization state of the atoms. Effectively, an amino acid was characterized as being RNA-binding if any atom was within 3.8 Å of an atom in the bound RNA.
Treger and Westhof observed that the most favorable amino acids in protein-RNA interfaces were arginine, lysine, asparagine, and serine. The least favored amino acids were alanine, isoleucine, leucine, and valine. In this analysis, positively charged and highly polar residue were preferred for binding RNA. Hydrophobic residues were not preferred, but because this study did not limit analysis to surface residues, this finding may simply result from including many buried residues in the analysis of binding sites. Treger and Westhof also observed that contacts to the RNA backbone were more numerous than contacts to the RNA bases. There was no preference for binding to any particular RNA base, in contrast to the study of Jones et al. (Jones et al., 2001) in which a preference for guanine was reported. Treger and Westhof further observed that van der Waals contacts dominated protein-RNA interactions, forming 72% of all contacts observed. Contacts with the side chain atoms of the amino acids were more frequent than main chain contacts.

The study of Treger and Westhof (2001) produced many observations that have held up as the datasets analyzed have become larger. The preferences for backbone rather than base contacts and side chain rather than main chain contacts have been observed in later studies (Bahadur et al., 2008).

Allers and Shamoo developed a program, ENTANGLE, for calculating interactions between proteins and RNA and used it to analyze a set of 45 protein-RNA complexes (Allers and Shamoo, 2001). ENTANGLE classifies interactions into five categories, hydrogen bonds, stacking interactions, electrostatic interactions, van der Waals interactions, and hydrophobic interactions. Each type of interaction is characterized by a distance cutoff between atoms, and, in the case of hydrogen bonds and stacking interactions, proper angles between the interacting atoms. In their analysis, Allers and Shamoo did not discuss van der Waals and hydrophobic interactions. They compared the frequency of hydrogen bonds to the RNA backbone versus base atoms and found that hydrogen bonds to the backbone make up
65% of all hydrogen bonds. They emphasize the importance of hydrogen bonds between the main chain atoms of amino acids with the RNA backbone.

The study of Allers and Shamoo emphasized specific hydrogen bond interactions observed between amino acids and each base of RNA. Their approach was to catalog all observed interactions rather than indicate general propensities for interactions. In this respect, their study was similar to work later done by Frankel’s group (Cheng et al., 2003), in which the goal was to identify all theoretically possible modes of binding and then analyze complexes to determine which of the possible modes were observed.

Two studies from the same group analyzed a set of 51 protein-RNA complexes for hydrogen bonding propensities (Jeong et al., 2003, Kim et al., 2003). Hydrogen bonds were identified with the program HBPLUS (MacDonald and Thornton, 1994) and both direct and water-mediated hydrogen bonds were considered. No clear preference for binding with any particular nucleotide was observed. The amino acids with the highest binding propensities were arginine, lysine, asparagine, threonine, serine, and tyrosine. Hydrophobic amino acids had the lowest hydrogen bonding propensities. This study limited the propensity calculations to surface residues, so the calculation was not biased by including buried residues. However, only considered hydrogen bonds were considered, and most hydrophobic residues do not have hydrogen bond donors or acceptors in their side chains. Several amino acids showed preferences for specific nucleotides. For example, arginine and asparagine had higher propensities for hydrogen bonding with uracil, and lysine and threonine had a preference for adenine. Side chain atoms were found to form 71% of all hydrogen bonds, while base and backbone atoms of the RNA were each found in about half of the hydrogen bonds.

Kim et al. (2003) also analyzed hydrogen bonding propensities based on whether the nucleotides were paired or unpaired. They found that 70% of nucleotides in their dataset were involved in some type of base pair but found that unpaired nucleotides were more likely to form hydrogen bonds to amino acids. These two related studies focused solely on
hydrogen bonds, which have been shown to be relatively few in number compared to van der Waals interactions. However, the general trends for favored and disfavored amino acids are similar for both types of contacts.

Lejeune et al. analyzed interfaces in both protein-RNA and protein-DNA complexes (Lejeune et al., 2005). Here, we focus on the results of their analysis of protein-RNA interactions. They created a dataset of 49 protein-RNA complexes by extracting all protein-RNA complexes from the PDB with at least 3 Å resolution and used the PISCES server (Wang and Dunbrack, 2003) to remove sequences with greater than 30% sequence identity. RNA-binding residues were defined as any residues involved in an electrostatic interaction, hydrogen bond, van der Waals interactions, or hydrophobic interaction with bound RNA. Hydrogen bonds made up 47% of the interactions observed in this dataset. Arginine, lysine, asparagine, histidine, glutamine, aspartate, and tyrosine were all found to have favorable interaction propensities. When considering interactions with only the RNA base atoms, arginine, lysine, asparagine, histidine, and aspartate had the highest interaction propensities. Interactions with the base made up one third of all interactions observed. Favored amino acid nucleotide pairs were asparagine with uracil, aspartate with guanine, and histidine with both guanine and uracil.

Morozova et al. performed a detailed analysis of protein-RNA interactions focusing on binding pockets in a protein surrounding a base (Morozova et al., 2006). They used a non-redundant dataset of 41 protein-RNA complexes from the PDB with resolution greater than 2.8 Å. Their approach was to superimpose each occurrence of each base and analyze the binding pockets formed by proteins around the base. The analysis focused on which interactions were required for specific recognition of each base; they found that a combination of hydrogen bonds, van der Waals interactions, and stacking interactions were involved in base-specific recognition. They noted that specific recognition can be achieved
with two hydrogen bonds to the base, or with as little as one hydrogen bond and one van der Waals contact.

Kim et al. analyzed the propensities for single amino acids and pairs of amino acids to bind RNA (Kim et al., 2006). Their dataset was a non-redundant set of 86 proteins from the PDB with at least 3 nucleotides and 50 amino acids. RNA-binding residues were defined as any amino acid that lost accessible surface area on binding RNA. The novel aspect of this study was the analysis of amino acid doublets, defined as pairs of amino acids within 7 Å of each other in the protein structure. The single amino acids with the highest RNA-binding propensities were arginine, lysine, tyrosine, methionine, histidine, glycine, and phenylalanine. The doublet propensities showed that when two hydrophobic residues are paired, the doublet propensity is high, but pairs of hydrophilic residues generally showed low doublet propensities.

The use of doublet propensities was an interesting and novel approach for analyzing RNA-binding sites, but the results have proven difficult to interpret from a physico-chemical standpoint. Kim et al. (2006) note several doublets with either high or low propensities that are difficult to explain. For example, tyrosine and lysine each have favorable singlet propensities, but the doublet of tyrosine and lysine has a very low propensity for RNA-binding.

Baker and Grant performed a detailed analysis of aromatic amino acids in protein-nucleic acid complexes (Baker and Grant, 2007). Here, we focus on their results with protein-RNA complexes. They created a dataset of 61 protein-RNA complexes with no more than 90% sequence identity. They defined interactions between the aromatic amino acids phenylalanine, tyrosine, tryptophan, and histidine with any of the nucleotides by finding residues with ring centers less than 7.5 Å apart. They found that most of the aromatic interaction pairs were observed more than expected, with the only exceptions being the phenylalanine-uracil, phenylalanine-guanine, and tryptophan-uracil pairs.
Ellis et al. investigated whether interface propensities differ in different functional classes of RNA (Ellis et al., 2007). Their dataset of 89 proteins was classified into five groups based on the type of RNA bound: ribosomal, viral, messenger, transfer, or ligand. Direct hydrogen bonds and van der Waals contacts were considered, with hydrogen bonds calculated by the program HBPLUS and van der Waals contacts defined by a reduction in solvent accessibility between the free protein and the protein-RNA complex. They found that van der Waals contacts were more numerous than hydrogen bonds and contacts to the RNA backbone were more frequent than contacts to the RNA base. Side chain atoms of amino acids formed more contacts than main chain atoms. They find that arginine, lysine, histidine, tryptophan, and serine had the highest RNA-binding propensities, while glutamate and aspartate had the lowest propensities. The same trends in amino acid propensities were observed for all five functional classes of RNA.

Ellis et al. (2007) concluded that it is important to consider the functional class of the RNA in analyzing protein-RNA interfaces. In general, their data on interaction propensities showed that there are few differences in amino acid and nucleotide binding propensities among proteins that bind different types of RNA. The largest difference they observed was in the number of contacts to RNA base atoms. In the ribosomal RNA class, they found fewer base contacts than expected, while other classes showed more base contacts. They concluded that ribosomal RNA is largely double-stranded and the base atoms are therefore not available for interactions with amino acids as often as in predominantly single-stranded RNA.

Ellis and Jones performed a study of conformational changes in proteins upon binding to RNA (Ellis and Jones, 2008). They constructed a set of 12 proteins for which a high resolution structure was available for both the unbound protein and the protein-RNA complex. They found that four proteins did not undergo any conformational change upon RNA binding, while eight proteins underwent significant conformational changes. Among the latter group, four had a greater change in the RNA-binding site, while four had a greater
change in other regions of the protein. Although the dataset used was small, the results showed that proteins often undergo conformational changes upon RNA binding, but not necessarily at the RNA binding site.

Bahadur et al. analyzed the interfaces of 81 transient binary protein-RNA complexes (Bahadur et al., 2008). They created their dataset by extracting protein-RNA complexes from the PDB with 3 Å resolution or better, then removing any proteins with more than 35% sequence identity. RNA-binding residues were defined as any amino acids that lost accessible surface area in the complex with RNA. Hydrogen bonds were also considered and were identified with the program HBPLUS. Bahadur et al. found that amino acid side chain atoms formed a larger fraction of the interfaces than main chain atoms and that RNA backbone atoms form about 65% of the interface area on the RNA side. They found that adenine and uracil form a larger portion of the interface than guanine or cytosine, indicating a preference for amino acids to interact with these nucleotides. The most favored amino acid in protein-RNA interfaces was arginine. Interestingly, they found that lysine was not overrepresented in interfaces compared to the protein surface as a whole. The aromatic amino acids were also found to be overrepresented in protein-RNA interfaces. The amino acid main chain atoms are involved in 26% of all hydrogen bonds, while the side chain atoms make up the remaining 74%. Arginine and lysine side chains account for 34% of all hydrogen bonds observed in the dataset. On the RNA side, the phosphate, ribose, and base atoms contribute almost equally to the number of hydrogen bonds, with 36% involving the phosphate atoms, 33% ribose atoms, and 31% base atoms. Bahadur et al. also considered water-mediated hydrogen bonds and found that 69% of water-mediated hydrogen bonds involve either the RNA phosphate or ribose atoms. Water-mediated hydrogen bonds were most frequently seen with main chain atoms of amino acids, which accounted for 38% of observed bonds.
Conclusions on computational analyses of protein-RNA interfaces

From this summary of computational analyses of protein-RNA interfaces, several themes emerge. First, every study identified the importance of positively charged residues arginine and lysine in binding RNA. There is also general agreement that asparagine, histidine, and tyrosine are favored RNA-binding residues. Most studies agree that van der Waals contacts are more numerous than hydrogen bonds, that amino acid side chains make more contacts with RNA than side chains, and that the majority of the contacts are with the RNA backbone rather than the bases.

There are several issues to note about these studies. Comparisons between the studies are difficult because each used a different dataset and different methods of defining RNA-binding residues. Early studies using small datasets produced a number of results that have been seen repeatedly in later studies, such as the prevalence of arginine and lysine in the interface, but other observations have not been borne out, such as the high propensity for isoleucine observed by Jones et al. in 2001 (Jones et al., 2001). Considering that the number of experimentally determined protein-RNA complexes is still small, current studies using the largest available datasets (such as the one presented in Chapter 5 of this dissertation) may still not produce an accurate picture of protein-RNA interfaces. While many observations and conclusions from the current studies may generalize well to protein-RNA complexes yet to be determined, the molecular details of the interaction data are likely to change.

Databases of protein-RNA interactions

Specialized databases of protein-RNA interactions are scarce. Several of the studies summarized above created databases containing interaction information, but few of them are publicly available, and those that are available have not been maintained. For example, Allers and Shamoo created a database in their 2001 work (Allers and Shamoo, 2001) but it
has not been updated since and still contains only 42 protein-RNA complexes. Here, we present a summary of the available databases of protein-RNA interactions.

The Amino Acid Nucleotide Interaction Database (AANT) contains information on both protein-DNA and protein-RNA interactions (Hoffman et al., 2004). The database is organized by type of interaction, allowing the user to specify the amino acid and nucleotide as well as whether the interaction involves the main chain or side chain of the amino acid, and the base, sugar, or phosphate of the nucleotide. AANT then displays either a table with counts of interactions, or an interactive display of all interactions of this type superimposed. Alternatively, a user can select an individual complex and view a list of all interactions observed in the structure.

The main drawback of AANT is that it contains only hydrogen bond interactions. Most studies have shown that hydrogen bonds are far less common in protein-RNA interfaces than van der Waals contacts. Also, in the tables summarizing the number of times a particular interaction was observed are for all complexes in the database; there is no option to specify only protein-DNA or only protein-RNA complexes. A final note is that the AANT database has not been updated since 2006.

NPIDB, a database of nucleic acid protein interactions, also contains information on both protein-DNA and protein-RNA interfaces (Spirin et al., 2007). The database is organized as a collection of PDB format files along with various Perl scripts for extracting and displaying interaction data. The interaction data are gathered by programs that identify either hydrogen bonds or hydrophobic interactions from the PDB files describing the complex. One major advantage of NPIDB is that it is automatically updated weekly. The main drawback of NPIDB is that the output is limited to interactions within a single complex. In order to extract information about many protein-nucleic acid complexes, the user has to manually go to the entry for each complex and compile the combined results.
OVERALL GOALS

The overall goal of this research is to characterize protein-RNA interaction sites and identify features that can be used to accurately predict RNA-binding sites in proteins. To achieve this goal, we have accomplished the following specific aims:

1. Predict RNA-binding sites in proteins using a sequence-based classifier and demonstrate the application of the prediction method on telomerase reverse transcriptase (Chapter 2).

2. Develop RNABindR, a web server for calculating RNA-binding residues in proteins given a protein-RNA complex from the PDB and predict RNA-binding residues given an amino acid sequence (Chapter 3).

3. Improve prediction of RNA-binding sites in proteins using information derived from the protein structure and sequence conservation (Chapter 4).

4. Develop the Protein-RNA Interaction Database (PRIDB), a comprehensive database of protein-RNA complexes, and analyze the interaction propensities for amino acids and nucleotides in protein-RNA complexes (Chapter 5).

We also propose to explore related problems, especially prediction of DNA-binding sites in proteins, and application of several different prediction methods to clinically important proteins, in studies described in the Appendices:

5. Predict DNA binding sites in proteins using protein sequence and sequence entropy (Appendix A).

DISSERTATION ORGANIZATION

The dissertation has six chapters and two appendices.

**Chapter 1** is a general introduction to protein-RNA interactions and describes the currently available methods for predicting RNA binding sites in proteins. A review of computational analyses of protein-RNA interactions is also presented.

**Chapter 2** is a paper published in the journal *RNA* in 2006 (Terribilini et al., 2006), in which a sequence-based classifier for predicting RNA-binding sites in proteins is described. The Naïve Bayes classifier used the local sequence information for each residue to predict whether it is likely to bind RNA or not. RNA-binding residues were predicted at 85% overall accuracy and the predictions made for the human telomerase reverse transcriptase protein were shown to be in good agreement with the available experimental data. I conceived of the experiment, created the dataset of protein-RNA interactions, carried out all computations, wrote the first draft of the paper and participated in revisions and editing. Jae-Hyung Lee and Changhui Yan contributed to discussions. Robert Jernigan contributed to discussions and manuscript reviews. Vasant Honavar and Drena Dobbs contributed to experimental design, discussions, and manuscript preparation.

**Chapter 3** is a paper published in the journal *Nucleic Acids Research* in 2007 (Terribilini et al., 2007). It describes RNABindR, a server for analyzing and predicting RNA-binding sites in proteins. Given a protein-RNA complex from the PDB, RNABindR calculated the RNA-binding residues based on a user-defined distance cutoff and displays the interactions using a Jmol applet. Given a protein sequence, RNABindR predicts the RNA-binding residues in the sequence. I developed the classification method, implemented the web server, wrote the first draft of the paper, and participated in revisions and editing. Jeffry Sander, Jae-Hyung Lee, and Peter Zaback contributed to discussions and tested the web
server functionality. Robert Jernigan and Vasant Honavar contributed to discussions and participated in manuscript reviews. Drena Dobbs contributed to the web server design and participated in manuscript preparation.

Chapter 4 describes several improved classifiers for predicting RNA-binding sites in proteins. We developed classifiers that use protein sequence or structure neighbors as input, as well as classifiers based on the position-specific scoring matrix (PSSM) derived from a multiple sequence alignment of related protein sequences. These classifiers achieved different levels of performance, raising the possibility that combining the output from them could lead to increased classification performance. The ensemble classifier developed with this approach in mind gave the best prediction performance. We analyzed the effect, if any, of increased dataset size on prediction performance. We analyzed the properties of PSSMs that lead to improved predictions and compared prediction performance for amino acids that contact only the RNA base versus those that contact the RNA backbone. I created the datasets used in the study, designed the input information for the classifiers, performed the analyses, carried out the computations, and drafted the manuscript. Jeffry Sander contributed to discussions and performed the initial ensemble classifications. Cornelia Caragea contributed to experimental design, carried out some of the computations, and contributed to manuscript preparation. Vasant Honavar and Drena Dobbs contributed to experimental design, discussions, and manuscript preparation.

Chapter 5 describes the Protein-RNA Interaction Database (PRIDB). PRIDB is a comprehensive collection of all protein-RNA complexes in the PDB. From PRIDB, we have created a non-redundant dataset of 181 proteins and carried out an analysis of the protein-RNA interface properties. This dataset is about twice as large as any previously studied in this manner. We analyzed the interaction propensities between amino acids and nucleotides in direct hydrogen bonds, water-mediated hydrogen bonds, and van der Waals interactions. We found few differences in propensities for interaction with Watson-Crick paired
nucleotides versus non-Watson-Crick paired nucleotides, with single-stranded nucleotides showing more differences in interaction propensities. We found that many amino acids have distinct preferences for binding with either main chain atoms or side chain atoms. We also found that some amino acids have preferences for binding to RNA base atoms versus RNA phosphate or ribose atoms. I created the database and non-redundant dataset, performed the computational analyses, and prepared the manuscript. Jeff Ferguson performed the computation of the hydrogen bond interactions. Drena Dobbs contributed to discussions.

Chapter 6 summarizes the general conclusions of this dissertation study, the contributions and impact of this work, and describes future directions.

In the Appendices, two studies related to the goals of this dissertation are described.

Appendix A is a paper published in the journal BMC Bioinformatics (Yan et al., 2006). It describes a machine learning classifier for predicting DNA-binding sites in proteins based on the identities of the target residue and the surrounding protein sequence. Prediction performance was improved by adding the sequence entropy and solvent accessibility of the target residue as input. Changui Yan carried out the computations and drafted the manuscript. I contributed to discussions, requested data and composed results from Sarai’s group, and participated in manuscript reviews. Feihong Wu contributed to discussions. Robert Jernigan contributed to discussions and manuscript reviews. Drena Dobbs and Vasant Honavar contributed to experimental design, discussions, and manuscript preparation.

Appendix B is a paper published by the Pacific Symposium on Biocomputing, PSB 2006 (Terribilini et al., 2006). It describes the application of protein-RNA and protein-protein binding site prediction methods to the HIV-1 and EIAV Rev proteins, which are essential for viral regulation and replication. The predicted binding sites were shown to be in good agreement with the available experimental data. I performed the RNA-binding site predictions, and participated in manuscript preparation. Jae-Hyung Lee experimentally mapped the EIAV Rev RNA-binding sites and contributed to manuscript preparation.
Changhui Yan performed the protein-protein binding site predictions. Robert Jernigan and Susan Carpenter contributed to discussions and manuscript reviews. Vasant Honavar and Drena Dobbs contributed to experimental design and manuscript preparation.

REFERENCES


Chen, Y. C., and C. Lim (2008), Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry, *Nucleic acids research.*


Terribilini, M., J. D. Sander, J. Lee, P. Zaback, R. L. Jernigan, V. Honavar, and D.
Dobbs (2007), RNABindR: a server for analyzing and predicting RNA-binding sites in

Thompson, J. D., D. G. Higgins, and T. J. Gibson (1994), CLUSTAL W: improving
the sensitivity of progressive multiple sequence alignment through sequence weighting,
position-specific gap penalties and weight matrix choice, *Nucleic Acids Research*, 22(22),
4673-80.

Tong, J., P. Jiang, and Z. Lu (2008), RISP: A web-based server for prediction of
RNA-binding sites in proteins, *Computer methods and programs in biomedicine*.

Towfic, F., C. Caragea, D. Gemperline, D. Dobbs, and V. Honavar (2008), Struct-NB:
Predicting Protein-RNA Binding Sites Using Structural Features, *International Journal of
Data Mining and Bioinformatics*. *In press*.

Treger, M., and E. Westhof (2001), Statistical analysis of atomic contacts at RNA-

Wang, G., and R. L. Dunbrack (2003), PISCES: a protein sequence culling server,

of DNA and RNA binding sites in amino acid sequences, *Nucleic acids research*, 34(Web
Server issue), W243-8.


CHAPTER 2. PREDICTION OF RNA BINDING SITES IN PROTEINS FROM AMINO ACID SEQUENCE

A paper published in *RNA*

Michael Terribilini, Jae-Hyung Lee, Changhui Yan, Robert L. Jernigan, Vasant Honavar, and Drena Dobbs

**ABSTRACT**

RNA–protein interactions are vitally important in a wide range of biological processes, including regulation of gene expression, protein synthesis, and replication and assembly of many viruses. We have developed a computational tool for predicting which amino acids of an RNA binding protein participate in RNA–protein interactions, using only the protein sequence as input. RNABindR was developed using machine learning on a validated non-redundant data set of interfaces from known RNA–protein complexes in the Protein Data Bank. It generates a classifier that captures primary sequence signals sufficient for predicting which amino acids in a given protein are located in the RNA–protein interface. In leave-one-out cross-validation experiments, RNABindR identifies interface residues with >85% overall accuracy. It can be calibrated by the user to obtain either high specificity or high sensitivity for interface residues. RNABindR, implementing a Naive Bayes classifier, performs as well as a more complex neural network classifier (to our knowledge, the only previously published sequence-based method for RNA binding site prediction) and offers the advantages of speed, simplicity and interpretability of results. RNABindR predictions on the human telomerase protein hTERT are in good agreement with experimental data. The availability of computational tools for predicting which residues in an RNA binding protein are likely to contact RNA should facilitate design of experiments to directly test RNA binding function and contribute to our understanding of the diversity, mechanisms, and
regulation of RNA–protein complexes in biological systems. (RNABindR is available as a Web tool from http://bindr.gdcb.iastate.edu.)

**INTRODUCTION**

Understanding the molecular mechanisms by which proteins recognize and discriminate between specific RNA molecules is critical for comprehending the functional implications of these interactions in cells. RNA-protein interactions, in addition to their importance in protein synthesis, mRNA processing and viral replication, have recently been shown to play critical roles in cellular defense and developmental regulation (Hall, 2002; Tian et al., 2004), underscoring the importance of understanding the molecular determinants of RNA-protein interactions.

At least 9 families of RNA binding proteins have been identified using sequence-based analyses of RNA binding proteins, together with functional characterization of mutations that affect the specificity or affinity of RNA binding (reviewed in Chen & Varani, 2005). In contrast, the number of experimentally determined structures for RNA-protein complexes is still relatively small and heavily biased (ribosomal proteins represent ~ 50% of all RNA binding proteins in the PDB). Nevertheless, several computational analyses of RNA-protein complexes have generated databases of RNA-protein contacts and provided valuable insights into the biophysical basis of interaction patterns between ribonucleotides and amino acids (Cusack, 1999; Draper, 1999; Jones et al., 2001; Kim et al., 2003; Hoffman et al., 2004; Jeong et al., 2004; Jeong & Miyano, 2006).

Because of the importance of RNA-protein interactions in biological regulation and the considerable effort required to identify RNA binding residues through biophysical analyses of RNA-protein complexes or in vitro binding studies, there is an urgent need for computational methods to identify RNA binding sites based on primary amino acid sequence alone. Machine learning techniques offer an attractive approach to construction of classifiers...
for this task, using datasets of experimentally well-characterized RNA-protein complexes. Three recent studies have reported the use of support vector machines (SVMs) to identify RNA binding proteins and assign them to functional classes (e.g., rRNA binding, mRNA binding, tRNA binding, viral RNA binding, etc.) using only the amino acid sequence (Han et al., 2004), a combination of sequence and pseudo-amino acid composition as input (Cai & Lin, 2003), or a variety of sequence based information including predicted solvent accessibility and predicted secondary structure (Yu et al., 2006). Our previous work (Yan et al., 2004a; Yan et al., 2004b; Yan et al., In press) has demonstrated the feasibility of constructing classifiers for protein-protein and protein-DNA binding site identification using machine learning approaches. However, there has been little work using machine learning approaches to construct classifiers for identifying RNA binding sites from primary amino acid sequence.

In this paper, we present RNABindR, a fast and simple tool for predicting RNA binding sites. In its current implementation, RNABindR requires only protein sequence information as input; no information regarding the structure of the protein or the sequence or structure of the RNA is required. Although inclusion of structure-derived information, when available, can improve predictions, we focus here on sequence-based prediction to provide a broadly applicable tool. To demonstrate the utility of RNABindR, we make predictions on the telomerase protein TERT, for which the structure of the protein-RNA complex has not been determined. The predictions are in good agreement with the experimentally characterized RNA binding regions of TERT.

The only previously published sequence-based method for predicting interface residues, to our knowledge, is a neural network classifier reported by Miyano’s group (Jeong et al.; 2004, Jeong & Miyano, 2006). The results of our experiments demonstrate that the performance of RNABindR, using a Naive Bayes classifier trained and tested on the same dataset, is comparable to that of the neural network classifier. Unlike the neural network
classifier, which requires multiple passes through the data during training, the Naive Bayes classifier requires only one pass through the training data, is easily updateable, and is rather straightforward to interpret.

RESULTS

Sequence characteristics of RNA binding sites

Arginine-rich motifs (Weiss, 1998) are abundant in RNA binding sites and other strong biases in the types of amino acids present in RNA-protein interfaces have been reported in several previous studies (Lustig et al., 1997; Jones et al., 2001; Kim et al., 2003; Jeong et al., 2004; Jeong & Miyano, 2006). To evaluate whether these primary sequence biases can be effectively exploited in a machine learning approach to identify amino acid sequence correlates of RNA binding sites, we generated a non-redundant dataset of 109 RNA binding proteins (see Materials and Methods) to estimate the interface propensity for each amino acid type as follows:

\[ \text{Interface Propensity}(x) = \log_2 \left( \frac{\text{percentage of residues of type } x \text{ in the interfaces}}{\text{percentage of residues of type } x \text{ in the entire dataset}} \right) \]

An interface propensity value greater than 0 indicates that an amino acid is overrepresented in RNA-protein interfaces relative to the protein sequence as a whole. Figure 2.1 shows the interface propensity (solid bars) for each of the 20 amino acids, as well as the frequency with which that amino acid occurs in each of the two positions immediately flanking a known interface residue (cross-hatched bars). Interface propensities, estimated from a smaller dataset of 55 ribosomal protein chains (data not shown) did not differ significantly from those estimated using the larger dataset of 109 RNA binding proteins, and
our results using both datasets are consistent with previously published data (Jones et al., 2001).

Figure 2.1 Certain amino acids are highly favored in RNA–protein interfaces. Interface propensities for the indicated amino acids are shown as solid bars; the hatched bars to the left and right of the solid bar are the propensities for the amino acid to occur in the position immediately before or after an interface residue, respectively. The residues are placed in the order of increasing hydrophobicity based on the (Kyte and Doolittle 1982) hydropathy index.

As expected, the positively charged amino acids arginine and lysine show the highest interface propensities, 1.29 and 1.17, respectively, consistent with their ability to participate in interactions both with bases and with the negatively charged phosphate backbone of RNA. Together arginine and lysine account for 32% of the interface residues in our dataset. While this is a significant fraction of interface residues, it also shows that one cannot focus solely
on positively charged amino acids to discover how RNA-protein recognition occurs. Another favored residue, histidine (0.60), also can be positively charged and can participate in stacking interactions with RNA bases through its imidazole ring. Tryptophan and tyrosine are slightly preferred, with propensities of 0.21 and 0.18 respectively. In contrast, phenylalanine (-0.60) and negatively charged amino acids glutamate (-1.13) and aspartate (-0.62) are significantly under-represented in interfaces, as are hydrophobic residues such as leucine, isoleucine, valine, and alanine (all below -0.84). Importantly, there are significant biases in the types of amino acids that tend to be "sequence neighbors" of interface residues. For instance, glycine is highly preferred on either side of an interface residue (0.50 and 0.47); its small size may enhance flexibility, allowing protein domains to adopt conformations that facilitate RNA binding.

If the biases in amino acid propensities noted above are frequently accompanied by clustering of interface residues within the primary sequence of an RNA binding protein, a machine learning algorithm should be able to "learn" sequence composition characteristics or other signals in the neighborhood surrounding interface residues, based on a validated training dataset, and generate a classifier for predicting likely interface residues in test sequences. The tendency of protein-protein interface residues to be clustered along the primary sequence of proteins has been noted previously (Jones et al., 2001; Ofran & Rost, 2003; Yan et al., 2004b). We examined the tendency of RNA-protein interface residues to be similarly clustered in our dataset of RNA binding proteins by calculating the log-likelihood that a residue is an interface residue, given that it is at a certain distance from another interface residue (Figure 2.2). The log-likelihood is given by \( \log_2(P_{\text{observed}}/P_{\text{background}}) \) where \( P_{\text{observed}} \) is the observed probability that a given neighbor of an interface residue is also an interface residue and \( P_{\text{background}} \) is the probability that the position is an interface residue by chance (~0.14 for our dataset, because ~14% of the residues in our dataset are interface residues).
Figure 2.2 RNA binding residues tend to occur in clusters within primary sequence. The log likelihood that a position neighboring an interface residue also contains an interface residue based on the nonredundant data set of 109 RNA binding proteins. The hatched portion of the bars represents the log likelihood for the entire data set of 109 proteins. The solid portion of the bars represents the log likelihood for the ribosomal protein subset of 55 proteins. Likelihood values >0 mean that the position has higher probability than random of also being an interface residue.

This analysis revealed that 95% of interface residues in the dataset of 109 RNA binding proteins have at least 1 additional interface residue among the 4 amino acids on either side, and 49% have at least 4. The tendency of interface residues to be clustered within the primary sequence is more pronounced in the subset of 55 ribosomal proteins: 97% of interface residues in the ribosomal dataset have at least one additional interface residue within 4 amino acids on either side and 63% have at least 4 neighboring interface residues. For the dataset of 54 non-ribosomal proteins, the corresponding values are 90% and 23%
respectively. Thus, this tendency of interface residues to cluster in primary sequence, together with the distinct interface propensities of individual amino acids, suggests that it should be possible to capture functionally relevant sequence signals in the neighborhood of interface residues and to exploit these using a machine learning approach to predict RNA binding sites in proteins.

**Using a Naive Bayes classifier, RNABindR, reliably predicts RNA-protein interface residues using only amino acid sequence information**

The performance of RNABindR, using a Naive Bayes classifier, was evaluated in leave-one-out cross-validation experiments as described in Materials and Methods. Table 2.1 summarizes an example of the results obtained using four different input window sizes and a threshold, \( \theta \), which was empirically determined to provide an optimal correlation coefficient on the training set. Using an input window of 25 amino acids, the classifier achieved an overall accuracy of 85% with a correlation coefficient of 0.35, specificity+ of 0.51 and sensitivity+ of 0.38 (see Materials and Methods for definitions). Adding information such as secondary structure, relative accessible surface area, sequence entropy, hydrophobicity, and electrostatic potential to the amino acid sequence inputs did not improve RNABindR performance. Performance on the ribosomal subset was better than the average performance over the entire dataset (data not shown). However, performance on the ribosomal subset was the same whether the training set used was the ribosomal subset or the entire dataset.
Table 2.2 Interface residue prediction performance of RNABindR. Examples of average results for 109 leave-one-out experiments using different input window sizes and optimizing the threshold, $\theta$, to maximize the correlation coefficient (CC) on the training set. The best performance, based on estimated CC, was obtained using an input window size of 25 and $\theta = 0.5$.

<table>
<thead>
<tr>
<th>Window Size</th>
<th>Accuracy (%)</th>
<th>CC</th>
<th>Specificity+ (%)</th>
<th>Sensitivity+ (%)</th>
<th>Specificity- (%)</th>
<th>Sensitivity- (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>5</td>
<td>80.7</td>
<td>0.26</td>
<td>37</td>
<td>37</td>
<td>89</td>
<td>88</td>
</tr>
<tr>
<td>15</td>
<td>85.6</td>
<td>0.33</td>
<td>55</td>
<td>37</td>
<td>88</td>
<td>91</td>
</tr>
<tr>
<td>25</td>
<td>84.8</td>
<td>0.35</td>
<td>51</td>
<td>38</td>
<td>89</td>
<td>93</td>
</tr>
<tr>
<td>27</td>
<td>84.5</td>
<td>0.33</td>
<td>46</td>
<td>37</td>
<td>90</td>
<td>93</td>
</tr>
</tbody>
</table>

In specific biological applications, such as identifying critical residues for site-specific mutagenesis, it may be more important to predict interface residues with high specificity (i.e., to produce a smaller number of "positive" interface residue predictions with high confidence) than to obtain a high correlation coefficient. We report results obtained with classifiers trained to obtain an optimal correlation coefficient (CC) because CC is a more meaningful measure than specificity or sensitivity for comparing different classifiers (see Materials and Methods, Baldi 2000). With a Naive Bayes classifier, it is straightforward to vary the threshold $\theta$ to increase specificity+ at the expense of a decrease in sensitivity+. This is illustrated in Figure 2.3, which shows a ROC (receiver operating characteristic) plot of sensitivity+ against false positive rate, defined as (1-specificity-). At the expense of lower sensitivity, a very low false positive rate can be achieved.
Figure 2.3 Receiver operating characteristic (ROC) curve for RNABindR predictions. The ROC curve illustrates how varying the cutoff threshold \(\theta\) determines the trade-off between sensitivity+ and false positive rate \((1 - \text{specificity}^-)\), where specificity– is defined as \(\text{FP}/(\text{FP} + \text{TN})\). Results shown are for an input window of 25 amino acids.

While these statistics allow evaluation of the performance of RNABindR in identifying RNA-protein interface amino acids on a per residue basis, an important criterion for evaluating its utility in practice is whether it correctly identifies a significant fraction of the total interface residues in individual RNA binding proteins. For the complete dataset, RNABindR effectively recognized binding sites in 59% of proteins by correctly identifying at least 20% of the interface residues (see below, Figure 2.5A).

**Evaluating RNABindR predictions in the context of 3-dimensional structures**

In developing RNABindR, we have not taken advantage of available structural information regarding the target protein or its cognate RNA because it is much more
common to have the sequence of a protein without a structure. Nevertheless, it is informative
to evaluate RNABindR results by visualizing them in the context of 3-dimensional structures
of known RNA-protein complexes. Figure 2.4 shows examples in which RNABindR was
tested on one protein chosen from each of the four different categories of complexes in the
complete dataset (see Table 2.2): i) rRNA; ii) mRNA, snRNA, dsRNA, siRNA; iii) tRNA;
iv) viral RNA. For each protein, the predicted versus actual interface residues, shown in red,
are mapped onto surface plots of PDB structures (compare left and middle panels). In the
panels on far right, a different coloring scheme is used to illustrate the performance of
RNABindR on individual residues in each protein (see below).
Figure 2.4 Predictions mapped onto three-dimensional structures of RNA binding proteins. Examples of RNABindR results for four different types of RNA–protein complexes are shown: (A) ribosomal protein L15, PDB 1JJ2:K (Klein et al. 2001); (B) Xenopus dsRNA binding protein, PDB
1DI2:A (Ryter and Schultz 1998); (C) Ebola virus Vp40, PDB 1H2C:A (Gomis-Ruth et al. 2003); (D) tRNA pseudouridine synthase, PDB 1R3E:A (Pan et al. 2003). Predicted RNA binding sites, with predicted interface residues shown in red and predicted noninterface residues in gray (left panels). Actual RNA binding sites, with actual interface residues in red and actual noninterface residues in gray (middle panels). The performance of RNABindR for individual residues, with true positives (TPs) shown in red, false positives (FPs) in blue, false negatives (FNs) in yellow, and true negatives (TNs) in gray (right panels). Thus, in this representation, red + yellow residues correspond to the actual interface (derived from the PDB structure), red + gray residues correspond to correctly predicted residues (both interface and noninterface), and blue + yellow residues correspond to misclassified residues. Results shown were predicted with RNABindR using an input window of 25 amino acids and $\theta = 0.5$. All structure diagrams were generated using PyMol (http://www.pymol.org).

Results obtained for ribosomal protein L15 (PDB 1JJ2:K), a structural component of the large ribosomal subunit from the archaebacterium H. marismortui (Klein et al., 2001) are shown in Figure 2.4a. This was the "best" prediction (ranked #1 out of 109) based on correlation coefficient (0.63). For clarity and because of its large size, the RNA partner is not included in this example. In L15, one of the two RNA binding sites was detected with very high specificity (Figure 2.4a, compare red residues representing the predicted interface in left panel with the actual interface in middle panel). In the right-most panel, interface residues of L15 that were correctly identified as such (true positives, TPs) are shown in red: 40 out of 42 predicted interface residues are, in fact, interface residues (specificity+ = 95%). There were only 2 false positive (FP) predictions, shown in blue. True negatives (TNs), in gray, and false negatives (FNs), in yellow, are also shown. Note that although the specificity for interface residues in this example is high (95%), the accuracy is relatively low (80%) compared with the average over the complete dataset (85%), largely due to failure of RNABindR to detect any interface residues in one of two RNA binding sites on the L15 protein. As described below, sensitivity (for the training dataset) can be enhanced by choosing a lower value for $\theta$. 
In the case of L15, this results in better coverage (i.e., higher sensitivity), allowing the second RNA binding domain to be detected, but at the loss of specificity (data not shown).

Results of similar analyses for a protein from each of the other three classes of RNA-protein complexes are shown in Figures 2.4b, 2.4c and 2.4d. Figure 2.4b shows results for the double-stranded RNA binding motif (dsRBM) domain of the Xenopus dsRNA binding protein A bound to RNA (in green wire frame). The prediction for this protein ranked 23rd (CC=0.38) with an overall accuracy of 83%. A simple search for RNA binding motifs on this protein reveals that the entire 69 amino acid sequence included in the crystal structure is the canonical dsRBM. However, there are only 13 actual interface residues within this motif, all clustered on one face of the protein shown. RNABindR correctly identified 5 of these 13 interface residues. Figure 2.5 illustrates how lowering the threshold $\theta$ significantly improves identification of the interface residue class. The interface residue predictions for the dsRNA binding protein shown in Figure 2.4 are shown for three different values of $\theta$. In Figure 2.5A when $\theta$ is relatively high, a small number of interface residues are predicted with high specificity. Figure 2.5B shows the predictions using the value of $\theta$ obtained by optimizing RNABindR on the training set. In Figure 2.5C when $\theta$ is low, many more interface residues are predicted but we sacrifice specificity to do so.
Figure 2.5 RNABindR sensitivity and specificity trade-off. Changing the value of the threshold parameter \( \theta \) causes a trade-off between specificity and sensitivity in predicting RNA binding residues. The example shown here is the double-stranded RNA binding protein from Xenopus, PDB ID 1DI2:A, also shown in Figure 2.4B. The color scheme in this figure is the same as in Figure 2.4.

The Ebola virus matrix protein, Vp40, bound to a 3 nt RNA ligand (accuracy 95%) is shown in Figure 2.4c, and a tRNA pseudouridine synthase bound to a tRNA ligand (51 nt) is shown in Figure 2.4d. These predictions were ranked 19th (CC=0.42) and 34th (CC=0.29) out of 109, respectively. Performance statistics provided in the figure captions illustrate that the specificity and sensitivity for non-interface residues are much higher than for interface residues in both cases.

**Comparison of RNABindR predictions with mapped RNA binding sites in the telomerase protein, TERT**

The primary motivation for developing RNABindR (which does not require structural information) was to provide a tool for identifying potential RNA binding sites in proteins when information regarding the RNA-protein complex or its interface is not available. To demonstrate the utility of RNABindR in such cases, we have applied it to the prediction of RNA binding residues in the human telomerase protein hTERT. Telomerase is the ribonucleoprotein complex responsible for maintaining telomere length by adding short
repeated sequences to the ends of chromosomes (recently reviewed in Blackburn 2005; Autexier & Lue, 2006). TERT is the reverse transcriptase component of telomerase and binds to the essential telomerase RNA subunit (TR), which serves as the template for synthesis of telomeric DNA repeats. The C-terminal half of hTERT contains the reverse transcriptase domain (RT) and two RNA interaction domains (RIDs) have been mapped to the N-terminal half of the protein (Lai et al., 2001; Bachand & Autexier, 2001; Moriarty et al., 2002; Moriarty et al., 2005). RID2 is a relatively high affinity RNA binding domain and RID1 is a lower affinity RNA binding domain (reviewed in Autexier & Lue, 2006). RID1 and RID2 each contain several elements that are conserved at the primary sequence level and, in some cases, have been shown to be important for RNA binding based on mutagenesis and in vitro binding experiments (Lai et al., 2001; Bachand & Autexier, 2001; Moriarty et al., 2002; Moriarty et al., 2005).

Figure 2.6A shows the RNA interface residues predicted by RNABindR mapped onto functional domains of hTERT defined by in vitro catalytic activity and/or RNA binding assays (Lai et al., 2001; Bachand & Autexier, 2001; Moriarty et al., 2002; Moriarty et al., 2005). The prediction that most residues involved in hTERT RNA binding lie outside the RT domain is in agreement with experimental results that have demonstrated that the RT and RNA binding domains of hTERT are separable (Lai et al., 2001; Moriarty et al., 2004). Most clusters of predicted RNA binding residues are located within the experimentally mapped RNA binding domains, RID1 and RID2, or correspond to arginine-rich portions of the variable "linker" region between them, which has been shown to contribute to hTERT RNA binding in vitro (Moriarty et al., 2002).
Figure 2.6 RNABindR predictions on telomerase reverse transcriptase (TERT). Mapped functional domains and conserved motifs of TERT are shown at the top. Shaded boxes on lines labeled "Predictions" show clusters of predicted RNA interface residues. (A) Human telomerase reverse transcriptase (hTERT). Boundaries of two major RNA interaction domains (RIDs) indicated by open boxes (Moriarty et al. 2005Go). The amino acid sequence that includes one of the clusters of predicted RNA-interface residues, located in RID2, is shown at the bottom. Two boxed regions, amino acids 481–490 and amino acids 508–517, correspond to deletion mutations that have been shown to decrease hTERT RNA binding activity by 60% and 70%, respectively (Moriarty et al. 2002). Individual interface residues predicted by RNABindR are indicated by + below the sequence. (B) Tetrahymena thermophila telomerase reverse transcriptase (tTERT). The two RNA binding domains are indicated by open boxes.
The amino acid sequence of the C-terminal end of the TEN RNA binding domain is shown, with individual interface residues predicted by RNABindR indicated by + below the sequence. Removing residues 1–12 and 182–191 (boxed in the sequence view) abolished RNA binding of the TEN domain construct (Jacobs et al. 2005Go, 2006). RNABindR predicts a cluster of interface residues in residues 182–191, but no interface residues are predicted in residues 1–12. (N) N terminus, (TEN) telomerase essential N-terminal domain, (GQ, CP, QFP, and T) conserved sequence motifs, (RT) reverse transcriptase domain.

The amino acid sequence of a conserved portion of RID2 containing a cluster of predicted RNA binding residues is shown in the lower portion of Figure 2.6A. This predicted cluster lies within the "QFP" motif in RID2 and encompasses amino acids whose deletion results in a 60% reduction in RNA binding (aa 481-490, in box) (Moriarty et al., 2002). Another cluster of interface residues within RID2 (but outside the region whose sequence is shown) also overlaps with sequences within the "T" motif required for full RNA binding activity based on deletion studies (Lai et al, 2001). Three clusters of predicted interface residues lie within or overlap the boundaries of RID1, which appears to comprise a lower affinity binding domain that contributes to, but is not absolutely required for, RNA binding (Moriarty et al., 2002). An example of a case in which RNABindR does not predict interface residues corresponding to amino acids whose deletion results in reduced RNA binding activity is also shown in the lower portion of Figure 2.6A (aa 508-517, in box). It is important to note, however, that loss of RNA binding activity in these experiments could be due either to deletion of residues that directly contact RNA or to loss of binding due to an indirect effect on the overall structure or stability of hTERT. Moreover, experimental data that provide evidence for or against the role of specific amino acids in the hTERT-TR interaction are not available for most residues within the mapped RNA binding domains. Overall, the RNABindR predictions are in very good agreement with currently available
Experimental data and identify several additional amino acids that could potentially contribute to hTERT RNA binding activity.

Tetrahymena TERT also has two RNA-binding domains. The higher affinity domain, residues 195-516, is essential for telomerase RNA binding (Lai et al., 2001) and mutagenesis experiments have demonstrated that specific residues within the CP and T motifs are important for RNA-binding (Bryan et al., 2000; Lai et al., 2002). Figure 2.6B shows RNABindR predictions mapped onto the functional regions of Tetrahymena TERT. RNABindR predicts two clusters of interface residues, one near the T motif, but none in the CP motif. A lower affinity RNA binding domain, referred to as the TEN domain, contributes to telomerase RNA binding (O’Connor et al., 2005). Residues 1-12 and 182-191 within this domain are especially important for RNA binding; they are susceptible to digestion by Lys-C in the absence of RNA, and protected in the presence of RNA (Jacobs et al., 2005; Jacobs et al., 2006). Also, the deletion of these two short segments abolishes RNA-binding (Jacobs et al., 2006). Notably, the only interface residues predicted by RNABindR in this domain are a cluster from 185-191. Thus, RNABindR predictions for Tetrahymena TERT agree well with the available experimental data.

The performance of RNABindR, implementing a Naive Bayes classifier, is comparable to that of a more complex Neural Network classifier.

To our knowledge, there is only one other published successful application of a machine learning approach to sequence-based prediction of interface residues in RNA-protein complexes. Using a dataset of 96 chains from RNA-protein complexes and a total of 4782 interface residues, Miyano's group (Jeong et al., 2004) used a neural network to predict interface residues in RNA binding proteins. Miyano’s group reported a CC=0.59 obtained using filtering and state-shifting. Both filtering and state-shifting take advantage of the fact that interface residues are clustered along the primary sequence. Filtering removes incorrect
interface predictions that are isolated, i.e., if a residue is predicted to be an interface residue, but no neighboring residues are predicted as interface residues, the prediction is changed to non-interface. State-shifting corrects predictions for residues that were misclassified as non-interface residues by changing the prediction to interface if a neighboring residue is predicted to be an interface residue. Both filtering and state-shifting use information that is generally unavailable to the classifier, i.e., there is no a priori way to determine the false positive and false negative predictions in a test sequence that is not part of the training set. Hence, we do not attempt a comparison of results obtained by filtering and state-shifting with our results. To facilitate direct comparison of RNABindR with the published neural network classifier, we trained and tested RNABindR using a Naive Bayes classifier on the same dataset used in Miyano’s study. Table 2.3 shows the best results reported by Miyano’s group (Jeong et al., 2004) using a neural network, without filtering and state-shifting, compared with the best results (in terms of correlation coefficient) obtained using RNABindR. Notably, the overall results are comparable, but the Naive Bayes method is considerably faster and easier to implement.

Table 2.3 Comparison of RNABindR (Naive Bayes classifier) with a neural network classifier.
Direct comparison of RNABindR Naïve Bayes classifier with the neural network method of Miyano, trained and tested on the Miyano data set (Jeong et al. 2004). The data presented here represent the average performance of the methods on the Miyano data set.

<table>
<thead>
<tr>
<th>Method</th>
<th>RNABindR</th>
<th>Neural Net</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlation Coefficient</td>
<td>0.30</td>
<td>0.29</td>
</tr>
<tr>
<td>Accuracy</td>
<td>76.6%</td>
<td>77.5%</td>
</tr>
<tr>
<td>Specificity</td>
<td>47%</td>
<td>47%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>43%</td>
<td>40%</td>
</tr>
</tbody>
</table>
**RNABindR detects known PROSITE RNA binding motifs**

To compare the motifs picked out by RNABindR with known RNA binding motifs, we identified all PROSITE motifs (Hulo et al., 2004) annotated as nucleic acid binding in the proteins from our dataset. PROSITE contains a collection of sequence patterns that are known to be associated with a particular protein family or function. By identifying all of the PROSITE motifs that are involved in nucleic acid binding in our non-redundant dataset, we can compare RNABindR performance with simply searching for known RNA binding sequence motifs. RNABindR identified 104 out of 109 proteins (95%) in the non-redundant dataset as RNA binding proteins, whereas PROSITE identified RNA binding motifs in only 17 out of 109 chains (15.6%). The interface residues predicted by RNABindR lie within the boundaries of the PROSITE motifs for 16 out of these 17 chains, demonstrating that RNABindR does identify known RNA binding motifs. Furthermore, the fact that RNABindR detected RNA binding sites in 88 proteins that do not contain any PROSITE motif whose annotation indicates a role in RNA binding, suggests that RNABindR could be used to identify novel RNA binding motifs.

The tRNA pseudouridine synthase protein shown in Figure 2.4D contains the PROSITE PUA domain (PS50890), which is predicted to be an RNA binding domain. The PROSITE PUA domain contains 77 amino acids, only 6 of which contact RNA. RNABindR predicted a cluster of three interface residues in this region, shifted relative to the cluster of three actual interface residues in the complex structure, but precisely overlapping one actual interface residue. This example illustrates that RNABindR is able to identify specific interface residues while a search for sequence motifs, such as a PROSITE search, may only identify larger RNA binding domains.
DISCUSSION

In this paper, we have presented RNABindR, a machine learning based tool for identifying RNA binding sites in proteins. To generate a widely applicable tool, we developed RNABindR to use only protein sequence information as input. This achievement is significant because the results presented here were obtained using a relatively small training set of non-redundant RNA-protein complexes chosen from the PDB to allow rigorous evaluation of classification performance. On this dataset of 109 diverse proteins (sequence identity below 30%), RNABindR performs well enough to be useful, with 85% accuracy, 0.35 CC, 0.51 specificity+, and 0.38 sensitivity+. Higher specificity values (with lower sensitivity) can be obtained in practice, if required, because RNABindR uses a Naive Bayes classifier, which allows the user to trade-off sensitivity against specificity by tuning the classification threshold.

To evaluate RNABindR’s ability to identify potential RNA binding sites in proteins for which structural information is not available, we predicted RNA binding residues in the telomerase TERT protein. To date, there is no high-resolution structure of the hTERT-TR complex, primarily because it has not been possible to obtain sufficient quantities of soluble full-length hTERT for detailed biophysical studies (Jacobs et al., 2005). Thus, we compared RNABindR predictions with available experimental data regarding conserved motifs and RNA binding domains in hTERT. The fact that RNABindR correctly predicted clusters of interface residues within known RNA binding domains of hTERT, and in several cases, precisely identified interface residues defined by mutagenesis experiments for hTERT, suggests that RNABindR could be valuable in designing experiments to identify RNA binding sites in other experimentally recalcitrant RNA-protein complexes. (See Terribilini et al., 2006, for an example of this). Although there is still no experimental structure for any TERT-RNA complex, the recent determination of the structure of a domain of Tetrahymena TERT prompted us to evaluate RNABindR predictions on Tetrahymena TERT. A cluster of
predicted interface residues from 185-191 in Tetrahymena TERT is confirmed by the available experimental evidence for RNA binding in this region of the protein. Several residues in both hTERT and Tetrahymena TERT predicted by RNABindR are located outside the boundaries of the essential RNA binding and RT catalytic domains so far defined by experiments. It will be interesting to determine whether these predicted RNA binding residues may, in fact, contact RNA to stabilize the complex or to assist in other functions, such as subnuclear localization of TERT (Blackburn, 2005).

We found that the performance of RNABindR, using a Naive Bayes classifier, was comparable to that of the only previously published sequence-based tool for predicting RNA binding sites, a neural network classifier developed by Miyano's group (Jeong et al., 2004). An advantage of RNABindR over the neural network classifier is that the latter method requires the exploration of several alternative neural network architectures (number of layers between the input and output layers, the number of neurons in such intermediate layers and the connectivity between layers) before settling on an optimal network structure. In contrast, a Naive Bayes classifier does not require such hand-tuning. A Naive Bayes classifier requires significantly less computational effort (a single pass through the training data) to train than a neural network classifier (which requires multiple passes through the training data), making it especially well suited for use with large datasets or in settings that call for incremental update of the classifier as new training data become available.

Several classes of RNA binding domains and motifs that mediate the recognition of RNA by proteins have been very well characterized (Draper, 1999). Two abundant and structurally-defined RNA binding motifs are the RDB or RNA-recognition motif (RRM), which is the most common single-stranded RNA binding motif, and the double-stranded RNA binding motif (dsRBM) (Hall, 2002), which recently has been shown to play important roles in regulatory interactions mediated by siRNAs and miRNAs (Tian et al., 2004). Shorter sequence motifs, including the arginine-rich-motif (ARM) motif and Arg-Gly-Gly (RGG)
box are also found in a large number of proteins (Mulder et al., 2003). Within the non-redundant dataset of 109 validated RNA binding proteins, only 17 PROSITE RNA binding motifs were identified. RNABindR predicted RNA binding residues in 104 of the 109 proteins and predicted interface residues within 16 of the 17 PROSITE RNA binding motifs. Additionally, most of the sequences "hit" by the 17 PROSITE motifs consist of relatively long stretches of amino acids that contain very few actual interface residues. Because the PROSITE motifs were not generated for the purpose of identifying interface residues, this comparison is not intended to prove "better performance" of RNABindR but instead to indicate that RNABindR may also be valuable for identifying novel RNA binding motifs.

A major challenge in post-genomics research is the functional annotation of novel proteins of known sequence (and, increasingly, known structure) but unknown function. For example, ORFans, orphan open reading frames that share no significant sequence similarity with any ORFs outside the genome in which they reside, represent 20-30% of genes in sequenced genomes, but their origins and functions are largely mysterious (Fischer & Eisenberg, 1999; Siew & Fischer, 2004). Recently, several groups have demonstrated success in automatic prediction of protein functional interactions and intermolecular interfaces based on primary sequence information (Rost et al., 2003; Pang et al., 2004). However, when additional types of information are available (e.g., structural motifs, physical interactions, expression profiles, cellular localization, phylogenetic relationships), they can be incorporated to improve the accuracy of functional annotation. For example, for DNA-binding proteins, the use of structure-derived features such as small binding motifs, solvent accessibility and positive electrostatic potential have been shown to improve detection of HTH, HhH and HLH DNA binding motifs (Shanahan et al., 2004). The prediction of protein-protein interface residues is also significantly improved by incorporating diverse types of information (de Vries et al., 2006; Bradford & Westhead, 2004; Haskins et al., 2006; Neuvirth et al., 2004; Nissink & Taylor, 2004; Sen et al., 2004).
In experiments not reported here, we did not obtain significant improvement in classifier performance by incorporating sequence conservation information derived from multiple sequence alignments or residue solvent accessibility information derived from known structures of proteins in the training dataset (see Materials and Methods; data not shown). This was unexpected because including sequence entropy or relative solvent accessibility of the target residue along with the input of amino acid identities does, in fact, enhance performance when a similar Naive Bayes classifier is used to predict interface residues in DNA-protein binding sites (Yan et al., In press). Current experiments are directed at investigating the basis for this difference between DNA and RNA binding site classifiers. We are also exploring different encodings that may result in classifiers that more effectively exploit additional types of information.

Even without using information derived from structure, it should be possible to enhance prediction of RNA-protein interface residues. Recent results from Jeong and Miyano (Jeong & Miyano, 2006) have shown that using position specific scoring matrices (PSSMs) derived from PSI-BLAST searches can improve prediction performance of neural network classifiers. Recent preliminary experiments using PSSMs as inputs for RNABindR resulted in improved prediction performance comparable with that of Jeong and Miyano (data not shown). Other methods to improve prediction of interface residues may include, for example, adding "filters" that eliminate false positives based on the estimated probability that a particular interface residue should be located near other interface residues within the primary sequence, as has been done to improve performance of classifiers for identifying protein-protein interface residues (Ofran & Rost, 2003; Yan et al., 2004b). Alternatively, training on larger datasets of structurally or functionally related RNA binding proteins, generated by relaxing the redundancy criterion may generate higher accuracy predictions for specific subclasses of RNA binding proteins.
The RNABindR results reported here, together with results of previous studies published by Jeong and Miyano (Jeong et al., 2004; Jeong & Miyano, 2006), demonstrate that computational approaches can successfully identify RNA-protein interface residues using only amino acid sequence as input. For many proteins - notably, the ORFans, mentioned above - the deduced amino acid sequence is often the only information available. The approach we propose here requires only the primary sequence of the protein partner, implying that many structural determinants of RNA binding sites can be captured by local sequence characteristics. The simplicity of RNABindR, together with the fact that a relatively high level of accuracy can be achieved using only protein sequence information (and no information about the identity, sequence or structure of the RNA partner), suggests that it may prove valuable for functional annotation of putative RNA binding proteins and for genome-wide identification of RNA binding residues in protein. RNABindR is available at http://bindr.gdcb.iastate.edu.

METHODS

Dataset

A dataset of RNA-protein interactions was extracted from structures of known RNA-protein complexes solved by X-ray crystallography in the Protein Data Bank (PDB) (Berman et al., 2000). Proteins with >30% sequence identity or structures with resolution worse than 3.5Å were removed using PISCES (Wang & Dunbrack, 2003). This resulted in a set of 109 non-redundant protein chains containing a total of 25,118 amino acids. Amino acids in the RNA-protein interface were identified using ENTANGLE (Allers & Shamoo, 2001). Using default parameters, 3518 (14%) of the amino acids in the dataset are defined as interface residues (positive examples). Table 2.2 lists the PDB identifiers for all 109 proteins in the non-redundant dataset, which includes four major classes of RNA-protein complexes. A smaller dataset extracted from only ribosomal proteins (55 chains) was used in some
experiments. The ribosomal protein dataset comprises a total of 7522 amino acids, 2363 (31%) of which are defined as interface residues. These datasets and others are available at http://bindr.gdcb.iastate.edu.

Table 2.4 RNA binding proteins in the nonredundant training data set. RNA binding proteins corresponding to each four major RNA classes are shown along with their PDB identifiers. The complete non-redundant data set contains all 109 protein chains. Protein names and additional details are provided online at http://bindr.gdcb.iastate.edu/RNABindR/datasetSummaryTable.htm.

<table>
<thead>
<tr>
<th>Chains</th>
<th>Type</th>
<th>PDB IDs</th>
</tr>
</thead>
<tbody>
<tr>
<td>19</td>
<td>tRNA</td>
<td>1A34:A, 1DDL:A, 1E6T:A, 1F8V:A, 1H2C:A, 1LAJ:A, 1N34:A, 1NB7:A, 1PGL:2, 1RMV:A, 2BBV:C, 2BBV:F</td>
</tr>
</tbody>
</table>

Naive Bayes classifier using only amino acid sequence information as input

The results reported in this paper were obtained using RNABindR implementing a Naive Bayes classifier (Mitchell, 1997), which was chosen based on exploration of several different machine learning algorithms, including support vector machines, decision trees, and Bayesian networks. The performance of the Naive Bayes classifier was comparable to or
better than that of all other methods tested (data not shown). The Naive Bayes classifier assumes the independence of attributes. This assumption greatly reduces the complexity of the classifier and improves the reliability of the estimated parameters when the dimensionality of the input is high relative to the size of the available training set. Despite its simplicity and the fact that the independence assumption may not apply in certain cases, the Naive Bayes classifier often performs at least as well as more sophisticated methods for many problems (Buntine, 1991). We used the Naive Bayes classifier from the Weka package (Witten & Frank, 2000). In RNABindR, the input to a Naive Bayes classifier is a window \( x = (x_{-n}, x_{-n+1}, \ldots, x_{T-1}, x_T, x_{T+1}, \ldots, x_{n-1}, x_n) \) of \( 2n+1 \) contiguous amino acid identities, with \( n \) amino acid sequence residues on either side of the target residue \( x_T \). The output is an instance \( c \in \{+,-\} \) where + indicates that the target residue \( x_T \) at the center of the window is an interface residue and – indicates \( x_T \) is a non-interface residue. A training example is an ordered pair \((x, c)\) where \( x = (x_{-n}, x_{-n+1}, \ldots, x_{T-1}, x_T, x_{T+1}, \ldots, x_{n-1}, x_n) \) and \( c \) is the corresponding class label (interface or non-interface). A training dataset \( D \) is simply a collection of labeled training examples. In our experiments, several values of \( n \) from 2 to 14 (corresponding to windows of width 5 to 29) were tried.

Let \( X = (X_{-n}, \ldots, X_T, \ldots, X_n) \) 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 Naive Bayes classifier assigns input \( x \) the class label + (interface) if:

\[
\frac{P(C = + | X = x)}{P(C = - | X = x)} \geq \theta
\]

and the class label – (non interface) otherwise. The choice of \( \theta = 1 \) corresponds to assigning the most probable class label. The desired trade-off of sensitivity against specificity can be achieved by varying \( \theta \).
Because the inputs are assumed to be independent given the class, we have:

\[
\frac{P(C = + | X = x)}{P(C = - | X = x)} = \frac{P(X = x | C = +)P(C = +)}{P(X = x | C = -)P(C = -)} \\
= \frac{P(C = +) \prod_{i=n}^{i=-n} P(X_i = x_i | C = +)}{P(C = -) \prod_{i=n}^{i=-n} P(X_i = x_i | C = -)}
\]

The relevant probabilities are estimated from the training set using the Laplace estimator (Mitchell, 1997). The resulting Naive Bayes classifier classifies a target amino acid residue \(x_T\) as an interface residue or as a non-interface residue based on the identities of the \(n\) amino acid residues on either side.

**Naive Bayes classifiers using sequence plus additional information**

We experimented with adding relative accessible surface area (rASA), sequence entropy, hydrophobicity, secondary structure, or electrostatic potential to the sequence based classifier described above. rASA for each residue in the absence of RNA was computed using the program Naccess (http://wolf.bms.umist.ac.uk/naccess/). Each training and test example for the Naive Bayes classifier with rASA added is as follows:

\[x = (x_{-n}, x_{-n+1}, \ldots, x_{T-1}, x_T, x_{T+1}, \ldots, x_{n-1}, x_n, r_T)\], where \(x_i\) is as defined above and \(r_T\) is the rASA of the target residue. Inputs are encoded similarly for all features. Sequence entropy was encoded using the relative entropy for each residue from the HSSP database (http://www.cmbi.kun.nl/gv/hssp/). Hydrophobicity of each residue was obtained from the consensus normalized hydrophobicity scale derived by Eisenberg et al. (Eisenberg et al., 1984). The secondary structure of each residue was extracted from the PDB (http://www.rcsb.org/pdb/). Electrostatic potentials were calculated using the program APBS.
The electrostatic potential for each residue is the average over all its atoms.

**Performance Evaluation**

The performance of RNABindR was evaluated using leave-one-out cross validation experiments. That is, in each of the 109 experiments, the Naive Bayes classifier was trained using data from 108 chains and evaluated on the 109th chain. The threshold $\theta$ was chosen to maximize the correlation coefficient on the training set. The performance measures reported represent averages over the 109 experiments. The performance of a classifier designed to classify protein residues into interface and non-interface residues is completely summarized by TP (true positives) i.e., the number of interface residues correctly identified as such by the classifier; FP (false positives) i.e., the number of non-interface residues misclassified as interface residues by the classifier; FN (false negatives) i.e., the number of interface residues that are misclassified as non-interface residues by the classifier; and TN (true negatives) i.e., the number of non-interface residues that are correctly identified as such by the classifier. Note that $N$, the total number of instances used for evaluation of the classifier is given by $N = TP + FP + FN + TN$.

Commonly used performance measures include accuracy, correlation coefficient (CC), specificity+, sensitivity+, specificity-, and sensitivity- (Baldi et al., 2000). Specificity+ is the fraction of positive predictions (residues predicted to be RNA-binding residues) that are actually RNA-binding residues. For example, if 100 interface residues are predicted to be RNA-binding residues by RNABindR and 50 of them are actually interface residues, specificity+ is 0.5. Sensitivity+ is the fraction of RNA-binding residues that are predicted to be RNA-binding residues by RNABindR. For example, if a protein contains 20 actual interface residues and RNABindR predicts that 15 of these 20 are interface residues, sensitivity+ is 0.75.
Specificity- and Sensitivity- are similarly defined. Each of these performance measures summarizes the information contained in the four numbers (TP, FP, FN, TN) with a single number (e.g., accuracy), with inevitable loss of information. In the case of datasets in which there is a large difference between the number of instances belonging to the two classes, using accuracy alone to evaluate the classifier can be misleading (Baldi et al., 2000; Yan et al., 2004a; Yan et al., 2004b). The RNA binding site dataset contains 14% interface and 86% non-interface examples. A classifier that simply predicts each residue to be non-interface would have an accuracy of 0.86, but such a classifier would be completely useless in correctly identifying the interface residues. Thus, it is desirable to consider multiple performance measures collectively to evaluate the performance of a classifier and compare its performance with other classifiers (Baldi et al., 2000; Yan et al., 2004b).

As noted earlier, it is possible (and in many settings desirable), to trade off the sensitivity of the classifier against its false positive rate. The Receiver Operating Characteristic curve (ROC curve), a plot of the sensitivity+ or the “hit rate” versus the false positive rate (1-specificity-) characterizes such tradeoff for a classifier. We used the Weka package (Witten & Frank, 2000) to obtain the ROC plot for RNABindR.

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

\[
\text{Specificity}+ = \frac{TP}{TP + FP}
\]

\[
\text{Sensitivity}+ = \frac{TP}{TP + FN}
\]

\[
\text{Corr.Coeff.} = \frac{(TP\times TN) - (FP\times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]
ACKNOWLEDGEMENTS

This research was supported in part by grants from the National Institutes of Health (GM066387) to V.H., D.D and R.J., research assistantships from USDA IFAFS Multidisciplinary Graduate Education Training Grant (2001-52100-11506) for M.T., and from Iowa State University Biotechnology Council and Center for Integrated Animal Genomics for J.L. We would like to thank Satoru Miyano, Euna Jeong, and I-Fang Chung for kindly providing their dataset and members of our research groups for ideas and discussions.

REFERENCES


CHAPTER 3 RNABINDR: A SERVER FOR ANALYZING AND PREDICTING RNA BINDING SITES IN PROTEINS

A paper published in Nucleic Acids Research

Michael Terribilini, Jeffry Sander, Jae-Hyung Lee, Peter Zaback, Robert L. Jernigan, Vasant Honavar, and Drena Dobbs

ABSTRACT

Understanding interactions between proteins and RNA is key to deciphering the mechanisms of many important biological processes. Here we describe RNABindR, a web-based server that identifies and displays RNA-binding residues in known protein–RNA complexes and predicts RNA-binding residues in proteins of unknown structure. RNABindR uses a distance cutoff to identify which amino acids contact RNA in solved complex structures (from the Protein Data Bank) and provides a labeled amino acid sequence and a Jmol graphical viewer in which RNA-binding residues are displayed in the context of the three-dimensional structure. Alternatively, RNABindR can use a Naive Bayes classifier trained on a non-redundant set of protein–RNA complexes from the PDB to predict which amino acids in a protein sequence of unknown structure are most likely to bind RNA. RNABindR automatically displays ‘high specificity’ and ‘high sensitivity’ predictions of RNA-binding residues. RNABindR is freely available at http://bindr.gdcb.iastate.edu/RNABindR.

INTRODUCTION

Protein-RNA interactions are vital to a wide range of biological processes, including regulation of gene expression, protein synthesis, and replication and assembly of many
viruses (1,2,3,4). A more detailed understanding of protein-RNA interactions is especially important for understanding how miRNA and other non-coding RNAs regulate gene expression. The ability to computationally predict which residues of a protein directly participate in RNA binding has already contributed to the design of wet-lab experiments to decipher mechanisms of protein-RNA recognition (5,6) and has the potential to enhance our fundamental understanding of how proteins recognize RNA.

Here we describe RNABindR, a web-based server that uses machine learning approaches to identify amino acids in a protein that are most likely to participate in RNA binding. In previous work, we demonstrated that RNABindR can predict RNA binding residues with high accuracy, using only the amino acid sequence of a query protein (and no information about the bound RNA) as input (7). In the current web-based implementation, RNABindR allows users to: i) Identify actual binding residues for a given protein-RNA complex in the Protein Data Bank (PDB) (8). ii) Predict RNA binding residues in a protein sequence whose RNA-bound structure is not available in the PDB. When calculating actual binding residues for a known structure, the only required input is the PDB ID of a protein-RNA complex and an interface distance cutoff in Angstroms (Å). The RNABindR server calculates which amino acids in the protein have atoms within the defined cutoff distance of atoms in the bound RNA. It returns a display of the labeled amino acid sequence and a Jmol (www.jmol.org) graphical viewer in which RNA binding residues are highlighted within the three-dimensional structure of the complex. To predict RNA binding residues for a protein of unknown structure, the user must provide the amino acid sequence of a protein of interest. The RNABindR server returns the amino acid sequence with the predicted RNA binding status (+ or -) for each residue. Three different prediction results, reflecting different expected specificity values, are provided for each query sequence, allowing users to compare results with high “specificity” versus high “sensitivity” for RNA binding residues. RNABindR is designed to be fast and easy to use; results are typically returned within a few
seconds. Output can be displayed as described above, or can be downloaded as a file to facilitate transfer into other programs.

**MATERIALS AND METHODS**

**Dataset of Protein-RNA interactions**

A training dataset of protein-RNA interactions was extracted from structures of known protein-RNA complexes in the PDB solved by X-ray crystallography. Proteins with >30% sequence identity or structures with resolution worse than 3.5 Å were removed using PISCES (9). This resulted in a dataset, RB147, containing 147 non-redundant protein chains and a total of 32,324 amino acids. This dataset is larger than the RB109 dataset used in our previously published work (5,7), where a different method was used to define RNA binding residues. Previously, we used the ENTANGLE program (10) to identify amino acids in contact with RNA. For the dataset used in the current implementation of RNABindR, RNA binding residues were identified according to a distance-based cutoff definition: an RNA binding residue is an amino acid containing at least one atom within 5 Å of any atom in the bound RNA. According to this definition, RB147 contains a total of 6157 RNA binding residues and 26,167 non-binding residues.

**Naive Bayes classifier**

RNABindR uses a Naive Bayes classifier (11) as implemented in the Weka software package (12) for all predictions. A detailed description of the algorithm and evaluation of its performance on several different datasets of RNA binding proteins has been published (7). Briefly, the Naive Bayes classifier assumes the independence of attributes. This assumption greatly reduces the complexity of the classifier. In RNABindR, the input to a Naive Bayes classifier consists of a window \( x = (x_{-n}, x_{-n+1}, \ldots, x_{-1}, x_0, x_1, \ldots, x_{n+1}, x_n) \) of \( 2n+1 \) contiguous amino acid identities, with \( n \) amino acid sequence residues on either side of the target residue.
The output is an instance \( c \in \{+, -\} \) where “+” indicates that the target residue \( x_T \) at the center of the window is a RNA binding residue and “−” indicates \( x_T \) is not a RNA binding residue. The Naive Bayes classifier assigns the class label “+” to input \( x \) if:

\[
\frac{P(C = + | X = x)}{P(C = - | X = x)} \geq \theta
\]

and the class label “−” otherwise. The desired trade-off of sensitivity versus specificity can be achieved by varying \( \theta \), which is the classification threshold. Specificity is the fraction of residues predicted to be RNA binding residues that are in fact RNA binding residues. Sensitivity is the fraction of actual RNA binding residues that are predicted to be RNA binding residues by RNABindR (7,13).

**Reliability of RNABindR predictions**

RNABindR has been evaluated using leave-one-out cross validation experiments with several different datasets of RNA binding proteins (7). For the Naive Bayes classifier implemented in the current web-based version of RNABindR, one protein sequence was used as the test set and the other 146 sequences in the RB147 dataset were used as the training set for each round of training. This process was repeated until each protein had been used as the test set. Figure 3.1 depicts RNABindR performance over all values of \( \theta \) and the inset table provides a summary of the average classification performance of RNABindR on the RB147 dataset, using three different values of the classification threshold, \( \theta \). The results illustrate that, as with other machine learning methods, in the RNABindR predictions there is a trade-off between the specificity (or “precision”) and sensitivity (or “recall”). Changing the value of \( \theta \) changes the number of predicted RNA binding residues and the “confidence” with which binding residues are predicted. In classification tasks that involve unbalanced training sets (i.e., unequal numbers of positive and negative examples), as is the case here, the correlation coefficient (CC) is perhaps the best single parameter for comparing the “overall”
performance of different machine learning algorithms (13); also see (7) for further discussion and precise definitions of performance parameters used in our work.)

As shown in Figure 3.1, using the “high specificity” classification threshold, RNABindR predicts a smaller number of RNA binding residues, with higher confidence: 80% of the RNA binding residues predicted for the RB147 dataset are, in fact RNA binding residues. In contrast, using the “high sensitivity” classification threshold, RNABindR predicts a larger number of RNA binding residues, but with lower confidence: only 28% of the RNA binding residues predicted for the RB147 dataset are actually RNA binding. Using this high sensitivity threshold, however, a much higher fraction (~80%) of the actual binding residues is identified. The third prediction provided by RNABindR, referred to as the “optimal” prediction, uses a threshold corresponding to the value of $\theta$ that maximizes the correlation coefficient for predictions on the RB147 dataset. The “optimal” prediction is not guaranteed to be the best prediction. Instead it is a prediction in which the trade-off between specificity and sensitivity has been optimized on the training dataset.
Figure 3.1 Summary of RNABindR performance in predicting RNA-binding residues.

Specificity versus sensitivity trade-off and the average performance statistics for RNABindR in leave-one-out cross-validation experiments on the RB147 dataset are shown. The plot shows the specificity and sensitivity values across the entire range of the classification threshold $\theta$, with the ‘Optimal,’ ‘High Specificity,’ and ‘High Sensitivity’ points marked. The columns in the table show results obtained using the three different classification thresholds employed by RNABindR. The ‘Optimal Prediction’ uses the threshold value that maximizes the correlation coefficient on the training dataset; this prediction represents a balance between the competing goals of identifying as many RNA-binding residues as possible and minimizing the number of false positives. The ‘High Specificity Prediction’ identifies fewer RNA-binding residues, but with higher confidence in the positive predictions. The ‘High Sensitivity Prediction’ identifies more RNA-binding residues, but at the cost of an increased false positive rate.

Definitions of performance measures are according to Baldi et al. (2001) (13). Specificity ‘$+$’ and ‘$-$’ refer
to specificity on the positive class (RNA-binding residues) and negative class (non-RNA-binding residues), respectively.

**SERVER DESCRIPTION**

RNABindR provides two main services: i) Identification of RNA binding residues, given the structure of a protein-RNA complex; ii) Prediction of RNA binding residues given a protein sequence. An overview of RNABindR is provided in Figure 3.2.

![Figure 3.2 RNABindR flowchart. The query sequence is first compared with every protein sequence in every protein–RNA complex structure in the PDB to search for an exact match. If a match is found, the prediction program is not run and the actual RNA-binding residues are calculated using a distance cutoff and returned, along with an interactive Jmol image highlighting interface residues within the protein–RNA complex structure. If an exact sequence match is not identified, the Naive Bayes classifier is run and the predicted RNA-binding and non-binding residues are returned (using three different classification threshold values, see text).](image-url)
Calculation of RNA binding residues in protein-RNA complexes of known structure

Input – To identify RNA binding residues (i.e., amino acid residues that lie in the interface between protein and bound RNA) in a known protein-RNA complex, the only required input is the PDB ID of the complex. RNABindR parses the PDB file to determine which chains in the complex are protein and which are RNA. Interactions are calculated for each protein chain with every RNA chain in the complex. For example, for a protein-RNA complex with two protein chains (A and B), and two RNA chains, (C and D), interactions will be calculated between the following pairs of chains: A and C, A and D, B and C, and B and D. If desired, the user can enter a single protein chain identifier to restrict the output to only those interactions between the specified protein chain and the RNA chain(s) in the complex.

By default, RNABindR uses a distance cutoff of 5 Å between any atom of the amino acid and any atom of the RNA to determine which residues interact with the RNA. However, RNABindR allows the user to change this parameter to any desired value (between 0 and 100 Å) to make the definition of RNA binding more or less stringent.

Output – Figure 3.3 shows an example of RNABindR output to identify RNA binding residues in a known protein-RNA complex. The output is a display of the sequence of each chain in the complex, with a label for each residue; “+” for residues that are within the specified distance cutoff and “-” for residues that do not have any atoms within the distance cutoff. The calculated RNA binding residues are also displayed on the PDB structure of the protein-RNA complex using Jmol (www.jmol.org). By default, the RNA binding residues are displayed in red space-fill representation, the rest of the protein is displayed in blue space-fill, and the bound RNA is displayed in green wireframe. Users can also print or download the text output to facilitate further analysis of the calculated RNA binding residues.
Figure 3.3 Example of RNABindR results: identifying actual RNA-binding residues in a known protein–RNA complex. RNABindR output includes the amino acid sequence of the identified protein chain(s) in the complex, with a ‘+’ label for each interacting residue (those having atoms within the selected RNA contact cutoff distance) and ‘−’ for non-binding residues. Below, a Jmol applet displays the structure of the protein–RNA complex. RNA-binding residues are displayed in red space-fill, non-
Prediction of RNA binding residues in proteins of unknown structure

Input – To predict RNA binding residues in a protein of interest, the only required input is the amino acid sequence of the protein. RNABindR accepts FASTA-formatted protein sequences in the single-letter amino acid representation, but is able to read any standard amino acid sequence format; any characters (e.g., sequence numbering or blank spaces) that are not part of the standard 20-letter amino acid alphabet are ignored. After processing the sequence to remove any extra characters, RNABindR determines whether the query sequence has an exact match in any protein-RNA complexes available in the PDB. If an exact match to the query sequence is identified, the prediction program is not run. Instead, RNABindR returns the actual RNA binding residues from the PDB complex and a Jmol image of its structure, in which the RNA binding residues are highlighted as described above. If no exact match is found, RNABindR predicts RNA binding residues in the query protein sequence. In the current implementation, RNABindR predictions are made using a Naive Bayes classifier trained on all 147 protein chains in the RB147 dataset; the input query sequence is used as the test case.

Output – Figure 3.4 shows an example of RNABindR output obtained for predicting RNA binding residues in a protein of unknown structure. The input amino acid sequence is shown at top, and labels “+” and “-” for predicted RNA binding and non-binding residues, respectively, are shown immediately below the sequence. Users can also print or download the text output to facilitate further analysis of the predicted RNA binding residues.
Figure 3.4 Example of RNABindR results: predicting RNA-binding residues in a protein of unknown structure. RNABindR output includes the query sequence and three predictions obtained using three different classification thresholds. Residues predicted to bind RNA are indicated by ‘+’ and non-binding residues by ‘−’ on the line below the sequence. The ‘optimal prediction’ uses the threshold value that maximizes the correlation coefficient on the RB147 dataset. The ‘high specificity prediction’ provides fewer predicted RNA-binding residues, with higher confidence, and the ‘high sensitivity prediction’ provides more predicted RNA-binding residues, but with lower confidence. Links are provided for downloading the predictions in a text-only format or a printer friendly format.
Typical users of RNABindR may have different goals in mind when predicting RNA binding residues: some may wish to identify a relatively small number of amino acids predicted to bind RNA with “high confidence,” while others may wish to identify as many potential RNA binding residues as possible, with more potentially “false positive” predictions. To accommodate these different uses, RNABindR displays three different predictions for each query sequence: an “optimal” prediction, a “high specificity” prediction, and a “high sensitivity” prediction. As discussed above, the high specificity prediction uses a more stringent classification threshold to identify the most likely RNA binding residues, whereas the high sensitivity prediction uses a less stringent threshold to identify more potential RNA binding residues. Because the reliability of RNABindR predictions for any particular protein depends on the extent to which the query protein shares features that are “captured” by the Naive Bayes classifier (during training on the RB147 dataset), prediction performance for any particular query sequence cannot be guaranteed. The three types of predictions are supplied as a guide to help the user make best use of RNABindR predictions.

**RELATED SERVERS**

Predicting RNA binding residues has proven to be an important and difficult computational task (7,14,15,16). Since RNABindR was developed, two other web-based servers for RNA binding site predictions have become available, BindN (14) and KYG (15). BindN (http://bioinformatics.ksu.edu/bindn) uses a support vector machine (SVM) to predict both RNA binding and DNA binding residues in a protein sequence. BindN is a sequence-based server, requiring only the amino acid sequence of a query protein. The feature vector used as input to the SVM classifier consists of the side chain pKa value, hydrophobicity index, and molecular mass for each amino acid in a window of 11 residues. The BindN server requires users to choose an estimated specificity or sensitivity, which is used to determine the classification threshold (14). KYG
(http://yayoi.kansai.jaea.go.jp/qbg/kyg/index.php) provides several methods for statistically analyzing a protein structure and predicting RNA binding residues. KYG is a structure-based server and relies on estimating the interface propensity for single amino acids and pairs of amino acids. KYG also utilizes evolutionary information in the form of a multiple sequence alignment profile, which must be supplied by the user. Users are allowed to choose among nine different predictions, each of which is based on a different combination of residue propensities and profile scores. The KYG server can predict RNA binding residues only for those proteins whose structures are known.

RNABindR offers some potential advantages over BindN and KYG. RNABindR has been designed to be user-friendly and widely applicable. Like BindN, RNABindR requires only a protein sequence as input, so researchers can obtain predictions for any protein sequence of interest. RNABindR does not require users to specify any parameters or choose between different methods. Also, RNABindR provides a quick and easy way to visualize RNA binding residues and examine the protein-RNA interface(s) within the three-dimensional structure of any known protein-RNA complex.

RNABindR, BindN and KYG each use different methods, are trained on different datasets, and often provide different predictions of RNA binding residues for the same query protein sequence. Users may use all three servers and apply their biological expertise regarding their protein of interest to determine which predictions warrant further investigation.

**SIGNIFICANCE AND FUTURE DIRECTIONS**

Over the last decade, there has been a dramatic increase in the number of available structures of protein-nucleic acid complexes: the Protein Data Bank (PDB) included only 198 protein-nucleic acid complexes in 1996, but by April 2007, this number had grown to 1734, of which 529 were protein-RNA complexes [PDB, accessed April 3, 2007,
http://www.pdb.org]. The resulting availability of larger and more diverse training sets can be expected to significantly improve the performance of RNABindR. RNABindR will be updated periodically to take advantage of the latest data available in the PDB. A beta-version with three types of enhancements is under development. In recent work, we have generated a comprehensive database that includes every protein-RNA interface for which structural information is available in the PDB. The next version of RNABindR will incorporate this complete database. Users will have the option of choosing a classifier that is trained on the comprehensive dataset or on one of several “non-redundant” datasets (e.g., RB 147). Alternatively, users will be able to train a new classifier using a “customized” training dataset (e.g., any subset of known protein-RNA complexes, chosen based on similarities in sequence or biological function). Recent unpublished and earlier published results (7) indicate that using such training datasets can provide a significant increase in the reliability of RNA binding site predictions. A second enhancement will be to allow users to choose among several machine learning algorithms (e.g., SVMs) or statistical methods that have been shown to be effective for RNA binding site prediction by our group and by others (5,7,14,15,16). Third, RNABindR will allow users to take advantage of structural and/or evolutionary information, when available. If the structure of a query protein is available in the PDB (but the structure of the query protein in complex with RNA is not), predicted RNA binding residues will be identified and displayed on the three-dimensional structure of the protein, as is done for calculated RNA binding residues in known protein-RNA complexes in the current implementation (see Figure 3.3). In the longer term, structural predictions will also be included for such RNA binding sites, based on structure fragment libraries and other homology modeling approaches.

Protein-RNA interactions play many essential and diverse roles in biological regulation, ranging from structural and catalytic roles in ribosomes and spliceosomes, to regulatory roles in microRNA-mediated gene regulation and cellular signaling, to storage and
propagation of genetic information (17,18,19,20). Despite their obvious functional importance, the details of the molecular mechanisms of protein-RNA recognition are still poorly understood. The impressive diversity of structures and functions of protein-RNA complexes makes understanding what dictates specificity in protein-RNA interaction an especially challenging problem (18). Hence, computational tools for analyzing protein-RNA interfaces and for predicting RNA binding sites in proteins are becoming increasingly important for deciphering the amino acid sequence and structural underpinnings of protein-RNA interactions (7,14,15,16,21,22,23,24,25). RNABindR predictions have already helped guide the experimental investigation of the RNA binding domains in proteins (5,6). Approaches that combine computational prediction and experimental validation of RNA binding sites in proteins will increase our understanding of the mechanisms of protein-RNA recognition.

ACKNOWLEDGEMENTS

We are grateful to members of our groups and colleagues who have tested RNABindR and provided valuable feedback. This work was supported in part by NIH grant GM066387 and by graduate research assistantships provided by USDA MGET grant 2001-52100-11506, the ISU Center for Integrated Animal Genomics (CIAG) and the ISU Center for Computational Intelligence, Learning, and Discovery (CCILD). Funding to pay the Open Access publication charges for this article was provided by NIH grant GM066387.

REFERENCES


CHAPTER 4. PREDICTING RNA BINDING SITES IN PROTEINS: USE OF STRUCTURAL AND EVOLUTIONARY INFORMATION LEADS TO IMPROVED PERFORMANCE

A paper to be submitted to *PLOS Computational Biology*.

Michael Terribilini, Cornelia Caragea, Jeffry Sander, Vasant Honavar, and Drena Dobbs

**ABSTRACT**

Because of the importance of protein RNA interactions in the diverse roles played by RNA in the cell and due to the high cost of experimental determination of protein-RNA interfaces, there is an urgent need for accurate methods for predicting protein-RNA interface residues in proteins. We compare several classifiers trained to predict RNA binding sites in proteins using information derived from a protein sequence or protein structure. The input to a classifier consists of an encoding of the target residue plus its sequence or structural neighbors. Each residue is encoded using its amino acid identity, or its PSSM (position specific scoring matrix) obtained using PSI-BLAST. We consider two classifiers, IDSeq and IDStruct, that encode each amino acid and its sequence or structural neighbors using the 20-letter amino acid alphabet; two classifiers, PSSMSeq and PSSMStruct, which encode each amino acid and its sequence or structural neighbors using their PSSMs; and Ensemble, which combines the results of the IDStruct, PSSMSeq, and PSSMStruct classifiers. The IDSeq, IDStruct, PSSMSeq, PSSMStruct, and Ensemble classifiers achieve AUCs of 0.74, 0.77, 0.79, 0.80, and 0.81. The difference in AUC between each pair of classifiers is statistically significant. We also find that number of sequences that are homologous to the target sequence plays a role in determining whether PSSM-based encoding of the residues improves
predictive performance of the resulting classifier; proteins that had at least 100 similar sequences were 1.6 times more likely to have improved prediction performance compared with proteins that had fewer than 100 similar sequences. We also analyzed three different classes of RNA-binding amino acids: i) amino acids that contact only atoms of the RNA base, ii) amino acids that contact only the RNA backbone, and iii) amino acids that contact atoms in both the RNA base and backbone. We find that amino acids that contact only the RNA base are the most difficult to predict correctly (AUC of 0.66), and amino acids that contact both the RNA base and backbone are the easiest to predict (AUC of 0.87). The classifiers are available online at http://bindr.gdcb.iastate.edu/RNABindR.

INTRODUCTION

Interactions between protein and RNA are required for virtually every function of RNA. RNA molecules play important roles in all phases of protein production in the cell, performing tasks including carrying the message from DNA to the ribosome, catalyzing the addition of amino acids to a growing peptide chain, and regulating expression through siRNA pathways. RNA molecules carry the genetic information of many viruses. In every living system known today, RNA molecules function through interactions with proteins. These interactions may involve sequence-specific recognition, recognition of structural features of the RNA by the protein, and non-specific interactions. Consequently, understanding the sequence and structural determinants of protein-RNA interactions is important for therapeutic applications.

The most definitive way to verify RNA binding sites in proteins is to determine the structure of the complex by X-ray crystallography or NMR spectroscopy, but protein-RNA complexes have proven to be difficult to solve experimentally (Ke and Doudna, 2004). Determination of RNA binding sites in proteins is costly and time consuming, usually requiring site-directed mutagenesis and low-throughput RNA binding assays. 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 May 2008, there were 1406 protein-DNA complexes in the Protein Data Bank (Berman et al., 2000) and 661 protein-RNA complexes, of which only about 100 are unique complexes (Oubridge et al., Methods Mol Biol 2007).

A computational method to identify RNA binding sites in proteins is extremely valuable due to the difficulty in experimentally determining the interactions and their biological importance. A computational method allows for rapid identification of the most likely RNA binding sites and can shorten the experimental process for determining them. Ideally, the computational method would rely on information already available about the protein, such as the amino acid sequence. Our previous method for predicting RNA binding residues used only the amino acid sequence of the protein as input to a Naive Bayes classifier. This simple classifier provided reasonably good performance and has already been used to guide wet-lab experiments for determining actual RNA binding sites (Terribilini et al., 2006, Bechara et al. 2007, Sunita et al. 2007, Keren et al. 2008).

In addition to reducing the cost and effort of experimental investigations, a computational method for predicting RNA binding sites in proteins may provide insights into a recognition code for protein-RNA interactions. Several studies have been aimed at analyzing protein-RNA complexes to define and catalogue properties of RNA-binding sites (Jones et al., 2001, Treger and Westhof, 2001, Kim et al., 2003, Jeong et al., 2003, Lejuene et al., 2005, Kim et al., 2006, Morozova et al., 2005, Ellis et al., 2007, Bahadur et al., 2008). These studies are important because they catalogue the observed interaction patterns between proteins and RNA. The drawback of these analyses is that they are primarily descriptive and not predictive.

Recently, several studies have attempted to predict RNA binding sites in proteins using only amino acid sequence information (Jeong et al., 2004, Terribilini et al., 2006, Wang and Brown, 2006), using conservation information (Jeong and Miyano, 2006, Wang
and Brown, 2006, Kumar et al. 2007, Wang et al., 2008, Tong et al., 2008, Wang et al., 2008) or using information from the protein structure (Kim et al. 2006, Wang et al. 2008, Chen and Lim, 2008, Shulman-Peleg et al., 2008, Towfic et al., 2008). These studies have shown that evolutionary information in the form of position specific scoring matrices (PSSMs) significantly improves prediction performance over single sequence methods. However, they have yet to demonstrate an effective use of protein structure to significantly improve predictions over evolutionary information alone.

In this study, we compare several classifiers trained to predict RNA binding sites in proteins using information derived from a protein sequence or protein structure. The input to the classifier consists of an encoding of the target residue plus its sequence or structural neighbors. Each residue is encoded using its amino acid identity, or its PSSM (position-specific scoring matrix) profile obtained using multiple sequence alignment. We consider two classifiers, IDSeq and IDStruct, that encode each amino acid and its sequence or structural neighbors using the 20-letter amino acid alphabet; two classifiers, PSSMSeq and PSSMStruct, which encode each amino acid and its sequence or structural neighbors using their PSSMs; and Ensemble, a classifier that combines the other classifiers. In addition, we sought to answer the following questions: To what extent (if any) does the recent increase in the number of protein-RNA complexes available in the Protein Data Bank (PDB) contribute to improved prediction of RNA binding residues? What is the effect of the number of sequences used to construct the PSSMs on predictive performance? What is the effect (if any) of using protein structures extracted from protein-RNA complexes as opposed to unbound structures on the predicted protein-RNA interfaces? Are some types of protein-RNA contacts, e.g., contacts between a protein and an RNA base as opposed to contacts between a protein and an RNA backbone, easier to predict?

We show that classifiers that utilize information derived from both the protein structure and evolutionary considerations, i.e., PSSMs, each used individually as well as in
combination, have significantly better prediction performance relative to classifiers that are trained using only the amino acid identities of the target residues and its sequence neighbors. Our results show that the increase in the number of protein-RNA complexes in the PDB has not resulted in improvement in our ability to reliably predict protein-RNA interfaces. We also find that the number of sequences that are homologous to the target sequence plays an important role in determining whether the PSSM-based encoding improves predictive performance of the resulting classifier. We show that the structural neighborhood-based encoding of the input to the classifier is relatively insensitive to conformational changes in the protein upon binding to RNA. We find that amino acids that contact only the RNA base are the most difficult to predict and amino acids that contact both the RNA base and backbone are the easiest to predict. Taken together, our results demonstrate significant improvement in prediction of RNA binding residues and provide valuable insights into factors affecting the prediction of RNA binding residues. They also suggest several potential avenues for further improvements.

RESULTS

We compare several classifiers trained to predict RNA binding sites in proteins using information derived from a protein sequence or protein structure. The input to the classifier consists of an encoding of the target residue plus its sequence or structural neighbors. Each residue is encoded using its amino acid identity, or its PSSM (position-specific scoring matrix) obtained using PSI-BLAST. We consider two classifiers, IDSeq and IDStruct, that encode each amino acid and its sequence or structural neighbors using the 20-letter amino acid alphabet; two classifiers, PSSMSeq and PSSMStruct, which encode each amino acid and its sequence or structural neighbors using their PSSM profiles; and Ensemble, a classifier that combines the other classifiers.
**Information derived from protein structure improves prediction of RNA binding residues**

The sequence based classifiers, IDSeq and PSSMSeq, use a window of amino acid identities based on residues that are contiguous in the protein sequence. The protein sequence windows may not reliably capture the information that is important for RNA binding because the binding surface is defined by amino acids that are close together in the protein structure with no requirement for them to be close in the primary amino acid sequence. In an attempt to capture this structural context for predicting RNA binding sites, we constructed classifiers (IDStruct and PSSMStruct) that use the structural neighbors of an amino acid as its input.

By comparing the ROC curves of the IDSeq and IDStruct classifiers (Figure 4.1), we see that the IDStruct classifier outperforms the IDSeq classifier at all classification thresholds. For most given levels of false positive rate, the IDStruct classifier has ~5-10% improvement in true positive rate. Similarly, the specificity-sensitivity plot (Figure 4.2) shows that the IDStruct classifier achieves a higher specificity at any given level of sensitivity. For example, at a sensitivity of 50%, the IDSeq classifier achieves a specificity of 35%, while the IDStruct classifier has a sensitivity of 43%. The area under the ROC curve (AUC), a good overall measure of classifier performance, is 0.771 for the IDStruct classifier compared to 0.737 for the IDSeq classifier. When we compare the performance of the two classifiers at the "optimal" classification threshold (Table 4.1), we find that the IDStruct classifier has a 0.051 higher correlation coefficient, 6% higher specificity for RNA-binding residues, and 3% higher sensitivity for RNA-binding residues.
Figure 4.1 curves for all five classifiers on the RB181 dataset show that the Ensemble classifier provides the best prediction of RNA-binding sites. The Ensemble classifier dominates the figure, providing a higher true positive rate for any given false positive rate than any other classifier.
Figure 4.2 - sensitivity plot for all classifiers on the RB181 dataset. On highly unbalanced datasets like RB181, it is helpful to visualize classifier performance as specificity-sensitivity plots in addition to the ROC curves shown in Figure 1. The x-axis measures the sensitivity of prediction of RNA-binding residues, that is, the fraction of actual RNA-binding sites predicted as such by the classifier. The y-axis measures the specificity of prediction for RNA-binding residues, that is, the fraction of predicted RNA-binding sites that actually are RNA-binding. From this plot, we see that the IDSeq classifier achieves the lowest levels of specificity for any given sensitivity. Including structural or evolutionary information as input to the classifiers increases the level of specificity for any level of sensitivity. The ensemble classifier that combines the predictions of the IDStruct, PSSMSeq, and PSSMStruct classifiers achieves the highest levels of specificity for any sensitivity.

PSSM-based encoding of the target residue and its structural neighbors does not yield better prediction performance than PSSM-based encoding of the target residue and its
sequence neighbors. Figure 4.1 shows the ROC curve of the PSSMSeq and PSSMStruct classifiers, which are nearly identical. For any given false positive rate, the PSSMStruct classifier has less than 2% higher true positive rate. The sensitivity-specificity plot in Figure 4.2 shows that the PSSMStruct classifier has a slightly higher specificity for most levels of sensitivity. The AUC for the two classifiers, shown in Table 4.1, is nearly identical, with the PSSMSeq classifier achieving an AUC of 0.793 and the PSSMStruct classifier having an AUC of 0.798. This improvement of 0.005 is much smaller than the improvement of 0.034 observed between the IDSeq and IDStruct classifiers. This suggests that the PSSM-based encoding of structural neighbors is not more informative in predicting protein-RNA interfaces than the PSSM-based encoding of sequence neighbors.

Table 4.1 Average performance for each of the five classifiers at the classification threshold that maximized the correlation coefficient on the training set. Averages are taken over the 10 different runs of 10-fold cross validation with the standard deviation shown in parenthesis. The AUC is the best measure of overall classifier performance because it is not dependent on the classification threshold.

<table>
<thead>
<tr>
<th></th>
<th>IDSeq</th>
<th>IDStruct</th>
<th>PSSMSeq</th>
<th>PSSMStruct</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accuracy</strong></td>
<td>85.3 (0.11)</td>
<td>86.4 (0.15)</td>
<td>85.0 (0.1)</td>
<td>85.0 (0.29)</td>
<td>85.5 (0.33)</td>
</tr>
<tr>
<td><strong>Correlation Coefficient</strong></td>
<td>0.325 (0.003)</td>
<td>0.376 (0.004)</td>
<td>0.382 (0.003)</td>
<td>0.394 (0.004)</td>
<td>0.425 (0.003)</td>
</tr>
<tr>
<td><strong>Specificity+</strong></td>
<td>54 (0.8)</td>
<td>60 (1.2)</td>
<td>51 (0.4)</td>
<td>51 (1.3)</td>
<td>53 (1.4)</td>
</tr>
<tr>
<td><strong>Specificity-</strong></td>
<td>88 (0)</td>
<td>89 (0)</td>
<td>90 (0)</td>
<td>90 (0.3)</td>
<td>91 (0)</td>
</tr>
<tr>
<td><strong>Sensitivity+</strong></td>
<td>30 (0.3)</td>
<td>33 (0.7)</td>
<td>43 (0.7)</td>
<td>45 (0.8)</td>
<td>49 (1.9)</td>
</tr>
<tr>
<td><strong>Sensitivity-</strong></td>
<td>95 (0.3)</td>
<td>96 (0.3)</td>
<td>92 (0.5)</td>
<td>92 (0.6)</td>
<td>92 (0.7)</td>
</tr>
<tr>
<td><strong>F measure</strong></td>
<td>0.384 (0.002)</td>
<td>0.426 (0.004)</td>
<td>0.467 (0.003)</td>
<td>0.48 (0.004)</td>
<td>0.508 (0.005)</td>
</tr>
<tr>
<td><strong>AUC</strong></td>
<td>0.737 (0.001)</td>
<td>0.771 (0.001)</td>
<td>0.793 (0.001)</td>
<td>0.798 (0.001)</td>
<td>0.811 (0.001)</td>
</tr>
</tbody>
</table>
PSSM-based encoding of the target residue and its sequence or structural neighbors improves the prediction of RNA binding residues

Sequence conservation is believed to be correlated with functional or structurally important residues. We incorporated sequence conservation of amino acids using their PSSMs. PSSMs have been shown to improve prediction performance in a number of tasks including protein-protein interaction site prediction, protein-DNA interaction site prediction, and protein secondary structure prediction (Ofran and Rost, 2007, Ahmad and Sarai, 2005, Kuznetsov et al., 2007, Pollastri et al., 2002, Proteins). PSSMs have been used to improve prediction of RNA binding sites as well (Jeong and Miyano 2006, Kim et al 2006, Kumar et al., 2008, Wang et al., 2008, Tong et al., 2008).

In this work, we constructed support vector machine classifiers that utilize PSSM-based encoding of the target residue and its sequence or structural neighbors as input. The input to the SVM is a window of PSSM vectors for the target residue and the neighboring residues in the sequence in the case of the PSSMSeq classifier, or structural neighbors in the case of the PSSMStruct classifier. PSSM-based encoding dramatically improves prediction performance in both the sequence based and structure based classifiers. Figure 4.1 shows the ROC curves of the IDSeq and PSSMSeq classifiers, which achieve AUCs of 0.738 and 0.795 respectively. The PSSMSeq classifier achieves higher true positive rates for any given false positive rate, including up to ~10% higher true positive rates over a wide range of false positive rates. The specificity-sensitivity plot in Figure 4.2 also shows that the PSSMSeq classifier has a higher specificity for almost all levels of sensitivity. Evolutionary information also improves performance in the structure based classifiers, as shown by the ROC curves in Figure 4.1 of the IDStruct and PSSMStruct classifiers. The PSSMStruct classifier has an AUC of 0.798, compared to 0.771 for the IDStruct classifier, an improvement of 0.027. While not as large an improvement as we see in the sequence based classifiers, the PSSMStruct classifier still achieves true positive rates of ~5% higher for a
given false positive rate over the IDStruct classifier. This improvement in performance over the IDSeq and IDStruct classifiers demonstrates the utility of sequence conservation for prediction of RNA binding sites. Interestingly, the specificity-sensitivity plot in Figure 4.2 shows that at levels of sensitivity below 28%, the IDStruct classifier achieves higher specificity. Above 28% sensitivity, the PSSMStruct classifier has a higher specificity. This indicates that combining the information from the two classifiers may further improve prediction of RNA-binding sites.

**Combining sequence, structure, and evolutionary information improves prediction of RNA binding residues**

By examining the predictions on each target residue, we observed that our four individual classifiers (IDSeq, IDStruct, PSSMSeq, and PSSMStruct) often produce different predictions for the same target residue. We investigated the possibility of improving prediction performance by combining the output of the classifiers and creating an ensemble classifier. There are 11 possible combinations of the four individual classifiers. To create these ensemble classifiers, we used the predicted probability of RNA binding from the individual classifiers as input to a Naive Bayes classifier. In general, combining the IDSeq classifier with any other classifier improved performance over IDSeq alone, but did not significantly improve performance over the other classifier's individual performance. The best performance was given by an ensemble classifier combining the IDStruct, PSSMSeq, and PSSMStruct classifiers. The ROC curve for this ensemble classifier is also shown in Figure 4.1. The AUC of the ensemble classifier is 0.811, an improvement of at least 0.013 over the best individual classifier (PSSMStruct) and 0.074 higher than the IDSeq classifier. The ensemble classifier gives true positive rates of up to ~5% higher than the PSSMStruct classifier for a given false positive rate. The improvement over the IDSeq classifier is quite dramatic, with true positive rate of up to ~15% higher for the same false positive rate. Figure
4.2 shows the specificity-sensitivity plots for all classifiers. The ensemble classifier dominates this plot, giving a higher specificity at all levels of sensitivity. The lowest sensitivity achieved by the ensemble classifier is 29%, indicating that even at the most strict classification threshold the ensemble classifier provides good coverage of RNA-binding residues. These results convincingly show that using sequence conservation and protein structural information dramatically improves prediction of RNA binding sites.

Statistical analysis of the differences in observed performance of the different classifiers

To determine whether the prediction performance of each pair of classifiers was different, we performed a paired t-test on the AUC values of each pair of classifiers over the 10-fold cross validation runs. Table 4.2 shows the two-tailed p-values of each paired t-test. A p-value below 0.05 is generally considered to reflect a significant difference. Each pair of classifiers has significantly different AUCs. The PSSMSeq and PSSMStruct classifiers achieve very similar AUCs, 0.793 and 0.798 respectively, but the paired t-test shows that this small difference is significant with a p-value of 7.88e-04.
Table 4.2 The performance of each pair of classifiers is significantly different. The table shows the two-tailed p-values from a paired t-test on the AUC from each pair of classifiers. Each classifier was trained and tested using 10-fold cross validation based on protein sequences, in which the proteins in the RB181 dataset were randomly split into 10 different sets, 9 sets were used for training, and the remaining one set was used as the test set. This procedure was repeated 10 times, with different random splits each time. We computed the AUC for each classifier on each different cross-validation run and used a paired t-test to determine if the differences in AUC between each pair of classifiers over the cross-validation runs were significant. Typically, a p-value less than 0.05 is deemed a significant difference.

<table>
<thead>
<tr>
<th></th>
<th>IDSeq</th>
<th>IDStruct</th>
<th>PSSMSeq</th>
<th>PSSMStruct</th>
<th>Ensemble</th>
</tr>
</thead>
<tbody>
<tr>
<td>IDSeq</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>IDStruct</td>
<td>1.88E-15</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PSSMSeq</td>
<td>1.14E-12</td>
<td>7.10E-09</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PSSMStruct</td>
<td>1.34E-16</td>
<td>1.41E-13</td>
<td>7.88E-04</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Ensemble</td>
<td>1.04E-16</td>
<td>1.82E-14</td>
<td>3.37E-08</td>
<td>1.83E-08</td>
<td>-</td>
</tr>
</tbody>
</table>

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

In our previous work, we have used smaller non-redundant datasets. We have updated the datasets periodically to take advantage of the ever increasing number of protein-RNA complexes in the PDB. We now have 3 non-redundant datasets, RB109, RB147, and RB181 that were created using the same criteria, namely no more than 30% sequence identity between any two proteins and at least 3.5 Å resolution of the structure. We wanted to investigate the effect of increasing the size of the non-redundant training set on prediction performance. To do this, we trained all of our classifiers on each of the three datasets, and compared performance on each. Figure 4.3 shows the ROC curves for the ensemble classifier on each of the three datasets. The ROC curves are nearly identical, with AUCs of 0.822, 0.821, and 0.811 for the RB109, RB147, and RB181 datasets, respectively. Prediction
performance has become slightly worse as the size of the non-redundant datasets has increased, but this difference is very small. Similar results were seen for each of our four individual classifiers (data not shown).

**Figure 4.3** Prediction performance has not improved as the non-redundant dataset has gotten larger. The ROC curves for the ensemble classifier on three datasets, RB109, RB147, and RB181.

The comparison of performance on the three datasets is not straightforward because they contain different proteins. However, the three datasets share 68 proteins and for these 68, the only difference in prediction performance results from having different training sets. By comparing the performance on just these common proteins, we can determine whether the increased size of the training set improves performance. Figure 4.4 shows the ROC curves for the ensemble classifier evaluated on the 68 identical proteins from each dataset. The AUCs are 0.801, 0.805, and 0.806 when training on RB109, RB147, and RB181, respectively. Thus, when tested on these shared proteins, the increased size of the training
set does not significantly improve prediction performance. Similar results were obtained
with the four individual classifiers (data not shown). This result shows that the increasing
size of the non-redundant dataset has not yet provided improved prediction of RNA-binding
residues.

![Ensemble Identities ROC Curve](image)

Figure 4.4 Prediction performance has not increased as the size of the non-redundant dataset has
increased. ROC curves for the ensemble classifier on the 68 identical proteins in RB109, RB147, and
RB181 show that as the size of the available training set has increased, performance has not increased.

**How does the PSSM help?**

We have shown that PSSM inputs improve prediction of RNA-binding sites overall,
but some proteins in our dataset do not have better predictions when using the PSSM inputs.
The PSSMSeq classifier achieves a higher AUC than the IDSeq classifier for 79% (142 out
of 181) of the proteins in RB181. Of the proteins for which 100 sequences or less used to
build the PSSM, 50% (10 out of 20) had a lower AUC with the PSSMSeq classifier than the
IDSeq classifier. Proteins with over 100 similar sequences available had a higher AUC in 82% (132 out of 161) of the cases. This represents a 1.6 fold increase in the percentage of proteins that show increased AUC over the group that had 100 or fewer similar sequences. This result shows that having more homologous sequences available tends to improve the chances of getting better predictions with the PSSM-based classifiers.

The improvement in prediction performance using PSSM inputs is due to sequence conservation; we assume that functional residues, such as RNA-binding residues are more likely to be conserved through evolution than non-functional residues. The PSSM inputs contain this information as scores for substituting each amino acid at specific positions in the sequence. Favored substitutions are given a positive score, while disfavored substitutions are given a negative score. By looking at the input residues to which the PSSMSSeq classifier assigned the highest probability of RNA-binding, we attempted to find the substitutions that were highly indicative of RNA-binding sites. We computed the average and standard deviation of the PSSM values in the top 10% of residues as ranked by the probability of RNA-binding assigned by the PSSMSSeq classifier and compared the average with the value in the blosum62 matrix. We use the blosum62 matrix as the expected value for a given substitution and consider anything significantly higher than the blosum62 value as a favored substitution and anything significantly lower than the blosum62 value as a disfavored substitution. Figure 4.5 shows the substitutions for each amino acid that are most favored and disfavored. Each cell in the table is colored based on the difference between the observed mean in our PSSMs and the blosum62 value. Red cells indicate disfavored substitutions for RNA-binding sites, white cells indicate substitutions observed at the same level as in the blosum62 matrix, and blue cells indicate substitutions observed at a higher rate in our PSSMs than in blosum62. The substitutions that scored more highly in our PSSMs than in the blosum62 matrix are the positively charged residues Arg and Lys for the most hydrophobic residues and Arg and Lys for the negatively charged residues. Glu and Asp are
also favored in predicted RNA-binding residues. His also shows favorable substitutions, especially for hydrophobic residues. The major disfavored substitutions in predicted RNA-binding sites are hydrophobic residues for hydrophobic residues. Other highly disfavored substitutions include the negatively charged Asp and Glu for any of the hydrophilic residues.

Figure 4.5 Favored and disfavored substitutions in the top predicted RNA-binding sites. Boxes filled in red are substitutions that are observed to be more disfavored than the blosum62 value for that substitution. Boxes filled in blue are observed to be more favored than the blosum62 value for that substitution.
Structure-based classifiers are relatively insensitive to bound versus unbound structure inputs

Ellis and Jones (2008) recently compared a set of 12 proteins for which both the RNA-bound and unbound structures were available to determine the extent of conformational change in the protein upon RNA-binding. They found that many proteins had at least some movement upon RNA-binding. In light of this analysis, we investigated the effect of using the unbound form of the protein as input to our structure-based classifiers compared to using the bound form. Table 4.3 compares the AUC achieved by our structure-based classifiers using the bound and unbound forms of the protein as input. The proteins are grouped into three sets, as defined by Ellis and Jones: proteins with no significant conformational change upon RNA-binding, proteins with RNA-binding residues that undergo more movement upon RNA-binding than non-RNA-binding residues, and proteins with non-RNA-binding residues that move more than RNA-binding residues. For the four proteins with no movement upon RNA-binding, the unbound form of the protein provided a much higher AUC for two proteins, while the other two had slightly lower AUCs when using the unbound form as input. For the other proteins, there was not much difference in using the unbound form for six out of eight proteins, despite the fact that there were significant differences in the protein structures used as input. The two that did have a difference in AUC, showed a much better AUC using the bound form of the protein. From these results on such a small number of proteins, it is difficult to draw any conclusions about the validity of using structural information from the bound form of the protein. In this small set, we only observed a significant decrease in prediction performance in two out of twelve proteins when using the unbound form, indicating that our structure based methods may be tolerant of conformational changes in the protein.
Table 4.3 The AUC for the IDStruct classifier for each of the 12 pairs of proteins identified by Ellis and Jones (Ellis and Jones, Proteins, 2008) to have both bound and unbound structures in the PDB. The paired structures are listed sequentially, with the bound structure identifier and AUC value in bold italics. The unbound identifier and AUC value are in regular type. Column A lists the pairs that had no conformational changes upon RNA-binding, column B lists the pairs that had non-binding residues move more than RNA-binding residues, and column C lists the pairs that had RNA-binding residues move more than non-binding residues. The two pairs highlighted in gray saw a substantial decrease in AUC when using the unbound form of the protein as input. All other pairs had similar or increased prediction performance when using the unbound structure as input.

<table>
<thead>
<tr>
<th>Pair</th>
<th>A</th>
<th>AUC</th>
<th>B</th>
<th>AUC</th>
<th>C</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1GTF_L</td>
<td>0.63</td>
<td>1K8W_A</td>
<td>0.82</td>
<td>1M5Q_C</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>1QAW_C</td>
<td>0.67</td>
<td>1R3F_A</td>
<td>0.77</td>
<td>1NU4_A</td>
<td>0.68</td>
</tr>
<tr>
<td>2</td>
<td>1JBR_A</td>
<td>0.68</td>
<td>1M8W_B</td>
<td>0.78</td>
<td>1SDS_A</td>
<td>0.73</td>
</tr>
<tr>
<td></td>
<td>1AQZ_A</td>
<td>0.79</td>
<td>1IB2_A</td>
<td>0.77</td>
<td>1XBI_A</td>
<td>0.75</td>
</tr>
<tr>
<td>3</td>
<td>1U0B_B</td>
<td>0.69</td>
<td>1N78_A</td>
<td>0.56</td>
<td>1TFW_A</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>1LI5_B</td>
<td>0.68</td>
<td>1J09_A</td>
<td>0.55</td>
<td>1R89_A</td>
<td>0.63</td>
</tr>
<tr>
<td>4</td>
<td>1UVL_A</td>
<td>0.76</td>
<td>2BUL_A</td>
<td>0.39</td>
<td>2A8V_B</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>1HHS_A</td>
<td>0.75</td>
<td>1MSC_A</td>
<td>0.37</td>
<td>1A8V_A</td>
<td>0.65</td>
</tr>
</tbody>
</table>

Classifier performance on different types of protein-RNA contacts

Protein-RNA interactions can be either sequence-specific or non-specific. To determine which type of contacts our classifiers are identifying, we analyzed prediction performance on three different types of contacts: i) contacts between an amino acid and atoms in only the RNA base, ii) contacts between an amino acid and atoms in only the RNA backbone, and iii) contacts between an amino acid and atoms in both the RNA base and
backbone. Table 4.4 shows the AUC of the ensemble classifier on the three different sets of contacts as well as the overall AUC. On all RNA-binding residues, the ensemble classifier achieves an AUC of 0.812. On backbone only contacts, the ensemble classifier achieves an AUC of 0.813, essentially the same as the overall AUC. On contacts with only the base, the ensemble classifier achieves an AUC of 0.667, much lower than overall or backbone only contacts. On contacts with both the base and backbone, the ensemble classifier achieves an AUC of 0.87, significantly higher than any other category of contact. This trend in AUC, namely, highest AUC on base and backbone contacts, lowest AUC on base only contacts, held for all of the individual classifiers (data not shown).

Table 4.4 Ensemble classifier AUC on all contacts and three subsets of contacts: i) contacts with only the RNA backbone, ii) contacts with only the RNA base, iii) contacts with both the RNA base and backbone. Amino acids that only contact the RNA base are the hardest to predict, while amino acids that contact both base and backbone atoms of the RNA are the easiest to predict.

<table>
<thead>
<tr>
<th>Type of contact</th>
<th>Ensemble AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>All contacts</td>
<td>0.812</td>
</tr>
<tr>
<td>Backbone only</td>
<td>0.813</td>
</tr>
<tr>
<td>Base only</td>
<td>0.667</td>
</tr>
<tr>
<td>Base and backbone</td>
<td>0.87</td>
</tr>
</tbody>
</table>

Evaluating predictions in the context of sequence and three dimensional structures

Figure 4.6 depicts the actual and predicted RNA binding residues for 30S ribosomal protein S5 (PDB ID 2UUB_E). The region of the sequence shown in this figure contains three binding regions, residues 10-23, 41-57, and 77-83. All five of the classifiers provide some useful predictions for this protein with some important differences between the
predictions. The IDSeq classifier predicts quite a few false positive predictions and does not predict any RNA binding residues in the third RNA binding region. The IDStruct classifier makes no false positive predictions, but misses most of the actual RNA binding residues. The PSSMSeq classifier identifies all three RNA binding regions, but makes some false positive predictions. The PSSMStruct classifier identifies most of the residues in region 1, but only one residue in region 2 and none in region 3, but with fewer false positive predictions than the PSSMSeq classifier. The Ensemble classifier identifies many true positive residues in regions 1 and 2 with very few false positives, but misses region 3.

**Figure 4.6** Correlation of protein-RNA binding site predictions using different types of sequence and structure-derived information with actual interface residues. For 30S ribosomal protein S5 from *T. thermophilus* the amino acid sequence is shown at the top, corresponding to the structure of 2UUB chain E. The amino acids in bold white text on black background correspond to the residues found in contact with RNA. The positions marked with a ‘+’ and highlighted in green, red, yellow, pink, and cyan correspond to residues that are predicted to be RNA interface residues by IDSeq, IDStruct, PSSMSeq, PSSMStruct, and Ensemble respectively.

Figures 4.7, 4.8 and 4.9 show the predictions plotted on the three dimensional structure of three proteins. The first example, Figure 4.7, shows the 30S ribosomal protein S5, the same protein shown in Figure 4.6. True positive predictions are shown in red, false positives in blue, false negatives in yellow, and true negatives are shown in grey. In order to better compare the predictions, each classifier had the threshold set such that the specificity was constant. The IDSeq and IDStruct predictions (Panels A and C) show low sensitivity,
while the PSSMSeq, PSSMStruct, and Ensemble predictions (Panels B, D, and E) identify almost all of the actual RNA-binding residues. There are very few differences between the PSSMSeq, PSSMStruct, and Ensemble predictions for this protein, but all three of these classifiers show dramatically improved sensitivity and correlation coefficient over the IDSeq and IDStruct classifiers.
Figure 4.7 Predictions for 30S Ribosomal Protein S5 (PDB ID 2UUB_E). In all panels, true
positives are colored red, false positives are blue, false negatives are yellow, true negatives are grey.
Performance statistics for each classifier are shown below the images.

Figure 4.8 shows predictions on the signal recognition particle 19 kDa protein (PDB
ID 1JID_A). This is an interesting example in which the IDSeq and IDStruct classifiers
identify only a few RNA binding residues in the true RNA binding patch, but they identify
different residues. The PSSMSeq, PSSMStruct, and Ensemble predictions show dramatic improvement; many more true positive (red) residues are identified, and the false positive (blue) residues are adjacent to the actual RNA binding patch.

Figure 4.8 Predictions for Signal Recognition Particle 19kDa protein (PDB ID 1JID_A). In all panels, true positives are colored red, false positives are blue, false negatives are yellow, true negatives
are grey, RNA is shown in green wireframe. Performance statistics for each classifier are shown below the images.

Figure 4.9 shows predictions on translation initiation factor IF1 (PDB ID 1HR0_W). This protein was chosen as an “average” example since the correlation coefficient achieved by the Ensemble classifier is very close to the overall average of 0.43. Panel A shows that the IDSeq classifier only identifies a small number of true positives (red) while missing a large number of actual RNA binding residues (false negatives, shown in yellow). The Ensemble classifier predicts more true positives (red) while still limiting the number of false positives (blue) and the false positive residues are adjacent to actual binding residues. The PSSMStruct classifier achieves the highest sensitivity (79%) on this protein at the specificity level depicted in the figure.
Figure 4.9 Predictions for Initiation Factor IF1 (PDB ID 1HR0_W). In all panels, true positives are colored red, false positives are blue, false negatives are yellow, true negatives are grey. Performance statistics for each classifier are shown below the images.
DISCUSSION

In this work we exploit protein structural and sequence conservation information to improve prediction of RNA binding residues. Our previously published method used only sequence information and a simple representation for amino acids. While this method has proven useful in practice (Terribilini et al., 2006, Sunita et al., 2007, Bechara et al., 2007, Keren et al., 2008), there is obviously room for improvement. We developed classifiers that utilize protein structural information and sequence conservation information. The resulting classifiers, IDStruct, PSSMSeq, and PSSMStruct, all show significantly improved performance over the simple IDSeq classifier. We also created an ensemble classifier to take advantage of differential predictions from our individual classifiers. The ensemble classifiers gives the best overall predictive performance. Protein structural information in the form of nearest neighbors of a target residue captures more useful information about RNA binding than the protein sequence context alone. This is shown by the significantly improved performance (p-value of 1.88E-15) of the IDStruct classifier over the IDSeq classifier. Sequence conservation in the form of PSSMs provides more valuable information for RNA binding sites than either sequence identities or structural context alone. This is supported by the improved performance of the PSSMSeq classifier over the IDSeq and IDStruct classifiers, with p-values of 1.14E-12 and 7.10E-09 respectively. Combining protein structural information and sequence conservation information as was done for the PSSMStruct classifier provides another significant improvement, with a p-value of 7.88E-04 when compared to the AUC of PSSMSeq classifier. Combining sequence conservation information and protein structural information in the form of the ensemble classifiers gives another significant improvement (p-value of 1.83E-08) in predictive performance over the PSSMStruct method of combining these types of information.
**Dataset size**

We have attempted to exploit more information about protein-RNA interactions by creating a new dataset, RB181, which incorporates recently solved protein-RNA complexes. We now have three datasets created using the same criteria for redundancy and resolution, created approximately 2 years apart. The size of the non-redundant dataset has grown from 109 to 147 to 181 proteins as more complexes have been deposited in the PDB. However, RNA binding site prediction performance has not improved as the non-redundant dataset has increased in size. There are two possible explanations for this. First, there may not be additional generalizable signals in RNA-binding proteins that will allow for better predictions. Second, the coverage of all possible protein-RNA interactions is still too small to observe any further generalizable signals that may exist. We favor the second explanation because there are still only about 500 high resolution protein-RNA complexes in the PDB, and many of these are redundant. As coverage of the protein-RNA interaction space grows more complete, we expect to see improvements in prediction performance.

**Reasons for improvement with PSSM based classifiers**

The number of sequences used to build the PSSM is a good indicator of how much improvement can be gained by the PSSMSeq method over the IDSeq method. Proteins with relatively few similar sequences were much less likely to show improved predictions with the PSSM-based classifiers compared with proteins that had a large number of similar sequences. This indicates that some improvement in prediction performance can be expected simply from continued increase in the number of homologous protein sequences. We also analyzed the values in the PSSM for the top RNA-binding sites as ranked by our classifiers. We found that the PSSM-based classifiers utilize information in the PSSM that is "makes sense" in terms of expected structural effects of specific substitutions. In general, we observed that substitution of hydrophobic residues is highly disfavored in predicted RNA-binding sites.
while substitution of positively-charged residues for hydrophobic residues is highly favored. His also shows a number of highly favored substitutions, including the His for Trp substitution that may preserve stacking interactions with RNA. Interestingly, substituting Arg for Lys or Lys for Arg is not more favored in RNA binding sites than in the blosum62 matrix. From this we conclude that in many RNA-binding sites, positive charge is not sufficient for RNA-binding; rather there is a specific structural requirement for either Arg or Lys.

**Using bound versus unbound structures as input**

To determine whether our predictions were being unduly influenced by using structural data from the bound form of RNA-binding proteins, we examined prediction performance using structural information from unbound forms. Ellis and Jones (Ellis and Jones, 2008) analyzed 12 pairs of protein structures and found that 4 had almost no conformational changes upon RNA-binding, while 8 had conformational changes upon RNA-binding. We used these same 12 pairs of structures and observed that only 2 proteins had substantially better performance with the bound form of the protein. While based on a small number of examples, this indicates that our method of using structural data is relatively tolerant of conformational changes in the protein that may result from RNA-binding.

**Classification performance is best on residues that contact both the base and backbone of RNA**

We compared the performance of our classifiers on three different sets of RNA-binding residues: contacts with only the RNA backbone, contacts with only the RNA base, and contacts with both the RNA base and backbone. Contacts with the RNA backbone are non-specific, whereas contacts that involve the RNA base are potentially sequence-specific. Classification performance was best for residues that contacts with both the base and the backbone, one of the categories that is potentially sequence specific. Performance on the
backbone-only category is about the same as the overall performance. Classification on residues with base-only contacts was much more difficult, achieving the lowest performance. We analyzed the interaction propensities for each of these sets of contacts and found that they did not fluctuate significantly from the overall interaction propensity; thus, the performance differences are not due to differences in types of amino acids involved in the contacts (data not shown).

RNA binding residues that contact only the base are the most challenging to predict, which may be due in part to the small number of observed contacts in this category. Only 1226 residues in RB181 are in this category, which represents only 2.5% of amino acids in the dataset, and only 16% of RNA-binding residues. As the number of experimentally determined protein-RNA complexes grows, we expect improvement on this category of contacts. As we showed in the comparative analysis of performance on RB109, RB147, and RB181, we have not yet reached the point where the growth of the dataset has improved classification performance. This likely indicates that there are simply not enough known structures at present and the currently available structures are not diverse enough to provide improved predictions and a complete picture of all RNA-binding sites.

**Comparison with other methods**

Published methods for predicting RNA-binding amino acids can be classified into three groups: i) single-sequence methods (Jeong et al., 2004, Terribilini et al., 2006, Wang and Brown, 2006), ii) multiple-sequence methods (Jeong and Miyano, 2006, Wang and Brown, 2006, Kumar et al. 2007, Wang et al., 2008, Tong et al., 2008, Wang et al., 2008), and iii) methods that combine sequence and structure (Kim et al. 2006, Wang et al. 2008, Chen and Lim, 2008, Shulman-Peleg et al., 2008, Towfic et al., 2008). Single-sequence methods use only one protein sequence as input, while all of the multiple sequence methods published to date have used evolutionary information in the form of PSSMs. The most
common use of structural information has been adding the solvent accessible surface area and/or secondary structure state of each residue as part of the input to a classifier. This approach has resulted in small improvements in prediction performance over single-sequence or multiple-sequence inputs alone. Chen and Lim (Chen and Lim, 2008) developed a novel approach for the use of structural data: they rank residues in terms of conservation and electrostatic stabilization and then predict RNA-binding sites based on the highest scoring surface patch or cleft in each protein. This method is so far unique in the RNA-binding site prediction field in it that relies directly on analysis of properties RNA-binding sites rather than on machine learning. Also, it attempts to predict only a single patch and cleft in each protein, rather than giving a prediction for each residue in a protein.

In this work, we presented classifiers from all four classes of prediction methods. We have shown that the performance of the multiple-sequence based and structure-based methods are significantly better than single-sequence based methods. We further demonstrated that the performance of our Ensemble method for combining sequence and structure information is significantly better than that of our multiple-sequence based methods. The comparison performed by Kumar et al. (2007) showed that a classifier essentially identical to our PSSMSeq classifier outperforms all single-sequence methods. In that study, performance was essentially identical for different machine learning methods that used the same input information. Using the comparison of Kumar et al. as a basis, we conclude that the PSSMSeq classifier presented in the current work outperforms all single-sequence methods tested to date and provides performance equivalent to other multiple-sequence methods. The PRINTR method (Wang et al., 2008) is the only other study of which we are aware in which a combination of evolutionary and structural information has been used. A small improvement was observed when using a PSSM, rASA, and secondary structure as input over using the PSSM input alone. The Ensemble classifier developed in the current work also uses both evolutionary and structural information as input and shows a
similar small, but significant improvement over PSSM inputs alone. We conclude that our Ensemble classifier provides significantly better performance than existing multiple-sequence based methods and comparable performance to published methods that combine evolutionary and structural information.

METHODS

Datasets

RB181 dataset: RB181 was created by using the PISCES server (Wang and Dunbrack, 2003) to cull a set of proteins with <30% sequence identity and 3.5 Angstrom resolution or better from all protein-RNA complexes in the PDB. All residues in the protein structure were given a label, either “+” for RNA binding or “-“ for non-binding. To be included in the dataset, proteins had to have at least 40 amino acids and at least 3 RNA-binding amino acids. Also, the RNA in the complex had to have at least 5 nucleotides. Our previous work (Terribilini et al., 2006) used the program ENTANGLE (Allers and Shamoo, 2001) to define RNA binding residues. In this work we utilized a distance-based definition of RNA binding in which a residue was labeled RNA binding if any atom in the amino acid was within 5 Angstroms of any atom in the RNA. The resulting dataset has 181 protein chains with a total of 48,791 residues, of which 7456 are labeled as RNA binding. We further defined three non-overlapping sets of RNA binding residues: i) residues that are within 5 Angstroms of atoms in only the RNA base, ii) residues that are within 5 Angstroms of atoms in only the RNA backbone, and iii) residues that are within 5 Angstroms of atoms in both the RNA base and backbone. There were 1226 residues that contact only the RNA base, 4229 that contact only the RNA backbone, and 2001 that contact both the RNA base and backbone.

We also used two datasets from our previous work (Terribilini et al., 2006, Terribilini et al., 2007) RB147 and RB109, created using the same sequence identity and resolution
criteria as RB181 and using the 5 Angstrom distance cutoff definition for RNA-binding residues.

**Classifiers**

Input representations: In this study, we use two different encodings for amino acids. First, amino acid identity (ID) is simply the one letter abbreviation for each of the 20 amino acids. The second encoding is a position specific scoring matrix vector (PSSM) 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 0.0001 for inclusion in the next iteration. Sequences from the RB181 dataset were removed from the NCBI nr database before running PSI-BLAST to ensure that no sequences from the training set were used to build the PSSM inputs.

Residue context: We employ two methods for capturing the context of an amino acid within the protein. Sequence based windows: we use 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. We also define structure based windows in which the context of each amino acid is based on neighboring residues in the protein 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.

Naïve Bayes classifier: A Naïve Bayes (NB) classifier is based on Bayesian statistics and makes the simplifying assumption that all attributes are independent given the class label. Despite this assumption, NB classifiers have been shown to perform as well as or better than more sophisticated methods for many problems, even when the independence
assumption may be violated (Buntine, 1991). We used the NB implementation in the Weka package (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 Naive Bayes classifier assigns input $x$ the class label + (interface) if:

$$P(C = + \mid X = x) \geq \theta$$

and the class label – (non interface) otherwise. The choice of $\theta = 1$ corresponds to assigning the most probable class label. The desired trade-off of sensitivity against specificity can be achieved by varying $\theta$.

Because the inputs are assumed to be independent given the class, we have:

$$= P(C = +) \prod_{i=-n}^{i=n} P(X_i = x_i \mid C = +)$$

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

Support vector machine: A support vector machine (SVM) classifier uses a kernel function to map the inputs into a higher dimensional space (Boser, Guyon, and Vapnik, 1992, Cortes and Vapnik, 1995). The SVM then attempts to find the maximal margin hyperplane to separate the instances of members of the two classes, in this case RNA binding and non-RNA binding. The kernel function used in this study is the radial basis function (RBF) kernel. The RBF kernel function is:

$$K(z_i, z_j) = \exp(-\gamma \|z_i - z_j\|^2)$$

where $z_i$ and $z_j$ are input instances and $\gamma$ is a training parameter. SVM also has a regularization factor, $C$. Both $\gamma$ and $C$ were tested for values that maximized the
correlation coefficient on the training set. We used the LIBSVM implementation in this study (Chang and Lin, 2001, http://www.csie.ntu.edu.tw/~cjlin/libsvm).

Four types of individual classifiers: Using the above input representations, methods for determining the context of the target amino acid, and algorithms, we developed four different types of classifiers, IDSeq, IDStruct, PSSMSeq, and PSSMStruct. The IDSeq and IDStruct classifiers use the amino acid identity input representation and the NB algorithm. IDSeq uses sequence based windows while IDStruct uses structure based windows. PSSMSeq and PSSMStruct use the PSSM vector input representation and the SVM algorithm, with PSSMSeq using sequence based windows and PSSMStruct using structure based windows.

Ensemble classifiers: We also developed ensemble classifiers to combine the output of the individual classifiers and generate an ensemble prediction. We experimented with various types of ensemble classifiers including simple voting schemes, weighted voting schemes, and Naïve Bayes combinations. Results presented here used the NB ensemble classifier. The input to the NB ensemble classifier is the predicted probability of RNA binding from some or all of the individual classifiers. We tested all possible combinations of the four individual classifiers.

Cross Validation Procedure

In a single round of cross validation, m protein sequences are randomly chosen to be the test set and all other sequences are used to train the classifier. Cross validation based on protein sequences has been shown to be more rigorous than cross validation based on windows of protein sequence (Caragea et al., 2007). Windows-based cross validation has the potential to bias the classifier because portions of the test sequence are used in the training set. The classification threshold, $\theta$ is determined by applying different values for $\theta$ to the training set, from 0 to 1 in increments of 0.01. The value of $\theta$ that gave the highest
correlation coefficient on the training set is used to make predictions on the test set. This process is repeated until all protein sequences have been used in the test set. Here we report results 10 fold cross validation where \( m = \text{size of dataset}/10 \). Cross validation runs were performed 10 different times with different random splits of the training and test sequences used.

**Performance measures**

We used a number of different measures of classifier performance. 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), false negative (FN), or true negative (TN). A residue is a TP if the predicted and actual labels are RNA binding. A residue is a FP if the predicted label is RNA binding and the actual label is non-binding. A residue is a FN if the predicted label is non-binding and the actual label is RNA binding. A residue is a TN if the predicted and actual labels are non-binding. We report the following performance measures:

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

\[
\text{Specificity}^+ = \frac{TP}{TP + FP}
\]

\[
\text{Specificity}^- = \frac{TN}{TN + FN}
\]

\[
\text{Sensitivity}^+ = \text{True Positive Rate} = \frac{TP}{TP + FN}
\]

\[
\text{Sensitivity}^- = \frac{TN}{TN + FP}
\]
\[
\text{Correlation Coefficient} = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}}
\]

\[
\text{False Positive Rate} = \frac{FP}{FP + TN}
\]

All machine learning methods have an inherent trade-off between specificity and sensitivity that is controlled through the classification threshold. A useful method of comparing classifiers across all classification thresholds is the receiver operating characteristic (ROC) curve. A ROC curve plots the false positive rate against the true positive rate. The area under the ROC curve can be used to compare the total performance of classifiers. A perfect classifier would have an AUC of 1, while a classifier that makes random guesses would have an AUC of 0.5.

**Statistical Analysis**

To determine if the performance of our classifiers differ significantly, we performed paired t-tests between each pair of classifiers. We chose the AUC as the metric for performing the paired t-tests. The AUC of each classifier over each of the 10 cross validation experiments was used as the metric for the paired t-tests. We conclude that two classifiers have significantly different performance if the p-value from the paired t-test is less than 0.05.

**REFERENCES**

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


Chen, Y. C., and C. Lim (2008), Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry, *Nucleic acids research*.


Tong, J., P. Jiang, and Z. Lu (2008), RISP: A web-based server for prediction of RNA-binding sites in proteins, *Computer methods and programs in biomedicine*.


CHAPTER 5. PRIDB: PROTEIN-RNA INTERACTION DATABASE

Michael Terribilini, Jeff Ferguson, and Drena Dobbs

ABSTRACT

A detailed knowledge of protein-RNA interfaces will help in understanding fundamental biological processes such as protein synthesis and the regulation of gene expression. We have created the Protein-RNA Interface Database, PRIDB, a comprehensive collection of all protein-RNA complexes from the Protein Data Bank (PDB). From PRIDB, we have extracted a non-redundant set of 181 proteins (RB181) and analyzed structural and physiochemical properties of their protein-RNA interfaces. Based on this dataset, we calculated RNA binding propensities for individual amino acids using three different definitions of RNA-binding residues, based on van der Waals contacts, direct hydrogen bonds, or water-mediated hydrogen bonds and observed several differences in propensities based on different definitions. We also provide a systematic analysis of amino acid binding propensities for Watson-Crick versus non-Watson-Crick base paired ribonucleotides and find that the overall mode of binding to the different types of base pairs is very similar, whereas the amino acid binding propensities for single stranded ribonucleotides are more divergent. Analysis of the protein-RNA contacts for preferential interactions with main chain or side chain atoms of the amino acid or with base, phosphate, and ribose atoms of the RNA revealed distinct preferences for many amino acids to bind RNA with side chain atoms. The amino acids Asn and Gln show a clear preference for base-specific contacts. PRIDB allows for flexible definitions of RNA-binding residues with user-specified distance cutoffs, direct and water-mediated hydrogen bonds. PRIDB also allows for users to specify a set of complexes to analyze at various levels of detail, from whole residue interactions to detailed atomic contacts. PRIDB is available online at http://bindr.gdcb.iastate.edu/PRIDB.
INTRODUCTION

RNA is one of the most versatile molecules in biology, performing functions that range from storage of genetic information to enzymatic catalysis. Protein-RNA interactions play diverse roles in the cell and are especially important in all aspects of gene expression. A detailed understanding of protein-RNA binding will aid in understanding fundamental biological processes including protein synthesis, RNA splicing, and regulation of gene expression by RNA interference.

The amount of high resolution structural information regarding protein-RNA interactions has increased rapidly over the past few years and several studies have identified trends in how proteins and RNA interact (Treger and Westhof, 2001, Jones et al., 2001, Allers and Shamoo, 2001, Cheng et al., 2003, Kim et al., 2003, Jeong et al., 2003, Hoffman et al., 2004, Baker and Grant, 2005, Lejeune et al., 2005, Kim et al., 2006, Morozova et al., 2006, Ellis et al., 2007, Spirin et al., 2007, Bahadur et al., 2008). For example, positively charged and polar amino acids are generally favored for RNA-binding and large hydrophobic residues are disfavored. There has also been general agreement that van der Waals contacts are more numerous than hydrogen bonds, protein side chain contacts are more frequent than main chain contacts, and the majority of protein atom contacts are with the ribonucleotide backbone rather than the base atoms.

In this work, we build on previous studies by creating the Protein-RNA Interaction Database (PRIDB), a database of all protein-RNA complexes in the Protein Data Bank (PDB) (Berman et al., 2000). From PRIDB, we extracted a dataset of 181 non-redundant RNA binding proteins (RB181) and performed a detailed analysis of their protein-RNA interfaces. RB181 represents the largest set of proteins analyzed for protein-RNA interactions to date.

We define three different types of RNA-binding residues: i) RNA-binding residues identified using a distance-based definition in which an atom of an amino acid must be
within a specified distance cutoff of an atom in the RNA, ii) direct hydrogen bonds, and iii) water-mediated hydrogen bonds. We distinguish between Watson-Crick paired ribonucleotides, non-Watson-Crick paired ribonucleotides, and single stranded ribonucleotides and compare the recognition of each type by proteins. We analyze the binding sites in terms of contacts made by main chain atoms and side chains of the amino acids.

We find several interesting differences in binding propensities for specific amino acids examined in terms of distance-based contacts, direct hydrogen bonds, and water-mediated hydrogen bonds. We present a novel analysis of Watson-Crick and non-Watson-Crick base paired ribonucleotides and find that they have very similar binding characteristics, whereas single stranded ribonucleotides exhibit greater differences in amino acid binding propensities. We find substantial differences in binding propensities by main chain and side chain atoms. We also find that the amino acids asn and gln have a preference for interacting with base atoms over phosphate and ribose atoms and that asp prefers base atoms in water-mediated hydrogen bonds.

PRIDB is available online at http://bindr.gdeb.iastate.edu/PRIDB. It allows users to obtain detailed information regarding any single protein-RNA complex or groups of complexes from the PDB. PRIDB allows for flexible definitions of RNA-binding residues and user-specified levels of interaction analysis, from whole residue to detailed atomic level contacts.

METHODS

Protein-RNA Complexes Included in PRIDB

To select the protein-RNA complexes for inclusion in the database, we extracted all complexes from the PDB (Berman et al., 2000) as of February, 2008. The criteria used were that the complex had to contain protein and RNA; DNA was ignored. We limited the search
to complexes with a resolution of 3.5 Å or better. We filtered out complexes that did not contain at least 5 ribonucleotides of RNA and those that contained fewer than 20 amino acids. The remaining 503 complexes were all included in PRIDB.

The Non-Redundant RB181 Dataset

RB181 was created by using the PISCES server to cull a set of proteins with <30% sequence identity and 3.5 Å resolution or better from all protein-RNA complexes in the PDB. To be included in the dataset, proteins had to have at least 40 amino acids and at least 3 RNA-binding amino acids as defined below. The final criterion was that there had to be at least 5 ribonucleotides of RNA in the complex.

Definition of RNA-Binding Residues

We have included three different definitions of RNA-binding residues in PRIDB, amino acids with:

1. Direct hydrogen bonds to RNA atoms
2. Water-mediated hydrogen bonds to RNA atoms
3. Atoms within a specified distance cutoff of the RNA

Hydrogen bonds: The HBPLUS program (McDonald and Thornton, 1994) was used to calculate all possible hydrogen bonds in each complex in the database. We used the default parameters for distance cutoffs and angles when running HBPLUS. Direct hydrogen bonds are defined as those between any amino acid atom and any RNA atom. Water-mediated hydrogen bonds were defined as those in which a single water molecule forms hydrogen bonds with both an atom from the protein and an atom from the RNA. We listed all hydrogen bonds between an amino acid atom and a water molecule. We then identified any other hydrogen bonds involving the same water molecule. If that water molecule also formed a hydrogen bond with an atom from the RNA, we defined the water-mediated hydrogen bond between the amino acid and the ribonucleotide.
**Distance cutoff:** a residue was labeled RNA binding if any atom in the amino acid was within 5 Å of any atom in the RNA. PRIDB allows for user defined distance cutoffs as well.

**Interaction Propensities**

We calculated a number of different interaction propensities based on each type of contact discussed above. All propensity calculations were limited to surface residues as defined by having a relative solvent accessibility of at least 5% as calculated by Naccess. The general form of all of the propensity calculations is:

\[
\text{Propensity}_i = \log_2 \left( \frac{\% \text{ binding residues of type } i}{\% \text{ all residues of type } i} \right)
\]

as reported previously (Terribilini et al., 2006). Other studies have used a similar propensity function (Jones et al., 2001, Kim et al., 2003, Jeong et al., 2003) that does not include the log term. Without the log transformation, the propensity of overrepresented residues is greater than 1 and underrepresented residues have a propensity between 0 and 1, but the scale above 1 and below 1 is not even. Reporting the final propensity value as the log of the ratio term, the scale is equivalent for over- and under-represented residues, with overrepresented residues having a propensity greater than 0 and underrepresented residues having a propensity less than 0.

We calculate overall propensities for each amino acid binding to any ribonucleotide as:

\[
\text{Propensity}_i = \log_2 \left( \frac{n_i / \sum n_j}{N_i / \sum N_j} \right)
\]
Where $n_i$ and $n_j$ are the number of amino acids of type $i$ or $j$ bound to RNA and $N_i$ and $N_j$ are the total number of amino acids of type $i$ or $j$.

We calculate propensities for binding of amino acid $i$ to a ribonucleotide $b$ as:

$$
\text{Propensity}_{ia} = \log_2 \left( \frac{n_{ia} / \sum_{j,b} n_{jb}}{N_i / \sum_j N_j \cdot N_a / \sum_b N_b} \right)
$$

Where $n_{ia}$ is the number of amino acid $i$ bound to ribonucleotide $a$, the sum $n_{jb}$ is the total number of amino acids bound to any ribonucleotide, $N_i$ and $N_j$ are the same as above, $N_a$ is the number of ribonucleotide $a$, and the sum $N_b$ is the total number of ribonucleotides. The \textit{counts} of amino acids interacting with RNA are based on the number of interacting residues, not the total number of interactions. For example, a single arg residue forming two hydrogen bonds to a ribonucleotide was only counted once for the propensity calculations. Therefore, the numbers presented throughout this work indicate how many \textit{residues} are involved in each type of interaction, not the total number of bonds involved.

Similar propensities were calculated for interactions between the main chain of each amino acid and each of the four ribonucleotides, and between the side chain of the amino acids and each ribonucleotide. We also calculated the propensity for interactions between each amino acid and the base, phosphate, and ribose portions of the ribonucleotides. Finally, we calculated the propensities for interactions between the main chain atoms of the amino acids and the base, phosphate, and ribose atoms of each ribonucleotide, and between the side chain atoms and the base, phosphate, and ribose atoms of each ribonucleotide. In addition to the propensities, we present the number of each interaction type, because the propensity calculation hides the raw frequencies of individual types. If fact, for many specific classes of bonding interactions, few or no examples were observed.
Other studies have defined the propensity for binding based on buried surface area rather than counts of interactions. Bahadur et al. (Bahadur et al., 2008) found that the buried surface area in protein-RNA interfaces is linearly correlated ($R^2 = 0.92$) with the number of residues in the interface. We found that interaction propensities are essentially equivalent when calculated using count-based versus buried surface area based definitions; thus only count-based propensities are presented in this work.

**Base Pairs**

Base pairs in RNA are defined and differentiated based on their hydrogen bonding patterns. Using hydrogen bond information obtained from HBPLUS (MacDonald and Thornton, 1994), we defined each ribonucleotide as either double stranded (DS) or single stranded (SS) as follows: A ribonucleotide in which at least one base atom is involved in a hydrogen bond with another ribonucleotide’s base atom(s) is defined as double stranded. If none of its base atoms form hydrogen bonds with base atoms of any other ribonucleotides, a ribonucleotide is defined as single stranded. We further differentiated between Watson-Crick base pairs and non-Watson-Crick base pairs. A ribonucleotide is defined to be in a Watson-Crick (WC) base pair if it meets the canonical WC hydrogen bonding pattern of either A-U (A N6 with U O4 and A N1 with U N3) or G-C (G O6 with C N4, G N1 with C N3, and G N2 with C O2). If a ribonucleotid e is defined as double stranded, but does not meet the WC standards, we define it to be in a non WC base pair.

**RESULTS**

We have created a database of all protein-RNA complexes from the PDB and analyzed protein-RNA interactions in a non-redundant set of 181 proteins, RB181. We present an analysis of protein-RNA interactions based on three different types of contacts: direct hydrogen bonds, water-mediated hydrogen bonds, and distance-based contacts. We distinguish between binding of amino acids to single stranded, Watson-Crick paired, and non
Watson-Crick paired ribonucleotides. We further analyze the specific atoms involved in contacts between amino acids and ribonucleotides by differentiating between contacts made by main chain and side chain atoms of amino acids, and base, phosphate, and ribose atoms of ribonucleotides.

Table 5.1 summarizes the distribution of ribonucleotides and bond types in RNA chains of interfaces in the RB181 dataset. G is the most frequent ribonucleotide in the RB181 dataset (32.5%), while U is the least frequent (18.5%). 46.6% of ribonucleotides in RB181 participate in canonical Watson-Crick base pairs, with CG pairs making up 75% of the Watson-Crick base pairs in the dataset. 31.7% of the ribonucleotides in RB181 are involved in non-Watson-Crick base pairs, while 21.7% are single stranded.
Table 5.1 Summary statistics for the ribonucleotides in the RB181 dataset. Counts of the ribonucleotides are presented with percentages in parentheses. Row 1 shows the counts of each ribonucleotide along with the percent of total ribonucleotides. The next three rows show the secondary structure state of each ribonucleotide, WC for Watson-Crick base paired ribonucleotides, nonWC for non-Watson-Crick based paired ribonucleotides, and SS for single stranded ribonucleotides. The percentages show the percent of each individual ribonucleotide that is in the secondary structure state, for example, A is seen 6154 times in RB181, of which 1484 are in WC base pairs, therefore 1484/6154 or 24.1% of the A’s in RB181 are in WC base pairs. The final three rows show the count and percentages of each ribonucleotide being involved in an interaction with protein. The distance row shows the number of ribonucleotides of each type that are within 5 Å of any atom in the protein, the H-bonds row shows the number of ribonucleotides involved in direct hydrogen bonds with the protein, and HOH shows the number of ribonucleotides involved in water-mediated hydrogen bonds with the protein.

<table>
<thead>
<tr>
<th></th>
<th>A</th>
<th>C</th>
<th>G</th>
<th>U</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Occurences</td>
<td>6154</td>
<td>6766</td>
<td>8568</td>
<td>4886</td>
<td>26374</td>
</tr>
<tr>
<td></td>
<td>(23.3%)</td>
<td>(25.7%)</td>
<td>(32.5%)</td>
<td>(18.5%)</td>
<td>(100%)</td>
</tr>
<tr>
<td>Double Stranded</td>
<td>3992</td>
<td>5726</td>
<td>7449</td>
<td>3479</td>
<td>20646</td>
</tr>
<tr>
<td></td>
<td>(64.9%)</td>
<td>(84.6%)</td>
<td>(86.9%)</td>
<td>(71.2%)</td>
<td>(78.3%)</td>
</tr>
<tr>
<td>WC</td>
<td>1484</td>
<td>4658</td>
<td>4665</td>
<td>1483</td>
<td>12290</td>
</tr>
<tr>
<td></td>
<td>(24.1%)</td>
<td>(68.8%)</td>
<td>(54.4%)</td>
<td>(30.4%)</td>
<td>(46.6%)</td>
</tr>
<tr>
<td>nonWC</td>
<td>2508</td>
<td>1068</td>
<td>2784</td>
<td>1996</td>
<td>8356</td>
</tr>
<tr>
<td></td>
<td>(40.8%)</td>
<td>(15.8%)</td>
<td>(32.5%)</td>
<td>(40.9%)</td>
<td>(31.7%)</td>
</tr>
<tr>
<td>Single Stranded</td>
<td>2162</td>
<td>1040</td>
<td>1119</td>
<td>1407</td>
<td>5728</td>
</tr>
<tr>
<td></td>
<td>(35.1%)</td>
<td>(15.4%)</td>
<td>(13.1%)</td>
<td>(28.8%)</td>
<td>(21.7%)</td>
</tr>
<tr>
<td>Within 5 Å</td>
<td>2225</td>
<td>2455</td>
<td>2776</td>
<td>1864</td>
<td>9320</td>
</tr>
<tr>
<td></td>
<td>(36.2%)</td>
<td>(36.3%)</td>
<td>(32.4%)</td>
<td>(38.1%)</td>
<td>(35.3%)</td>
</tr>
<tr>
<td>H-bonds</td>
<td>698</td>
<td>809</td>
<td>990</td>
<td>589</td>
<td>3086</td>
</tr>
<tr>
<td></td>
<td>(11.3%)</td>
<td>(12%)</td>
<td>(11.6%)</td>
<td>(12.1%)</td>
<td>(11.7%)</td>
</tr>
<tr>
<td>HOH</td>
<td>483</td>
<td>483</td>
<td>617</td>
<td>443</td>
<td>2026</td>
</tr>
<tr>
<td></td>
<td>(7.8%)</td>
<td>(7.1%)</td>
<td>(7.2%)</td>
<td>(9%)</td>
<td>(7.7%)</td>
</tr>
</tbody>
</table>

Using a distance-based definition to identify interface residues, 35% of ribonucleotides in the analyzed complexes are in contact with atoms of the bound protein. 11.7% of the ribonucleotides are involved in direct hydrogen bonds to protein atoms, and
7.7% have water-mediated hydrogen bonds to protein atoms. The percentage of each ribonucleotide involved in each type of contact is approximately equal, with a few exceptions. Only 32.4% of G ribonucleotides are bound to protein using the distance-based definition, slightly lower than the overall average of 35.3%, while 38.1% of U’s are bound. 9% of U’s are involved in water-mediated hydrogen bonds, slightly higher than the 7.7% overall average.

Table 5.2 summarizes the characteristics of the amino acids in the RB181 dataset. The RB181 dataset contains 48,791 amino acids, of which 36,487 are surface residues (rASA of greater than 5%). 7084 residues (19.4% or surface residues) are in contact with RNA based on the distance cutoff definition, 2585 (7.1%) have direct hydrogen bonds to RNA, and 1451 (4%) had water-mediated hydrogen bonds to RNA. We also analyzed the interaction propensities based on the RNA-binding definition used, the secondary structure state of the ribonucleotides, and whether the interactions involved the main chain or side chain of the amino acid and the base, phosphate, or ribose atoms of the ribonucleotide.
Table 5.2 RNA-binding residue counts, percentages, and propensities. The first column lists each amino acid, the Total column gives the total number of occurrences in RB181, the Surface column lists the number of residues with rASA of at least 5%, Distance lists the number of residues with any atoms within 5 Å of any RNA atom, HB lists the number of residues making a direct hydrogen bond with RNA, HOH lists the number of residues making a water-mediated hydrogen bond with RNA. The % columns give the percentage of surface residues in each type of contact with RNA. The Prop columns give the binding propensity for each type of contact with RNA.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>Surface</th>
<th>Distance</th>
<th>%</th>
<th>Prop</th>
<th>HB</th>
<th>%</th>
<th>Prop</th>
<th>HOH</th>
<th>%</th>
<th>Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>3699</td>
<td>2190</td>
<td>371</td>
<td>16.9%</td>
<td>-0.20</td>
<td>77</td>
<td>3.5%</td>
<td>-1.01</td>
<td>48</td>
<td>2.2%</td>
<td>-0.86</td>
</tr>
<tr>
<td>Cys</td>
<td>489</td>
<td>220</td>
<td>35</td>
<td>15.9%</td>
<td>-0.29</td>
<td>11</td>
<td>5.0%</td>
<td>-0.50</td>
<td>11</td>
<td>5.0%</td>
<td>0.33</td>
</tr>
<tr>
<td>Asp</td>
<td>2650</td>
<td>2411</td>
<td>332</td>
<td>13.8%</td>
<td>-0.50</td>
<td>117</td>
<td>4.9%</td>
<td>-0.55</td>
<td>112</td>
<td>4.6%</td>
<td>0.22</td>
</tr>
<tr>
<td>Glu</td>
<td>3883</td>
<td>3693</td>
<td>334</td>
<td>9.0%</td>
<td>-1.10</td>
<td>115</td>
<td>3.1%</td>
<td>-1.19</td>
<td>83</td>
<td>2.2%</td>
<td>-0.82</td>
</tr>
<tr>
<td>Phe</td>
<td>1879</td>
<td>1058</td>
<td>174</td>
<td>16.4%</td>
<td>-0.24</td>
<td>13</td>
<td>1.2%</td>
<td>-2.53</td>
<td>15</td>
<td>1.4%</td>
<td>-1.49</td>
</tr>
<tr>
<td>Gly</td>
<td>3403</td>
<td>2641</td>
<td>584</td>
<td>22.1%</td>
<td>0.19</td>
<td>183</td>
<td>6.9%</td>
<td>-0.03</td>
<td>119</td>
<td>4.5%</td>
<td>0.18</td>
</tr>
<tr>
<td>His</td>
<td>1096</td>
<td>915</td>
<td>239</td>
<td>26.1%</td>
<td>0.43</td>
<td>95</td>
<td>10.4%</td>
<td>0.55</td>
<td>56</td>
<td>6.1%</td>
<td>0.62</td>
</tr>
<tr>
<td>Ile</td>
<td>2754</td>
<td>1383</td>
<td>193</td>
<td>14.0%</td>
<td>-0.48</td>
<td>26</td>
<td>1.9%</td>
<td>-1.91</td>
<td>12</td>
<td>0.9%</td>
<td>-2.20</td>
</tr>
<tr>
<td>Lys</td>
<td>3461</td>
<td>3375</td>
<td>870</td>
<td>25.8%</td>
<td>0.41</td>
<td>444</td>
<td>13.2%</td>
<td>0.89</td>
<td>173</td>
<td>5.1%</td>
<td>0.37</td>
</tr>
<tr>
<td>Leu</td>
<td>4506</td>
<td>2421</td>
<td>331</td>
<td>13.7%</td>
<td>-0.51</td>
<td>37</td>
<td>1.5%</td>
<td>-2.21</td>
<td>38</td>
<td>1.6%</td>
<td>-1.34</td>
</tr>
<tr>
<td>Met</td>
<td>1055</td>
<td>610</td>
<td>122</td>
<td>20.0%</td>
<td>0.04</td>
<td>25</td>
<td>4.1%</td>
<td>-0.79</td>
<td>14</td>
<td>2.3%</td>
<td>-0.79</td>
</tr>
<tr>
<td>Asn</td>
<td>1655</td>
<td>1469</td>
<td>332</td>
<td>22.6%</td>
<td>0.22</td>
<td>166</td>
<td>11.3%</td>
<td>0.67</td>
<td>102</td>
<td>6.9%</td>
<td>0.80</td>
</tr>
<tr>
<td>Pro</td>
<td>2405</td>
<td>2002</td>
<td>313</td>
<td>15.6%</td>
<td>-0.31</td>
<td>29</td>
<td>1.4%</td>
<td>-2.29</td>
<td>30</td>
<td>1.5%</td>
<td>-1.41</td>
</tr>
<tr>
<td>Gln</td>
<td>1684</td>
<td>1518</td>
<td>317</td>
<td>20.9%</td>
<td>0.11</td>
<td>145</td>
<td>9.6%</td>
<td>0.43</td>
<td>84</td>
<td>5.5%</td>
<td>0.48</td>
</tr>
<tr>
<td>Arg</td>
<td>3446</td>
<td>3253</td>
<td>1091</td>
<td>33.5%</td>
<td>0.79</td>
<td>606</td>
<td>18.6%</td>
<td>1.39</td>
<td>260</td>
<td>8.0%</td>
<td>1.01</td>
</tr>
<tr>
<td>Ser</td>
<td>2459</td>
<td>1954</td>
<td>426</td>
<td>21.8%</td>
<td>0.17</td>
<td>197</td>
<td>10.1%</td>
<td>0.51</td>
<td>111</td>
<td>5.7%</td>
<td>0.51</td>
</tr>
<tr>
<td>Thr</td>
<td>2399</td>
<td>1870</td>
<td>380</td>
<td>20.3%</td>
<td>0.07</td>
<td>161</td>
<td>8.6%</td>
<td>0.28</td>
<td>75</td>
<td>4.0%</td>
<td>0.01</td>
</tr>
<tr>
<td>Val</td>
<td>3658</td>
<td>1882</td>
<td>283</td>
<td>15.0%</td>
<td>-0.37</td>
<td>20</td>
<td>1.1%</td>
<td>-2.74</td>
<td>43</td>
<td>2.3%</td>
<td>-0.80</td>
</tr>
<tr>
<td>Trp</td>
<td>561</td>
<td>391</td>
<td>84</td>
<td>21.5%</td>
<td>0.15</td>
<td>19</td>
<td>4.9%</td>
<td>-0.54</td>
<td>16</td>
<td>4.1%</td>
<td>0.04</td>
</tr>
<tr>
<td>Tyr</td>
<td>1649</td>
<td>1231</td>
<td>273</td>
<td>22.2%</td>
<td>0.19</td>
<td>99</td>
<td>8.0%</td>
<td>0.18</td>
<td>49</td>
<td>4.0%</td>
<td>0.00</td>
</tr>
<tr>
<td>Total</td>
<td>48791</td>
<td>36487</td>
<td>7084</td>
<td>19.4%</td>
<td></td>
<td>2585</td>
<td>7.1%</td>
<td></td>
<td>1451</td>
<td>4.0%</td>
<td></td>
</tr>
</tbody>
</table>

Comparison of the frequency and propensity of distance-based binding residues, direct hydrogen bonds, and water-mediated hydrogen bonds
The interaction propensities for each amino acid with any ribonucleotide using all three definitions of RNA-binding are shown in Figure 5.1, with the counts of interactions shown in Table 5.2. The positively charged residue arg has the highest interaction propensity of any amino acid for all three types of contacts. Other residues with positive RNA-binding propensities for all three types of contacts are lys, his, asn, gln, and ser. The highly polar side chain of asn is a preferred way for proteins to interact with RNA through water-mediated hydrogen bonds, with a propensity of 0.80, compared to 0.67 for direct hydrogen bonds, and only 0.22 for distance-based interactions. His and gln also show their highest propensity for RNA-binding through water-mediated hydrogen bonds compared to the other two types of contacts. Interestingly, the negatively charged aspartic acid has a positive propensity for water-mediated hydrogen bonding of 0.22, with a total of 112 water-mediated hydrogen bonds observed, the third highest total by counts of any of the amino acids. A water-mediated hydrogen bond is able to overcome the unfavorable charge interaction, whereas a direct hydrogen bond between aspartic acid and RNA is unfavorable, with a propensity of -0.55.
Figure 5.1 Interface propensities for binding any ribonucleotide for distance-based, direct hydrogen bonds, and water-mediated hydrogen bonds definitions of RNA-binding residues.

In general, the hydrophobic amino acids have negative RNA-binding propensities. This is not simply due to burial in the interior of the protein because this analysis is limited to residues on the protein surface. The general trend is for hydrophobic residues to have slightly negative propensities for distance-based contacts, more negative propensities for water-mediated hydrogen bonds, and the lowest propensities for direct hydrogen bonds. Trp and tyr actually have positive propensities for distance-based contacts due to their ability to form stacking interactions with ribonucleotides.

Figure 5.2 shows the interaction propensities for each amino acid with the four ribonucleotides. From these propensities we observe that certain amino acids have preferences for specific bases. For favorable distance-based contacts (Figure 5.2A), the largest difference in propensity for the different bases is shown by his. The his-A pair has a propensity of 0.72, whereas the contact propensity for his with C, G, and U is at most 0.47.
For unfavorable distance-based contacts, the most striking preference is shown by val, with a propensity of -1.1 for guanine, while propensities for the other three ribonucleotides are between -0.18 and -0.27.

Figure 5.2 Interaction propensities for each amino acid with each ribonucleotide. Panel A shows the propensities based on the distance-based definition of RNA-binding residues. Panel B shows direct hydrogen bonding propensities. Panel C shows water-mediated hydrogen bonding propensities.

For contacts via direct hydrogen bonds (Figure 5.2B), his favors cytosine, and asn favors uracil. Gln has positive propensities for hydrogen bonding to C, G, and U, but a
slightly negative propensity for A. Ser also shows positive propensities for three out of four
ribonucleotides, with the only unfavored interaction being with U.

Water-mediated hydrogen bonds (Figure 5.2C) show the most variation in propensity
between the ribonucleotides. His is an interesting example; for direct hydrogen bonds, his
prefers C, but in water-mediated hydrogen bonds, the his C pair is the only one of the four
ribonucleotides with a negative propensity. The amino acids gly, his, lys, asn, gln, ser, thr,
trp, and tyr all have their highest propensity for water-mediated hydrogen bonds with uracil.

**Interactions with single stranded, Watson-Crick, and non Watson-Crick paired
ribonucleotides**

We classified each ribonucleotide as being single stranded, Watson-Crick base paired,
or non-Watson-Crick base paired based on its hydrogen bonds to other bases. Single
stranded bases had no hydrogen bonds to other bases, Watson-Crick paired bases had all of
the hydrogen bonds required by the Watson-Crick base pairing rules, and non-Watson-Crick
paired bases had at least one hydrogen bond to another base. Our dataset contains 5728
single stranded bases, 12,290 bases in Watson-Crick base pairs, and 8356 bases in non
Watson-Crick base pairs (see Table 5.1).

Figure 5.3 shows the interaction propensities for each amino acid with single
stranded, Watson-Crick, and non Watson-Crick paired ribonucleotides for all three types of
interactions. The counts of observations of each type of interaction along with the
propensities are shown in Table 5.3 for distance-based interactions, Table 5.4 for direct
hydrogen bonds, and Table 5.5 for water-mediated hydrogen bonds.
Figure 5.3 Interface propensities for each amino acid with ribonucleotides in Watson-Crick base pairs (WC), non-Watson-Crick base pairs (nonWC), and single stranded ribonucleotides (SS). Panel A shows propensities calculated using the distance-based definition of RNA-binding residues, Panel B shows direct hydrogen bonding propensities, and Panel C shows water-mediated hydrogen bonding propensities.
Table 5.3 Distance-based counts and propensities for each amino acid binding to Watson-Crick (WC), non-Watson-Crick (nonWC), and single stranded (SS) ribonucleotides.

<table>
<thead>
<tr>
<th></th>
<th>WC Bound</th>
<th>WC Prop</th>
<th>nonWC Bound</th>
<th>nonWC Prop</th>
<th>SS Bound</th>
<th>SS Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>153</td>
<td>-0.34</td>
<td>107</td>
<td>-0.24</td>
<td>189</td>
<td>-0.04</td>
</tr>
<tr>
<td>Cys</td>
<td>10</td>
<td>-0.96</td>
<td>5</td>
<td>-1.34</td>
<td>21</td>
<td>0.11</td>
</tr>
<tr>
<td>Asp</td>
<td>141</td>
<td>-0.60</td>
<td>70</td>
<td>-0.99</td>
<td>165</td>
<td>-0.37</td>
</tr>
<tr>
<td>Glu</td>
<td>141</td>
<td>-1.21</td>
<td>68</td>
<td>-1.65</td>
<td>156</td>
<td>-1.07</td>
</tr>
<tr>
<td>Phe</td>
<td>62</td>
<td>-0.59</td>
<td>50</td>
<td>-0.29</td>
<td>98</td>
<td>0.06</td>
</tr>
<tr>
<td>Gly</td>
<td>256</td>
<td>0.13</td>
<td>187</td>
<td>0.30</td>
<td>230</td>
<td>-0.03</td>
</tr>
<tr>
<td>His</td>
<td>110</td>
<td>0.44</td>
<td>76</td>
<td>0.53</td>
<td>106</td>
<td>0.39</td>
</tr>
<tr>
<td>lle</td>
<td>74</td>
<td>-0.73</td>
<td>53</td>
<td>-0.59</td>
<td>103</td>
<td>-0.25</td>
</tr>
<tr>
<td>Lys</td>
<td>468</td>
<td>0.65</td>
<td>310</td>
<td>0.67</td>
<td>353</td>
<td>0.24</td>
</tr>
<tr>
<td>Leu</td>
<td>142</td>
<td>-0.59</td>
<td>83</td>
<td>-0.75</td>
<td>162</td>
<td>-0.41</td>
</tr>
<tr>
<td>Met</td>
<td>50</td>
<td>-0.11</td>
<td>28</td>
<td>-0.33</td>
<td>69</td>
<td>0.35</td>
</tr>
<tr>
<td>Asn</td>
<td>158</td>
<td>0.28</td>
<td>98</td>
<td>0.21</td>
<td>143</td>
<td>0.13</td>
</tr>
<tr>
<td>Pro</td>
<td>154</td>
<td>-0.20</td>
<td>95</td>
<td>-0.28</td>
<td>134</td>
<td>-0.41</td>
</tr>
<tr>
<td>Gln</td>
<td>157</td>
<td>0.22</td>
<td>72</td>
<td>-0.28</td>
<td>134</td>
<td>-0.01</td>
</tr>
<tr>
<td>Arg</td>
<td>550</td>
<td>0.93</td>
<td>393</td>
<td>1.07</td>
<td>474</td>
<td>0.72</td>
</tr>
<tr>
<td>Ser</td>
<td>178</td>
<td>0.04</td>
<td>129</td>
<td>0.19</td>
<td>196</td>
<td>0.18</td>
</tr>
<tr>
<td>Thr</td>
<td>163</td>
<td>-0.02</td>
<td>107</td>
<td>-0.01</td>
<td>179</td>
<td>0.11</td>
</tr>
<tr>
<td>Val</td>
<td>120</td>
<td>-0.47</td>
<td>72</td>
<td>-0.59</td>
<td>129</td>
<td>-0.37</td>
</tr>
<tr>
<td>Trp</td>
<td>42</td>
<td>0.28</td>
<td>29</td>
<td>0.36</td>
<td>36</td>
<td>0.05</td>
</tr>
<tr>
<td>Tyr</td>
<td>100</td>
<td>-0.12</td>
<td>73</td>
<td>0.04</td>
<td>158</td>
<td>0.53</td>
</tr>
<tr>
<td>Total</td>
<td>3229</td>
<td>2105</td>
<td>3235</td>
<td>2772</td>
<td>2975</td>
<td>3235</td>
</tr>
</tbody>
</table>
Table 5.4 Direct hydrogen bond counts and propensities for each amino acid binding to Watson-Crick (WC), non-Watson-Crick (nonWC), and single stranded (SS) ribonucleotides.

<table>
<thead>
<tr>
<th></th>
<th>WC Bound</th>
<th>WC Prop</th>
<th>nonWC Bound</th>
<th>nonWC Prop</th>
<th>SS Bound</th>
<th>SS Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>32</td>
<td>-1.23</td>
<td>17</td>
<td>-1.38</td>
<td>31</td>
<td>-0.78</td>
</tr>
<tr>
<td>Cys</td>
<td>3</td>
<td>-1.33</td>
<td>4</td>
<td>-0.15</td>
<td>5</td>
<td>-0.10</td>
</tr>
<tr>
<td>Asp</td>
<td>38</td>
<td>-1.12</td>
<td>26</td>
<td>-0.91</td>
<td>55</td>
<td>-0.09</td>
</tr>
<tr>
<td>Glu</td>
<td>49</td>
<td>-1.37</td>
<td>19</td>
<td>-1.98</td>
<td>49</td>
<td>-0.87</td>
</tr>
<tr>
<td>Phe</td>
<td>3</td>
<td>-3.60</td>
<td>3</td>
<td>-2.83</td>
<td>7</td>
<td>-1.88</td>
</tr>
<tr>
<td>Gly</td>
<td>97</td>
<td>0.10</td>
<td>42</td>
<td>-0.35</td>
<td>50</td>
<td>-0.36</td>
</tr>
<tr>
<td>His</td>
<td>50</td>
<td>0.67</td>
<td>26</td>
<td>0.49</td>
<td>26</td>
<td>0.23</td>
</tr>
<tr>
<td>Ile</td>
<td>10</td>
<td>-2.25</td>
<td>3</td>
<td>-3.22</td>
<td>13</td>
<td>-1.37</td>
</tr>
<tr>
<td>Lys</td>
<td>217</td>
<td>0.91</td>
<td>163</td>
<td>1.26</td>
<td>127</td>
<td>0.63</td>
</tr>
<tr>
<td>Leu</td>
<td>19</td>
<td>-2.13</td>
<td>8</td>
<td>-2.61</td>
<td>11</td>
<td>-2.42</td>
</tr>
<tr>
<td>Met</td>
<td>14</td>
<td>-0.58</td>
<td>2</td>
<td>-2.63</td>
<td>9</td>
<td>-0.72</td>
</tr>
<tr>
<td>Asn</td>
<td>84</td>
<td>0.74</td>
<td>48</td>
<td>0.69</td>
<td>48</td>
<td>0.43</td>
</tr>
<tr>
<td>Pro</td>
<td>12</td>
<td>-2.52</td>
<td>7</td>
<td>-2.53</td>
<td>10</td>
<td>-2.28</td>
</tr>
<tr>
<td>Gin</td>
<td>83</td>
<td>0.67</td>
<td>26</td>
<td>-0.24</td>
<td>51</td>
<td>0.47</td>
</tr>
<tr>
<td>Arg</td>
<td>322</td>
<td>1.53</td>
<td>209</td>
<td>1.67</td>
<td>213</td>
<td>1.43</td>
</tr>
<tr>
<td>Ser</td>
<td>85</td>
<td>0.34</td>
<td>56</td>
<td>0.50</td>
<td>74</td>
<td>0.64</td>
</tr>
<tr>
<td>Thr</td>
<td>73</td>
<td>0.19</td>
<td>43</td>
<td>0.19</td>
<td>60</td>
<td>0.40</td>
</tr>
<tr>
<td>Val</td>
<td>11</td>
<td>-2.55</td>
<td>6</td>
<td>-2.67</td>
<td>3</td>
<td>-3.93</td>
</tr>
<tr>
<td>Trp</td>
<td>12</td>
<td>-0.16</td>
<td>5</td>
<td>-0.66</td>
<td>3</td>
<td>-1.66</td>
</tr>
<tr>
<td>Tyr</td>
<td>38</td>
<td>-0.15</td>
<td>25</td>
<td>0.01</td>
<td>41</td>
<td>0.46</td>
</tr>
<tr>
<td>Total</td>
<td>1252</td>
<td>738</td>
<td>886</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.5 Water-mediated hydrogen bond counts and propensities for each amino acid binding to Watson-Crick (WC), non-Watson-Crick (nonWC), and single stranded (SS) ribonucleotides.

<table>
<thead>
<tr>
<th></th>
<th>WC Bound</th>
<th>WC Prop</th>
<th>nonWC Bound</th>
<th>nonWC Prop</th>
<th>SS Bound</th>
<th>SS Prop</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ala</td>
<td>24</td>
<td>-0.98</td>
<td>20</td>
<td>-0.57</td>
<td>16</td>
<td>-0.99</td>
</tr>
<tr>
<td>Cys</td>
<td>5</td>
<td>0.07</td>
<td>2</td>
<td>-0.58</td>
<td>5</td>
<td>0.65</td>
</tr>
<tr>
<td>Asp</td>
<td>58</td>
<td>0.15</td>
<td>31</td>
<td>-0.08</td>
<td>41</td>
<td>0.23</td>
</tr>
<tr>
<td>Glu</td>
<td>46</td>
<td>-0.80</td>
<td>23</td>
<td>-1.12</td>
<td>33</td>
<td>-0.70</td>
</tr>
<tr>
<td>Phe</td>
<td>10</td>
<td>-1.20</td>
<td>2</td>
<td>-2.84</td>
<td>4</td>
<td>-1.94</td>
</tr>
<tr>
<td>Gly</td>
<td>63</td>
<td>0.14</td>
<td>37</td>
<td>0.05</td>
<td>35</td>
<td>-0.13</td>
</tr>
<tr>
<td>His</td>
<td>28</td>
<td>0.50</td>
<td>21</td>
<td>0.76</td>
<td>25</td>
<td>0.91</td>
</tr>
<tr>
<td>Ile</td>
<td>7</td>
<td>-2.10</td>
<td>2</td>
<td>-3.23</td>
<td>3</td>
<td>-2.74</td>
</tr>
<tr>
<td>Lys</td>
<td>105</td>
<td>0.52</td>
<td>68</td>
<td>0.57</td>
<td>60</td>
<td>0.29</td>
</tr>
<tr>
<td>Leu</td>
<td>17</td>
<td>-1.62</td>
<td>13</td>
<td>-1.34</td>
<td>12</td>
<td>-1.55</td>
</tr>
<tr>
<td>Met</td>
<td>7</td>
<td>-0.92</td>
<td>2</td>
<td>-2.05</td>
<td>6</td>
<td>-0.56</td>
</tr>
<tr>
<td>Asn</td>
<td>53</td>
<td>0.74</td>
<td>33</td>
<td>0.73</td>
<td>42</td>
<td>0.98</td>
</tr>
<tr>
<td>Pro</td>
<td>12</td>
<td>-1.85</td>
<td>12</td>
<td>-1.18</td>
<td>10</td>
<td>-1.54</td>
</tr>
<tr>
<td>Gln</td>
<td>51</td>
<td>0.63</td>
<td>21</td>
<td>0.03</td>
<td>32</td>
<td>0.54</td>
</tr>
<tr>
<td>Arg</td>
<td>152</td>
<td>1.11</td>
<td>100</td>
<td>1.18</td>
<td>97</td>
<td>1.04</td>
</tr>
<tr>
<td>Ser</td>
<td>60</td>
<td>0.50</td>
<td>39</td>
<td>0.56</td>
<td>42</td>
<td>0.57</td>
</tr>
<tr>
<td>Thr</td>
<td>42</td>
<td>0.05</td>
<td>30</td>
<td>0.24</td>
<td>24</td>
<td>-0.18</td>
</tr>
<tr>
<td>Val</td>
<td>21</td>
<td>-0.96</td>
<td>14</td>
<td>-0.87</td>
<td>14</td>
<td>-0.96</td>
</tr>
<tr>
<td>Trp</td>
<td>4</td>
<td>-1.08</td>
<td>11</td>
<td>1.05</td>
<td>4</td>
<td>-0.50</td>
</tr>
<tr>
<td>Tyr</td>
<td>25</td>
<td>-0.09</td>
<td>14</td>
<td>-0.25</td>
<td>24</td>
<td>0.43</td>
</tr>
<tr>
<td>Total</td>
<td>790</td>
<td>495</td>
<td>529</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Differences in distance-based interactions depending on secondary structure of ribonucleotides

The overall distance-based interaction propensities (Figure 5.3A and Table 5.3) for each amino acid with RNA are almost identical when comparing Watson-Crick to non-Watson-Crick pair ribonucleotides. Gln has an interaction propensity of 0.22 for Watson-Crick base pairs and -0.28 for non Watson-Crick base pairs. The only other amino acid with a substantial difference in propensity is tyr, with a propensity of -0.12 for Watson-Crick base pairs and 0.04 for non Watson-Crick base pairs.
There are many more differences in propensities when comparing the double stranded and single stranded bases. Tyr has a propensity of 0.53 for single stranded ribonucleotides, while the propensities for Watson-Crick and non Watson-Crick ribonucleotides are much lower at 0.04 and -0.12. Phe shows a slightly positive propensity of 0.06 for single stranded ribonucleotides, while phe has a strongly negative propensity for both Watson-Crick and non Watson-Crick ribonucleotides. Met also shows much higher propensity for single stranded ribonucleotides, 0.35 versus -0.11 for Watson-Crick and -0.33 for non Watson-Crick.

**Differences in direct hydrogen bonds depending on secondary structure of ribonucleotides**

The amino acids show very similar propensities for interacting with all three classes of ribonucleotides by direct hydrogen bonds (Figure 5.3B and Table 5.4). Gln has a higher propensity for direct hydrogen bonds to Watson-Crick and single -stranded ribonucleotides than non Watson-Crick ribonucleotides. Aspartic acid has a strongly negative propensity for both Watson-Crick and non Watson-Crick ribonucleotides, but a propensity of only -0.1 for single stranded ribonucleotides. Tyr also has a higher propensity for single stranded ribonucleotides, 0.46 compared to -0.15 for Watson-Crick and 0.01 for non Watson-Crick ribonucleotides.

**Differences in water-mediated hydrogen bonds**

The propensities for water-mediated hydrogen bonds are also very similar between Watson-Crick, non Watson-Crick, and single stranded ribonucleotides (Figure 5.3C and Table 5.5). Gln shows the largest differences, with propensities of 0.63, 0.03, and 0.54 for Watson-Crick, non Watson-Crick, and single stranded ribonucleotides respectively. Tyr has negative propensities for both Watson-Crick and non Watson-Crick ribonucleotides but a propensity of 0.43 for single stranded ribonucleotides. Trp is the only other amino acid to show a difference, with propensities of -1.1, 1.05, and -0.5 for Watson-Crick, non Watson-
Crick, and single stranded ribonucleotides, although the counts of those interactions are low, with only 4, 11, and 4 bonds observed.

In summary, the overall propensities for all types of interactions are very similar regardless of the secondary structure state of the ribonucleotide. The propensities are especially similar for Watson-Crick and non Watson-Crick base pairs, while more differences are seen between the base-paired and single stranded ribonucleotides.

**Binding preferences between main chains and side chains of amino acids**

We further specified each interaction between amino acids and ribonucleotides as being between the ribonucleotide and either the main chain or side chain atoms of the amino acid. The interaction counts and propensities for main chain and side chain interactions based on all three definitions of RNA-binding are given in Table 5.6 and Figure 5.4. For each definition of RNA-binding, we observe more contacts involving amino acid side chain atoms than main chain atoms. For distance-based interactions, we observe 4632 amino acids with main chain atoms within the distance cutoff and 6497 amino acids with side chain atoms within the distance cutoff. 989 amino acids have direct hydrogen bonds with RNA made by main chain atoms, while 2251 amino acids have direct hydrogen bonds with RNA made by side chain atoms. For water-mediated hydrogen bonds, there are 995 and 1210 amino acids with main chain and side chain interactions, respectively. We note that the interactions made by main chain and side chain atoms are not exclusive; a single amino acid can be counted in both groups if it makes contacts to the RNA with both the main chain and the side chain atoms. Gly has no side chain, and therefore has no side chain binding propensities. Furthermore, the amino acids ala, phe, ile, leu, pro, and val do not have hydrogen bond donors or acceptors in their side chains, and therefore have no hydrogen bonding propensities for their side chains.
Table 5.6 Interaction propensities and counts for main chain and side chain contacts with RNA.

<table>
<thead>
<tr>
<th>Distance-based</th>
<th>Direct Hydrogen Bonds</th>
<th>Water-mediated Hydrogen Bonds</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Main Chain Prop</td>
<td>Side Chain Prop</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Prop</td>
</tr>
<tr>
<td>397</td>
<td>0.51</td>
<td>0.51</td>
</tr>
<tr>
<td>30</td>
<td>0.10</td>
<td>0.10</td>
</tr>
<tr>
<td>205</td>
<td>-0.58</td>
<td>-0.58</td>
</tr>
<tr>
<td>163</td>
<td>-1.52</td>
<td>-1.52</td>
</tr>
<tr>
<td>80</td>
<td>-0.75</td>
<td>-0.75</td>
</tr>
<tr>
<td>760</td>
<td>1.18</td>
<td>1.18</td>
</tr>
<tr>
<td>106</td>
<td>-0.13</td>
<td>-0.13</td>
</tr>
<tr>
<td>121</td>
<td>-0.54</td>
<td>-0.54</td>
</tr>
<tr>
<td>421</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>207</td>
<td>-0.57</td>
<td>-0.57</td>
</tr>
<tr>
<td>72</td>
<td>-0.11</td>
<td>-0.11</td>
</tr>
<tr>
<td>202</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>282</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>150</td>
<td>-0.36</td>
<td>-0.36</td>
</tr>
<tr>
<td>405</td>
<td>-0.03</td>
<td>-0.03</td>
</tr>
<tr>
<td>390</td>
<td>0.65</td>
<td>0.65</td>
</tr>
<tr>
<td>282</td>
<td>0.25</td>
<td>0.25</td>
</tr>
<tr>
<td>197</td>
<td>-0.28</td>
<td>-0.28</td>
</tr>
<tr>
<td>51</td>
<td>0.04</td>
<td>0.04</td>
</tr>
<tr>
<td>111</td>
<td>-0.49</td>
<td>-0.49</td>
</tr>
<tr>
<td>4632</td>
<td>6497</td>
<td>989</td>
</tr>
</tbody>
</table>

154
Figure 5.4 Interaction propensities for main chain and side chains of amino acids. Panel A shows the distance-based interaction propensities, Panel B shows the direct hydrogen bonding propensities, and Panel C shows the water-mediated hydrogen bonding propensities.

There are considerable differences in main chain and side chain RNA-binding propensities of individual amino acids. Arg, lys, and his have the overall highest RNA-binding propensities of any amino acids, largely due to side chain interactions. These three amino acids have much higher propensities for RNA-binding by their side chain atoms than their main chain atoms. For the distance-based definition, Arg and Lys each have propensities of -0.03 for main chain interactions while the propensities for side chain interactions are 1.28 for Arg and 0.72 for Lys. This preference for main chain interactions is most extreme for the distance-based definition of RNA-binding and not as pronounced for the water-mediated hydrogen bond definition. Arg, Lys, and His all have large side chains
that make favorable contacts with RNA, which makes their main chain atoms much less likely to fall within the distance-cutoff. However, the water molecule in water-mediated hydrogen bonds allows for longer range interactions with the main chain atoms, and the difference in propensities for interacting with the main chain atoms versus side chain atoms of Arg and His via water-mediated hydrogen bonds is smaller than with distance-based interactions.

The negatively charged Asp residue has a highly negative interaction propensity in general, but has a positive interaction propensity for water-mediated hydrogen bonds made by its side chain. The bridging water molecule allows for a favorable interaction whereas direct hydrogen bonds are unfavorable due to the charge repulsion of the negatively charged Asp side chain and the negatively charged RNA backbone.

Ala also shows a difference in binding propensities between main chain and side chain atoms. For the distance-based definition, ala has an interaction propensity of 0.51 for main chain interactions and -0.86 for side chain interactions. The small hydrophobic side chain of ala is not likely to be found in the protein-RNA interface, but the main chain atoms are often found in binding sites.

The large aromatic amino acid tyr also shows a large difference in binding propensities with main chain and side chain atoms. The side chain propensities for tyr are positive, while the main chain propensities are negative. The large side chain again precludes interactions with the main chain atoms, but the side chain can participate in both types of hydrogen bonds and can form stacking interactions with the RNA ribonucleotides.

**Binding preferences between base, phosphate, and ribose atoms of ribonucleotides**

We analyzed the interactions for each ribonucleotide based on whether the interacting atom was in the base, phosphate, or ribose of the ribonucleotide. Figure 5.5 shows the
propensities for each amino acid to interact with the base, phosphate, or ribose groups using
all three definitions of RNA-binding. The counts and propensities for these interactions
based on distance are shown in Table 5.7, direct hydrogen bonds in Table 5.8, and water-
mediated hydrogen bonds in Table 5.9. As with the main chain and side chain interactions,
these categories are not mutually exclusive; a ribonucleotide can have interactions with any
combination of the three groups of atoms. The total counts of interactions with each of the
three groups of atoms are shown at the bottom of Tables 5.7, 5.8, and 5.9. We see that the
large majority of interactions are with the phosphate group or the ribose, rather than the base
atoms for all three definitions of RNA-binding. Typically only the interactions with the base
atoms are potentially sequence-specific. Based on the observed counts of interactions, we
note that the non-sequence specific contacts with the sugar-phosphate backbone of RNA are
highly important in protein-RNA recognition. These contacts increase the affinity of the
interaction and provide stability for the complex whereas contacts to only the base atoms
may not be stable enough to be effective. We also note that several RNA-binding proteins
are non-sequence specific binders and have functions that depend on being able to bind any
RNA sequence.
Figure 5.5 Interaction propensities for base, phosphate, and ribose atoms of ribonucleotides.

Panel A shows the distance-based interaction propensities, Panel B shows the direct hydrogen bonding propensities, and Panel C shows the water-mediated hydrogen bonding propensities.
Table 5.7 Interaction propensities and counts for contacts with RNA base, phosphate, and ribose atoms using the distance-based definition of RNA-binding.

<table>
<thead>
<tr>
<th></th>
<th>Distance-based</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td></td>
<td>Phosphate</td>
<td></td>
<td>Ribose</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td></td>
</tr>
<tr>
<td>Ala</td>
<td>135</td>
<td>-0.20</td>
<td>155</td>
<td>-0.45</td>
<td>268</td>
<td>-0.18</td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>19</td>
<td>0.29</td>
<td>8</td>
<td>-1.41</td>
<td>16</td>
<td>-0.93</td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>134</td>
<td>-0.35</td>
<td>104</td>
<td>-1.17</td>
<td>204</td>
<td>-0.71</td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>143</td>
<td>-0.87</td>
<td>93</td>
<td>-1.94</td>
<td>205</td>
<td>-1.32</td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>78</td>
<td>0.06</td>
<td>58</td>
<td>-0.82</td>
<td>128</td>
<td>-0.19</td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>189</td>
<td>0.02</td>
<td>253</td>
<td>-0.02</td>
<td>449</td>
<td>0.30</td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>109</td>
<td>0.75</td>
<td>106</td>
<td>0.26</td>
<td>180</td>
<td>0.51</td>
<td></td>
</tr>
<tr>
<td>Ile</td>
<td>79</td>
<td>-0.31</td>
<td>58</td>
<td>-1.21</td>
<td>156</td>
<td>-0.29</td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>226</td>
<td>-0.08</td>
<td>678</td>
<td>1.05</td>
<td>600</td>
<td>0.36</td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>136</td>
<td>-0.33</td>
<td>99</td>
<td>-1.25</td>
<td>248</td>
<td>-0.43</td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>51</td>
<td>0.24</td>
<td>30</td>
<td>-0.98</td>
<td>102</td>
<td>0.27</td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>146</td>
<td>0.49</td>
<td>161</td>
<td>0.18</td>
<td>225</td>
<td>0.15</td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>133</td>
<td>-0.09</td>
<td>119</td>
<td>-0.71</td>
<td>230</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>132</td>
<td>0.30</td>
<td>126</td>
<td>-0.22</td>
<td>221</td>
<td>0.07</td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>365</td>
<td>0.67</td>
<td>809</td>
<td>1.36</td>
<td>788</td>
<td>0.81</td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>139</td>
<td>0.01</td>
<td>236</td>
<td>0.32</td>
<td>291</td>
<td>0.11</td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>125</td>
<td>-0.08</td>
<td>171</td>
<td>-0.08</td>
<td>252</td>
<td>-0.04</td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>95</td>
<td>-0.48</td>
<td>99</td>
<td>-0.88</td>
<td>216</td>
<td>-0.27</td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>36</td>
<td>0.38</td>
<td>37</td>
<td>-0.03</td>
<td>69</td>
<td>0.35</td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>107</td>
<td>0.30</td>
<td>137</td>
<td>0.20</td>
<td>199</td>
<td>0.22</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>2577</td>
<td></td>
<td>3537</td>
<td></td>
<td>5047</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.8 Interaction propensities and counts for contacts with RNA base, phosphate, and ribose atoms for direct hydrogen bonds.

<table>
<thead>
<tr>
<th></th>
<th>Base</th>
<th></th>
<th>Phosphate</th>
<th></th>
<th>Ribose</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
</tr>
<tr>
<td>Ala</td>
<td>19</td>
<td>-1.15</td>
<td>36</td>
<td>-1.37</td>
<td>28</td>
<td>-1.20</td>
</tr>
<tr>
<td>Cys</td>
<td>5</td>
<td>0.24</td>
<td>4</td>
<td>-1.22</td>
<td>6</td>
<td>-0.10</td>
</tr>
<tr>
<td>Asp</td>
<td>56</td>
<td>0.27</td>
<td>12</td>
<td>-3.09</td>
<td>69</td>
<td>-0.04</td>
</tr>
<tr>
<td>Glu</td>
<td>43</td>
<td>-0.72</td>
<td>11</td>
<td>-3.83</td>
<td>79</td>
<td>-0.46</td>
</tr>
<tr>
<td>Phe</td>
<td>3</td>
<td>-2.76</td>
<td>6</td>
<td>-2.90</td>
<td>5</td>
<td>-2.63</td>
</tr>
<tr>
<td>Gly</td>
<td>41</td>
<td>-0.31</td>
<td>92</td>
<td>-0.29</td>
<td>69</td>
<td>-0.17</td>
</tr>
<tr>
<td>His</td>
<td>31</td>
<td>0.82</td>
<td>37</td>
<td>-0.07</td>
<td>46</td>
<td>0.78</td>
</tr>
<tr>
<td>Ile</td>
<td>13</td>
<td>-1.03</td>
<td>8</td>
<td>-2.88</td>
<td>7</td>
<td>-2.53</td>
</tr>
<tr>
<td>Lys</td>
<td>88</td>
<td>0.44</td>
<td>375</td>
<td>1.39</td>
<td>124</td>
<td>0.33</td>
</tr>
<tr>
<td>Leu</td>
<td>8</td>
<td>-2.54</td>
<td>12</td>
<td>-3.10</td>
<td>19</td>
<td>-1.90</td>
</tr>
<tr>
<td>Met</td>
<td>10</td>
<td>-0.23</td>
<td>4</td>
<td>-2.69</td>
<td>15</td>
<td>-0.25</td>
</tr>
<tr>
<td>Asn</td>
<td>53</td>
<td>0.91</td>
<td>80</td>
<td>0.36</td>
<td>85</td>
<td>0.98</td>
</tr>
<tr>
<td>Pro</td>
<td>9</td>
<td>-2.10</td>
<td>0</td>
<td>N/A</td>
<td>20</td>
<td>-1.55</td>
</tr>
<tr>
<td>Gln</td>
<td>62</td>
<td>1.09</td>
<td>54</td>
<td>-0.26</td>
<td>76</td>
<td>0.77</td>
</tr>
<tr>
<td>Arg</td>
<td>133</td>
<td>1.09</td>
<td>536</td>
<td>1.96</td>
<td>227</td>
<td>1.25</td>
</tr>
<tr>
<td>Ser</td>
<td>58</td>
<td>0.63</td>
<td>105</td>
<td>0.34</td>
<td>83</td>
<td>0.53</td>
</tr>
<tr>
<td>Thr</td>
<td>42</td>
<td>0.22</td>
<td>95</td>
<td>0.26</td>
<td>64</td>
<td>0.22</td>
</tr>
<tr>
<td>Val</td>
<td>3</td>
<td>-3.59</td>
<td>8</td>
<td>-3.32</td>
<td>9</td>
<td>-2.62</td>
</tr>
<tr>
<td>Trp</td>
<td>3</td>
<td>-1.33</td>
<td>11</td>
<td>-0.59</td>
<td>8</td>
<td>-0.52</td>
</tr>
<tr>
<td>Tyr</td>
<td>22</td>
<td>-0.11</td>
<td>63</td>
<td>0.27</td>
<td>31</td>
<td>-0.22</td>
</tr>
<tr>
<td>Total</td>
<td>702</td>
<td>1549</td>
<td>1070</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5.9 Interaction propensities and counts for contacts with RNA base, phosphate, and ribose atoms for water-mediated hydrogen bonds.

<table>
<thead>
<tr>
<th></th>
<th>Water-Mediated Hydrogen Bonds</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Base</td>
<td>Phosphate</td>
<td>Ribose</td>
<td>Base</td>
<td>Phosphate</td>
<td>Ribose</td>
<td>Base</td>
<td>Phosphate</td>
<td>Ribose</td>
<td>Base</td>
<td>Phosphate</td>
<td>Ribose</td>
<td>Base</td>
<td>Phosphate</td>
<td>Ribose</td>
<td>Base</td>
<td>Phosphate</td>
</tr>
<tr>
<td></td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
<td>Prop</td>
<td>Count</td>
</tr>
<tr>
<td>Ala</td>
<td>10</td>
<td>-1.87</td>
<td>33</td>
<td>-0.85</td>
<td>25</td>
<td>-0.97</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cys</td>
<td>4</td>
<td>0.13</td>
<td>6</td>
<td>0.01</td>
<td>2</td>
<td>-1.30</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asp</td>
<td>54</td>
<td>0.43</td>
<td>57</td>
<td>-0.20</td>
<td>69</td>
<td>0.35</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Glu</td>
<td>62</td>
<td>0.01</td>
<td>37</td>
<td>-1.43</td>
<td>58</td>
<td>-0.51</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phe</td>
<td>6</td>
<td>-1.55</td>
<td>9</td>
<td>-1.67</td>
<td>9</td>
<td>-1.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gly</td>
<td>46</td>
<td>0.07</td>
<td>71</td>
<td>-0.01</td>
<td>65</td>
<td>0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>His</td>
<td>16</td>
<td>0.07</td>
<td>38</td>
<td>0.62</td>
<td>31</td>
<td>0.60</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lle</td>
<td>4</td>
<td>-2.52</td>
<td>5</td>
<td>-2.90</td>
<td>8</td>
<td>-1.95</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Lys</td>
<td>67</td>
<td>0.25</td>
<td>133</td>
<td>0.54</td>
<td>95</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Leu</td>
<td>14</td>
<td>-1.52</td>
<td>17</td>
<td>-1.95</td>
<td>17</td>
<td>-1.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Met</td>
<td>4</td>
<td>-1.34</td>
<td>8</td>
<td>-1.05</td>
<td>10</td>
<td>-0.45</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Asn</td>
<td>53</td>
<td>1.12</td>
<td>66</td>
<td>0.73</td>
<td>60</td>
<td>0.87</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pro</td>
<td>8</td>
<td>-2.06</td>
<td>13</td>
<td>-2.06</td>
<td>17</td>
<td>-1.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Gln</td>
<td>46</td>
<td>0.87</td>
<td>52</td>
<td>0.34</td>
<td>54</td>
<td>0.67</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Arg</td>
<td>100</td>
<td>0.89</td>
<td>235</td>
<td>1.42</td>
<td>148</td>
<td>1.02</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ser</td>
<td>43</td>
<td>0.40</td>
<td>78</td>
<td>0.56</td>
<td>58</td>
<td>0.40</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Thr</td>
<td>32</td>
<td>0.04</td>
<td>54</td>
<td>0.09</td>
<td>39</td>
<td>-0.10</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Val</td>
<td>11</td>
<td>-1.51</td>
<td>26</td>
<td>-0.97</td>
<td>22</td>
<td>-0.94</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trp</td>
<td>5</td>
<td>-0.38</td>
<td>12</td>
<td>0.18</td>
<td>8</td>
<td>-0.13</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Tyr</td>
<td>22</td>
<td>0.10</td>
<td>38</td>
<td>0.19</td>
<td>23</td>
<td>-0.26</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>607</td>
<td>988</td>
<td>818</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

There are some clear differences in binding propensities for base, phosphate, and ribose atoms for some amino acids. Arg has a high propensity to interact with any of the three groups of atoms, but lys has a much higher propensity for interacting with phosphate and ribose atoms compared to base atoms. This indicates that arg may be used just as often
for sequence specific and non-sequence specific contacts, while lys is preferred mainly for non-sequence specific contacts.

There are some amino acids that have a clear preference for binding base atoms over either phosphate or ribose atoms. Asn and gln have higher propensities for interacting with base atoms than with phosphate or ribose atoms by all three definitions of RNA-binding. This preference for base atoms indicates that Asn and Gln may often be used for sequence-specific binding of RNA. The negatively charged asp shows negative propensities for interacting with the negatively charged phosphate atoms, but has a slightly positive propensity for both direct and water-mediated hydrogen bonds with base atoms. This indicates that a well placed asp residue may be important for sequence-specific binding of RNA.

**DISCUSSION**

In this work, we have created PRIDB, a database of protein-RNA complexes from the PDB. We have performed a detailed analysis of the characteristics of protein-RNA interfaces of RB181, a non-redundant set of 181 proteins, which is the largest dataset of protein-RNA complexes analyzed to date. Defining RNA-binding residues in three ways, amino acids with atoms within 5 Å of RNA atoms (distance-based contacts), residues making direct hydrogen bonds, and residues making water-mediated hydrogen bonds, we analyzed the binding propensities for each type of contact. We differentiated between contacts to single stranded ribonucleotides, Watson-Crick paired ribonucleotides, and non-Watson-Crick paired ribonucleotides and analyzed each contact as being made by main chain or side chain atoms on the protein side, and by base, phosphate, or ribose atoms on the RNA side.

78.3% of ribonucleotides in the RB181 dataset are in base pairs, with 46.6% in Watson-Crick pairs and 31.7% in non-Watson-Crick pairs. 75% of the Watson-Crick base pairs are CG pairs. The percentages of ribonucleotides in the different secondary structure
states may not be a reflection of the naturally occurring distribution of each type. Many protein-RNA complexes in the PDB contain synthetic RNA oligomers chosen specifically for their stability in a double helix, which may explain the high percentage of CG base pairs in the dataset. Also, the ribosomal proteins in RB181 are mostly from thermophilic organisms, which may have evolved a higher GC content to confer more stability at high temperatures (Das et al., 2006).

In general, we observe similar trends in protein-RNA interfaces using three definitions of binding residues: approximately the same percentage of each ribonucleotide is bound by amino acids, and the overall propensities for amino acids to bind are also similar. By any definition of RNA-binding, the amino acids arg, lys, his, asn, gln, and ser are favored in RNA-binding. Generally, the hydrophobic amino acids are disfavored for RNA-binding. However, there are some important differences in RNA-binding propensities between the different definitions. In particular, asp is a favored residue for water-mediated hydrogen bonds, whereas it is disfavored in distance-based contacts or direct hydrogen bonding. The extra distance afforded by the bridging water molecule likely allows this negatively charged amino acid to be preferred in this type of interaction.

Several preferences for specific ribonucleotides to be contacted by specific amino acids were also observed. His has a higher propensity for binding with adenine than with the other ribonucleotides, using the distance-based definition of contact. Based on direct hydrogen bonding, the his-adenine pair is favored as is the asn-uracil pair, while the gln-adenine and ser-uracil pairs are disfavored. There are also some interesting differences in water-mediated hydrogen bond propensities between the different ribonucleotides. The his-cytosine pair is disfavored compared with other ribonucleotides, the opposite situation from direct hydrogen bonds. Another interesting feature of the water-mediated hydrogen bond propensities is that uracil is preferred for several amino acids.
We observe a substantial difference in the binding propensities of main chain atoms versus side chain atoms for many amino acids. The highly preferred RNA-binding residues arg and lys accomplish most of their binding interactions via their side chains. Most of the other amino acids that show a difference in binding propensity show higher propensities with their side chains than with main chain atoms. However, there are examples of main chain contacts being preferred, especially for ala. The small hydrophobic side chain has a highly negative propensity for binding RNA using the distance-based definition, but has a favorable propensity for main chain contacts.

**The influence of ribonucleotide secondary structure on protein-RNA recognition**

Non-Watson-Crick base pairs have long been recognized as important in protein-RNA interactions (Bartel et al., 1991, Hermann and Westhof, 1999, Draper, 1999, Westhof and Fritsch, 2000, Varani and McClain, 2000, Abad et al., 2008, Moulinier et al., 2001, McClain, 2006). Kim et al. (Kim et al., 2003) analyzed the hydrogen bonding propensities of single stranded and double stranded ribonucleotides without differentiating between Watson-Crick and non-Watson-Crick base pairs. We have performed a detailed analysis of the interaction propensities of single stranded, Watson-Crick paired, and non-Watson-Crick paired ribonucleotides in order to understand whether there are differences in how proteins interact with these different classes of ribonucleotides. We found few differences in binding propensities between non-Watson-Crick and Watson-Crick paired ribonucleotides. Gln showed the largest difference in binding propensities between Watson-Crick and non-Watson-Crick base paired ribonucleotides. The other amino acids had largely the same propensities for binding either type of paired ribonucleotide. While some proteins may require non-Watson-Crick base pairs for recognition, the mode of interaction for these pairs does not appear to be fundamentally different than recognition of Watson-Crick base paired ribonucleotides.
We observed more differences in binding propensities with single stranded ribonucleotides than with Watson-Crick and non-Watson-Crick paired ribonucleotides, indicating that recognition of single stranded segments of RNA is more divergent than Watson-Crick and non-Watson-Crick segments. This finding has important implications for predicting RNA-binding sites in proteins because it suggests that improved predictions could be obtained by generating distinct classifiers for single stranded versus double stranded segments of RNA.

**Sequence-specific interactions**

We analyzed the contacts between proteins and RNA based on which atoms in the ribonucleotide made contact with the amino acid. We differentiated between contacts made by the base atoms, the phosphate atoms, and the ribose atoms and observed that the majority of contacts between amino acids and ribonucleotides were with the phosphate and ribose atoms. The amino acids with the highest propensity for binding the base atoms were his, arg, asn, and gln. Asn and gln had substantially higher propensities for binding base atoms than either phosphate or ribose atoms, and asp had a clear preference for making both direct and water-mediated hydrogen bonds with base atoms. Based on their preference for base atoms over phosphate and ribose atoms, we conclude that asn, gln, and asp are important residues in sequence-specific recognition of RNA.

**Functional classes of RNA**

Ellis et al. (Ellis et al., 2007) performed an analysis of protein-RNA interactions and differentiated between five different functional classes of RNA, mRNA, rRNA, tRNA, viral RNA, and RNA ligands. They observed that rRNA complexes had fewer base specific contacts than the other functional classes. Ellis et al. attribute this difference to differences in the functional class of the RNA. They also noted that rRNA complexes are largely composed of double stranded RNA while the other classes contain large sections of single
stranded RNA; thus, the difference may be attributable simply to availability of base atoms for contact with proteins.

In light of these results, we split the RB181 dataset into two groups, rRNA complexes and all other complexes and performed an analysis of each group. RB181 contains 74 proteins that are bound to rRNA and 107 proteins bound to other types of RNA. We found that the rRNA group contained 81.3% dsRNA and the other group contained 62.3% dsRNA indicating that there is a substantial difference in the secondary structure states of the ribonucleotides in the two groups. Using the distance-based definition of RNA-binding, we found that the rRNA group had 3763 binding residues out of 8016 surface residues, or 47% of the amino acids interacted with RNA. Of these binding residues, 18.3% had contacts with base atoms. In the other group, 3321 out of 28471 surface residues, or 11.7% interacted with RNA, with 30.3% of them interacting with base atoms. When comparing the binding propensities of amino acids with Watson-Crick, non-Watson-Crick, and single stranded ribonucleotides between the rRNA group and the other group, we observed few differences. Our results are in agreement with those of Ellis et al.; we also observed a difference in the percentage of base specific contacts in the rRNA groups versus the other group. We note that this difference is probably attributable to the observed difference between single stranded versus double stranded RNA, not to the functional class of the RNA per se. We see the same trends in propensities of interaction between the two groups if we account for single stranded versus double stranded state of the ribonucleotides.

**PRIDB: An online resource for protein-RNA interactions**

The Protein-RNA Interaction Database (PRIDB) is an online database of all protein-RNA complexes in the PDB. The web server allows for users to input a list of proteins and obtain a detailed report describing the protein-RNA interfaces, in a format similar to the tables presented in this paper. The user can specify the level of interactions, from whole
residue interactions to detailed atomic level contacts and the definition of RNA-binding residues is also flexible, allowing for user-specified distance cutoffs, direct hydrogen bonds, and water-mediated hydrogen bonds. PRIDB is available online at http://bindr.gdcb.iastate.edu/PRIDB.

REFERENCES


Baker, C. M., and G. H. Grant (2005), Role of aromatic amino acids in protein-nucleic acid recognition, Biopolymers, 85(5-6), 456-70.


CHAPTER 6. GENERAL CONCLUSIONS

Protein-RNA interactions are responsible for carrying out and regulating a variety of vital cellular processes. Characterization of these interactions, including prediction of RNA-binding sites in proteins, is essential for understanding how these events occur and how they are regulated. In this dissertation, we have presented: several new methods for predicting RNA-binding sites in proteins, a web server for analyzing and predicting RNA-binding sites in proteins, a comprehensive non-redundant database of all protein-RNA complexes in the PDB, and an analysis of the protein-RNA interfaces from the largest dataset studied to date.

CONTRIBUTIONS

Classifiers that accurately predict RNA-binding sites in proteins

We have developed methods to predict RNA-binding sites in proteins using protein sequence-derived information alone as well as methods that use a combination of protein sequence and structure-derived information. We have shown that IDStruct, which exploits the context of a residue in terms of its nearest neighbors within the protein structure, gives better predictions of RNA-binding sites than IDSeq, which uses only the sequence-based context. In addition, we have shown that PSSMSeq, which takes advantage of evolutionary information in the form of PSSMs, has increased prediction performance over IDSeq, which uses single sequence-based inputs. Further, we have developed an ensemble classifier that combines the predictions of the IDStruct, PSSMSeq, and PSSMStruct classifiers to provide significantly improved predictions over those obtained with any of the individual classifiers. Finally, for two clinically important RNA binding proteins for which high resolution structures are not yet available, the telomerase reverse transcriptase and lentiviral Rev proteins, we have demonstrated that predictions of RNA-binding sites in agree well with all available experimental data.
A web server for analyzing and predicting RNA-binding sites in proteins

We have developed RNABindR, a server for analyzing and predicting RNA-binding sites in proteins (http://bindr.gdcb.iastate.edu/RNABindR). RNABindR allows users to identify known RNA-binding sites in protein-RNA complexes from the PDB and interactively displays the complex with RNA binding residues highlighted. For proteins without a known protein-RNA complex, RNABindR predicts the RNA binding sites. RNABindR has been accessed from 780 unique IP addresses in 45 countries. Tens of thousands of sequences have been submitted to RNABindR for predictions.

Several groups have published the results of using RNABindR predictions. One example is described by Sunita et al. (Sunita et al., 2007), who studied the structure and RNA-binding domains of the 16S rRNA methyltransferase RsmC, which specifically modifies G1207 of *E. coli* 16S rRNA. Mutations of G1207 are lethal, indicating the importance of this modified nucleotide. Predictions from RNABindR indicated a potential RNA-binding site in the N-terminal domain (NTD) of the protein, and further experimental analysis of this region showed several point mutations that decreased methyltransferase activity by 30-50%. The authors conclude that the NTD, predicted to be RNA-binding by RNABindR, is the main RNA-binding domain of the RsmC protein.

In a second example, Bechara et al. (Bechara et al., 2006) studied the RNA-binding properties of three isoforms of Fragile X Related Protein 1 (FXR1P). FXR1P is involved in Fragile X syndrome, the most common form of inherited mental retardation. FXR1P has been proposed to specifically bind mRNAs containing a G-quartet structure. The three isoforms are the result of alternative splicing, with the isoform Isoe containing exon 15, and the isoforms Isoa and Isod lacking this exon. RNABindR strongly predicted RNA-binding residues in the amino acids encoded by exon 15, among other regions, indicating that Isoe would bind RNA with this region that is missing in the other two isoforms. All three isoforms were shown to bind RNA using filter binding assays, but the authors demonstrated
that Isoe binds specifically to G-quartet containing RNAs, while the other two isoforms do not specifically recognize G-quartet containing RNAs. The authors conclude that exon 15 encodes the portion of the protein that gives the required functional specificity for FXR1P. This indicates that RNABindR is able to identify important RNA-binding regions of proteins.

A third case in which RNABindR provided valuable predictions is provided by Keren et al. (Keren et al., 2008), who studied the RNA-binding domains of CRS1, a chloroplast RNA splicing protein that specifically recognizes group II introns. CRS1 contains four CRM (chloroplast RNA splicing and ribosome maturation) domains, and have been shown to be required for splicing of plastid genes with group II introns. RNABindR predictions implicated three conserved motifs, GXXG, WKHK, and YRP as potential RNA-binding sites. The authors found that mutations in each of these sequences dramatically reduced RNA-binding activity in both filter binding assays and gel shift assays. The authors appear to have tested these particular mutations on the basis of their predicted RNA-binding capabilities from RNABindR and the fact that they were conserved in multiple sequence alignments, although we note that the version of RNABindR used by the authors does not include conservation information. Taken together, these studies highlight the utility of RNABindR predictions in guiding experimental investigations of RNA-binding sites in proteins.

**A comprehensive database of protein-RNA interfaces**

The Protein-RNA Interaction Database, PRIDB, provides the most up to date and flexible database of protein-RNA interactions to date (http://bindr.gdcb.iastate.edu/PRIDB). It is currently the only available database that allows a user to input a list of proteins and generate a summary of the protein-RNA interfaces. PRIDB also allows for flexible definitions of RNA-binding residues; all other existing databases have preset definitions or
only include hydrogen bond data. PRIDB will be a valuable resource for those interested in studying properties of protein-RNA interactions.

**Analysis of the characteristics of protein-RNA interaction sites**

We have performed an analysis of the protein-RNA interfaces in a non-redundant dataset of 181 proteins, the largest set analyzed to date. Through this analysis, we have calculated the interaction propensities for each amino acid to interact through van der Waals interactions, direct hydrogen bonds, and water-mediated hydrogen bonds. We have investigated the interaction propensities for amino acids with single-stranded, Watson-Crick paired, and non Watson-Crick paired nucleotides and found that the two classes of double-stranded nucleotides have almost identical interaction propensities while the single-stranded nucleotides have more divergent interaction propensities. Amino acids are more likely to interact with RNA through their side chain atoms than with their main chain atoms, and most contacts are to the RNA backbone rather than the bases. Through this analysis, we have confirmed many previous observations and made several new observations.

The analysis presented in this dissertation was based on a dataset of 181 proteins, while the largest previous study was done on a dataset of 89 proteins. Carrying out this analysis on larger datasets, as we have, is essential because early studies have shown that small datasets can provide biased results that do not generalize well as more structures of protein-RNA complexes become available. The current scarcity of structural information on protein-RNA complexes means that any significantly larger dataset can provide much valuable information.
FUTURE STUDIES

Protein-RNA interaction sites are complex and highly variable. It has been a great challenge to identify common themes in protein-RNA binding that can be used for prediction. Prediction performance is certainly not optimal, and there are several avenues to pursue for improvements.

Effectively use protein structural information for prediction of RNA-binding sites: To date the use of information from protein structure has not provided a large improvement in prediction performance. We have shown an improvement in prediction performance when using spatial neighbors as input rather than sequence neighbors. Others have attempted to use specific aspects of protein structure for predicting RNA-binding sites, including residue doublet propensity, solvent accessibility, secondary structure, and surface curvature (Kim et al. 2006, Wang et al. 2008, Chen and Lim, 2008, Shulman-Peleg et al., 2008, Towfic et al., 2008). These efforts have produced some gains in prediction performance, but there is much room for improvement. Future work should include a systematic analysis of additional structural features that can be used as input to classifiers and different representations of structural information that may improve prediction performance.

One potentially rich source of information is the B factors in crystallographic data. The B factor measures the disorder and amount of motion allowed for each atom in a structure. We might expect the atoms of interacting residues to have lower B factors than the atoms of non-interacting residues, due to the structural stability of protein-RNA interfaces, but a detailed analysis of this correlation needs to be performed. One complicating factor is that the structural data we have used so far comes from for protein-RNA complexes; the B factors may be quite different in the complexes than in the unbound proteins. We have shown our prediction methods to be rather insensitive to whether the protein structure is derived from a complex or from an unbound protein. However, a systematic analysis of the
differences in B factors for cases in which both bound and free structures are available will be required to draw definitive conclusions regarding the potential contribution of B-factor information to interface residue prediction.

**Incorporate longer range interactions between protein and RNA:** For our prediction methods, we have defined RNA-binding residues using a distance cutoff of 5 Å. While this distance is sufficient to include many types of protein-RNA interactions, it may neglect some important long range interactions. For example, the Lennard-Jones potential shows favorable energies of interaction for distances beyond our 5 Å cutoff. By incorporating these longer-range interactions into our definition of RNA-binding residues, we may be able to define a more realistic set of residues that are important for complex stability and therefore be able to capture more biologically relevant signals using machine learning classifiers.

**Use all available protein-RNA complexes to train customized classifiers:** In our efforts to date, we have used a non-redundant dataset of proteins in order to allow for a rigorous statistical evaluation and to generate classifiers that are not biased towards any given type of RNA-binding protein. These classifiers may well provide the best possible performance on novel proteins, but the best possible prediction of RNA-binding residues for a protein of interest will likely result from training *only* on similar proteins. In preliminary results, we have generated customized classifiers by training only on proteins highly similar to the test protein instead of the entire dataset (D. Reyon, personal communication). For a given test protein, a BLAST search is performed against all protein-RNA complexes in the PDB, and a customized classifier is generated from the training set similar proteins and used to make predictions on the test protein. For those proteins for which this procedure can find several highly similar sequences for building the training set, the prediction of RNA-binding residues is much improved over predictions made using a "generic" classifier. Further
developments of this method, including identification of proteins with similar *structures*, are likely to provide highly accurate predictions of RNA-binding residues.

**Incorporate RNA secondary structure information:** Our analysis of interaction propensities between amino acids and Watson-Crick paired nucleotides, non-Watson-Crick paired nucleotides, and single-stranded nucleotides indicated that interactions between paired nucleotides and single-stranded nucleotides have some significant differences. Future studies should exploit this difference by creating different classifiers for paired and unpaired nucleotides. A systematic analysis of value of using RNA secondary structure information from different sources will be necessary. RNA secondary structures can be probed experimentally or predicted computationally (Fabian and White, 2008, Shapiro et al., 2007). Ideally, predicted secondary structure would provide accurate enough information to improve prediction of RNA-binding sites.

**Develop a meta-server for RNA-binding site prediction:** Meta-servers have been successfully developed for tasks such as protein structure prediction (Fischer, 2006). Several methods for predicting RNA-binding sites in proteins have been published; all use slightly different input information and computational methods (Jeong et al., 2004, Terribilini et al., 2006, Wang and Brown, 2006, Jeong and Miyano, 2006, Kim et al., 2006, Kumar et al. 2007, Wang et al., 2008, Tong et al., 2008, Chen and Lim, 2008, Shulman-Peleg et al., 2008, Towfic et al., 2008). We have shown that combining different types of classifiers can lead to increased prediction performance. It seems likely that a meta-server method for RNA-binding site prediction will provide better predictions than any single method. Experiments with meta-servers would need to systematically evaluate all possible combinations of available methods as well the most effective ways of performing the combination. A variety of machine learning methods should be investigated, in addition to simple or weighted voting schemes. In our work, we have found that a Naïve Bayes combination method provides
better performance than a simple voting scheme, but this may not be the case with combining different methods.

**Further development of PRIDB:** The current implementation of PRIDB allows users to enter a list of selected proteins, choose a definition of RNA-binding residues, and obtain a detailed analysis of the protein-RNA interfaces for the selected proteins. In the future, we plan to include information on well-characterized RNA-binding domains to allow users to select a domain of interest and obtain an analysis of the interfaces formed by the domain. We also plan to provide links to other databases, such as Prosite (Hulo et al., 2008) and NCBI’s Conserved Domain Database, CDD (Marchler-Bauer et al., 2007). Other features planned for PRIDB include interactive viewing of protein-RNA interfaces and searching by keyword and sequence similarity. Finally, an important new direction will be to expand PRIDB to include detailed information regarding the RNA side of protein-RNA interfaces.

**REFERENCES**


Chen, Y. C., and C. Lim (2008), Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry, *Nucleic acids research*.


Tong, J., P. Jiang, and Z. Lu (2008), RISP: A web-based server for prediction of RNA-binding sites in proteins, *Computer methods and programs in biomedicine.*


APPENDIX A. PREDICTING DNA-BINDING SITES OF PROTEINS FROM AMINO ACID SEQUENCE

A paper published in *BMC Bioinformatics*.

Changhui Yan, Michael Terribilini, Feihong Wu, Robert L. Jernigan, Drena Dobbs, and Vasant Honavar

**ABSTRACT**

**Background**

Understanding the molecular details of protein-DNA interactions is critical for deciphering the mechanisms of gene regulation. We present a machine learning approach for the identification of amino acid residues involved in protein-DNA interactions.

**Results**

We start with a Naïve Bayes classifier trained to predict whether a given amino acid residue is a DNA-binding residue based on its identity and the identities of its sequence neighbors. The input to the classifier consists of the identities of the target residue and 4 sequence neighbors on each side of the target residue. The classifier is trained and evaluated (using leave-one-out cross-validation) on a non-redundant set of 171 proteins. Our results indicate the feasibility of identifying interface residues based on local sequence information. The classifier achieves 71% overall accuracy with a correlation coefficient of 0.24, 35% specificity and 53% sensitivity in identifying interface residues as evaluated by leave-one-out cross-validation. We show that the performance of the classifier is improved by using sequence entropy of the target residue (the entropy of the corresponding column in multiple
alignment obtained by aligning the target sequence with its sequence homologs) as additional input. The classifier achieves 78% overall accuracy with a correlation coefficient of 0.28, 44% specificity and 41% sensitivity in identifying interface residues. Examination of the predictions in the context of 3-dimensional structures of proteins demonstrates the effectiveness of this method in identifying DNA-binding sites from sequence information. In 33% (56 out of 171) of the proteins, the classifier identifies the interaction sites by correctly recognizing at least half of the interface residues. In 87% (149 out of 171) of the proteins, the classifier correctly identifies at least 20% of the interface residues. This suggests the possibility of using such classifiers to identify potential DNA-binding motifs and to gain potentially useful insights into sequence correlates of protein-DNA interactions.

Conclusion

Naïve Bayes classifiers trained to identify DNA-binding residues using sequence information offer a computationally efficient approach to identifying putative DNA-binding sites in DNA-binding proteins and recognizing potential DNA-binding motifs.

BACKGROUND

Protein-DNA interactions play a pivotal role in gene regulation. The ability to identify amino acid residues that are responsible for the specificity and affinity of the interactions can significantly improve our understanding of macromolecular functions and contribute to advances in drug discovery [1,2]. Hence, the discovery of the principles of protein-DNA interactions has been a topic of significant interest for many years [3]. Current approaches to uncovering such principles rely on experimental analysis of the structures of protein-DNA complexes in order to understand the molecular details of specific residue-residue contacts that mediate protein-DNA recognition [4-6]. In addition to biophysical methods for structure determination, biochemical and molecular genetic approaches have been widely used to identify DNA-binding sites on proteins and to investigate the interaction
modes between proteins and DNA. For example, alanine-scanning mutagenesis has been used to identify the amino acids important for target recognition by the m5C methyltransferase [7] and to distinguish specific amino acids important for DNA binding and transcription activation by SoxS [8]. More recently, methods for precisely identifying protein-DNA contacts by coupling photochemical crosslinking with mass spectrometry have also been developed [9].

With increasing availability of protein sequence data, there is an urgent need for computational tools that can rapidly and reliably identify DNA-binding sites. Hence, there has been significant recent interest in developing computational methods for identification of amino acid residues that participate in protein-DNA interactions based on combinations of sequence, structure, evolutionary information, and chemical or physical properties. For example, Jones et al. [10] analyzed residue patches on the surface of DNA-binding proteins and used electrostatic potentials of residues to predict DNA-binding sites. They recently applied this method to the identification of three specific classes of DNA-binding proteins, based on the presence of solvent accessible DNA-binding structural motifs [11]. In related work, Tsuchiya et al. [12] used a structure-based method to identify protein-DNA binding sites based on electrostatic potentials and surface shape, and Keil et al. [13] trained a neural network classifier to identify patches likely to be DNA-binding sites based on physical and chemical properties of the patches. Neural network classifiers have also been used to identify protein-DNA interface residues based on a combination of sequence neighbor and structure information [14]. More recently, Ahmad and Sarai have proposed a sequence-based method for predicting DNA-binding residues that incorporates sequence alignment profiles into the input [15].

Against this background, this paper describes a machine-learning approach to developing a classifier for identifying amino acid residues that are likely to be involved in protein-DNA interactions.
RESULTS

Identification of interface residues based on local sequence information

A Naïve Bayes classifier was trained to predict whether or not a target residue in a protein sequence is an interface residue based on local protein sequence information. Several input encodings based on local sequence information were tried, with input consisting of: (a) the identities of 9 amino acid residues, corresponding to a window containing the target residue and 4 neighboring residues on each side of the target residue; and (b) the identities of 9 amino acid residues and the sequence entropy of the target residue (the entropy of the corresponding column in multiple alignment obtained by aligning the target sequence with its sequence homologs). In each case, Naïve Bayes classifiers were trained and evaluated using leave-one-out cross-validation on a set of 171 DNA-binding proteins.

Table A.1 shows that the classifier using amino acid identities as input achieved an overall accuracy of 71% with a correlation coefficient of 0.24, 35% of the residues predicted to be interface residues are actually interface residues, and 53% of interface residues are correctly identified. Adding the sequence entropy of the target residue (the entropy of the corresponding column in multiple alignment obtained by aligning the target sequence with its sequence homologs) to the input improved the performance of the classifier (Table A.1). The resulting classifier achieved an overall accuracy of 78% with a correlation coefficient of 0.28, 44% specificity, and 41% sensitivity. In 33% (56 of 171) of the proteins, the classifier recognizes the interaction site by correctly identifying at least half of the interface residues, and in 87% (149 of 171) of the proteins, by correctly identifying at least 20% of the interface residues.
Table A.1 The performance of the Naive Bayes classifiers.

<table>
<thead>
<tr>
<th></th>
<th>Identities (ID)</th>
<th>ID + Entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (%)</td>
<td>71</td>
<td>78</td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.24</td>
<td>0.28</td>
</tr>
<tr>
<td>Specificity (%)</td>
<td>35</td>
<td>44</td>
</tr>
<tr>
<td>Sensitivity (%)</td>
<td>53</td>
<td>41</td>
</tr>
</tbody>
</table>

Inclusion of other features of the target residue, including relative solvent accessibility, secondary structure, electrostatic potential, and hydrophobicity as additional inputs to the classifier did not yield performance improvements (data not shown) relative to the classifier trained using only the amino acid identities of the target residue and its sequence neighbors. Classifiers trained using features other than the amino acid identities of target residue and its neighbors as input achieved performance that was lower than that of the classifier using amino acid identities of the corresponding residues as input (data not shown).

**Evaluation of the predictions in the context of 3-dimensional structures of proteins**

We examined in the context of the 3-dimensional structures of the protein-DNA complexes, the DNA-binding residue predictions generated by a Naïve Bayes classifier trained to identify such residues based on the amino acid identities of the target residue and its sequence neighbors. Two representative examples are shown in figure A.1. Figure A.1A shows the predictions on the transcription factor C/Ebpβ from PDB complex 1gu4. The predictions of the classifier rank the 3rd best in terms of correlation efficient among the 171 proteins. We note that the classifier is able to recognize the DNA-binding site on the protein on the basis of sequence information alone. Figure A.1B shows the predictions on the intron-associated endonuclease I-TevI from PDB complex 1i3j. The predictions of the classifier in this case rank the114th best among the 171 proteins in terms of correlation efficient. I-TevI
wraps around the DNA and has an unusually extended binding site. We note that the predicted DNA-binding residues cover the long segment of the protein that binds to the DNA.

Figure A.1 Visualization of predicted DNA-binding residues on 3-D Structure. The predicted interface residues are shown in red on protein surface. DNA molecules bound to the proteins are shown in blue. A: The predictions on C/Ebpb from PDB complex 1gu4, the 3rd best out of the 179 proteins in terms of correlation coefficient. B: The predictions on I-TevI from PDB complex 1i3j, the 114th best out of the 179 proteins. Figures are generated using Protein Explorer [38].

**Receiver operating characteristic (ROC) curve**

In some situations (e.g., identification of critical interface residues for site-specific mutagenesis), it is desirable to predict interface residues with high precision at the cost of reduced coverage. In other situations, discovering more potential interface residues might be more useful. These different requirements can be met by modifying the threshold \( \theta \) used by the Naïve Bayes classifier in this study. The Naïve Bayes classifier predicts a residue to be an interface residue if

\[
\frac{P(c = 1 \mid X = x_1x_2\ldots x_n)}{P(c = 0 \mid X = x_1x_2\ldots x_n)} > \theta
\]
Figure A.2 shows the Receiver Operating Characteristic curve (ROC curve) of the DNA-binding site predictor.

Naïve Bayes classifier using only local sequence identities as input can discover DNA-binding motifs

The results summarized above show that a Naïve Bayes classifier trained on a set of DNA-binding proteins can successfully identify protein-DNA interface residues from amino acid sequence. This raises the question as to how the sequence features that are identified as predictive of DNA-binding residues by Naïve Bayes classifier relate to known DNA-binding
motifs. To explore this question, we used the ps_scan program to search for PROSITE motifs in our data set of 171 DNA-binding proteins. PROSITE motifs were found in 53 of the 171 proteins (a total of 73 hits). Of these 73 hits, 61 overlap with actual protein-DNA binding sites. The DNA-binding site predictions produced by the Naïve Bayes classifier (in the leave-one-out cross-validation setting) using the identities of a window of 9 residues and the sequence entropy of the target residue as input, substantially overlap with 56 of the 61 PROSITE DNA-binding motifs (Figure A.3). It is worth noting that 118 of the 171 DNA-binding proteins in our data set contain no PROSITE motif whose annotation suggests a role in protein-DNA interactions. PROSITE motifs cover more than 50% of interface residues in only 11% (18 out of 171) of the proteins and cover at least 20% of interface residues in only 20% (34 out of 171) of the proteins. In contrast, the Naïve Bayes classifier identifies at least 50% of the interface residues in 33% (56 out of 171) of the proteins and at least 20% of the interface residues in 87% (149 out of 171) of the DNA-binding proteins used in this study. These results suggest the possibility of using a Naïve Bayes classifier trained to predict DNA-binding residues to identify putative DNA-binding motifs.

>1dh3A
Sequence : KREVRLMKNREARERSSRRKKEYVKSLENRVAVLQNKTLIEELKALDKLYSHK
Interface : * *** **** *** ***
Predictions: * * * * ********
Motif : *******************

BZIP_BASIC

Figure A.3 Comparison of actual and predicted DNA-binding site residues for transcription factor CREB (PDB 1dh3A). PROSITE motif BZIP_BASIC (bottom row) covers many of the actual interface residues (the first row below sequence). Note that the predictions of Naïve Bayes classifier (the second row below sequence) overlap with the PROSITE motifs, but more closely correspond to the actual interface residues.
Comparison with previously published methods

Ahmad and Sarai have developed a Position Specific Scoring Matrix (PSSM) based neural network classifier for predicting DNA-binding sites [15]. To the best of our knowledge, this is the only previously published study which reports the performance of a DNA-binding site prediction using only sequence information on a "per residue" basis. Ahmad and Sarai have made available an online server that predicts DNA-binding residues using a PSSM-based neural network classifier [16]. The server makes predictions for protein sequences that are 40 to 200 amino acid residues in length. In our data set of 171 DNA-binding proteins, 86 have length in this range. The predictions of the PSSM-based classifier on these 86 proteins were obtained by submitting the sequences to the online server. The server returns, for each residue in the submitted sequence, the estimated probability that the residue is a DNA-binding residue. These probabilities can be compared with a threshold to obtain a prediction as to whether a residue is a DNA-binding residue. Different choices of threshold yield different predictions. We varied the threshold from 0.01 to 0.99 in increments of 0.02 to generate an ROC curve for the PSSM-based neural network classifier. For comparison, we trained and evaluated using leave-one-out cross-validation, a Naïve Bayes classifier using as input the identities of 9 amino acid residues on the subset of 86 proteins (ranging from 40 to 200 amino acids in length). Figure A.4 shows the comparison of the ROC curves of the PSSM-based neural network classifier with that of the Naïve Bayes classifier on the data set of 86 proteins. The results show that the Naïve Bayes classifier achieves higher hit rate, for any given choice of the false alarm rate, than the current implementation of the PSSM-based neural network classifier in the online server.
Naïve Bayes classifier uses the identities of 9 amino acid residues as input. The ROC for the Naïve Bayes classifier is obtained using Weka on 86 DNA-binding proteins with lengths ranging from 40 to 200 residues with pairwise sequence similarity less than 30%. The ROC for the PSSM-based classifier is generated using the true positive, false positive, true negative, and false negative predictions obtained by submitting the 86 sequences to the online server [16] that implements PSSM-based classifier developed by Ahmad and Sarai [15].

Identification of DNA-binding residues in type I restriction-modification system

Restriction-modification (R-M) systems play important role in the recognition and elimination of foreign DNA. In type I R-M systems, S subunit determines the specificity of DNA recognition. The interaction mode between S subunit and DNA is still unknown. Recently, Kim et al. [17] solved the crystal structure of the S subunit from \textit{M. jannaschii}, the only crystal structure ever reported for the S subunit of type I (R-M) systems. To further
evaluate the Naïve Bayes classifier, we used the classifier trained on our data set of 171 DNA-binding proteins (using identities of the target residue, and 4 sequence neighbors on either side along with the sequence entropy of the target residue as input) to identify DNA-binding residues on the S subunit of the type I R-M system from *M. jannaschii*. Figure A.5 shows the predicted DNA-binding residues in red and spacefill. Note that Kim et al. [17] reported, based on the solved crystal structure of the S subunit of *M. jannaschii*, that the structures of the two target recognition domains (TRD1, residue 1–168 and TRD2, residue 209–378) of the S subunit are similar to the DNA binding domain of TaqI-MTase. By aligning the structures of TRD1 and TRD2 with the structure of TaqI-MTase/DNA complex, Kim et al. [17] proposed a model for the interaction between the S subunit and DNA. In figure A.5, the DNA molecules in Kim's model are shown in blue. Comparison of Kim's model with the DNA-binding site predictions produced by our Naïve Bayes classifier shows that the Naive Bayes classifier agrees with the locations of the two potential DNA-binding sites on the S subunit in Kim's interaction model.
Predictions of the Naïve Bayes classifier on proteins for which there is no experimental evidence suggesting a DNA-binding role

Given that the Naïve Bayes classifier was trained to identify DNA-binding residues in proteins that are known to bind to DNA, it is interesting to examine their predictions on a set of proteins for which at present, there is no evidence suggesting a DNA-binding role. We assembled a non-redundant data set of 2,323 proteins which, based on our analysis of Gene Ontology annotations, appear to have no evidence suggesting a DNA-binding role. A Naïve Bayes classifier trained on our data set of 171 DNA-binding proteins to identify the DNA-
binding residues (using amino acid identities of the target residue and its sequence neighbors together with the sequence entropy of the target residue as input) was applied to the 2,323 proteins with no known DNA-binding role. The Naïve Bayes classifier predicted 11% of the 613,754 residues from these 2,323 proteins as potentially DNA-binding residues. It would be inappropriate to conclude that 11% is a per residue basis false positive rate of our classifier because absence of DNA-binding evidence in GO annotation does not necessarily imply that the protein in question does not have a DNA-binding role. It is quite possible that at least some of these 2,323 proteins indeed bind to DNA. It should be emphasized that our classifier was not trained to distinguish the class of DNA-binding proteins from those that are not DNA-binding (Training such a classifier would involve using representatives of both DNA-binding and non DNA-binding proteins in the training set). It is interesting to note that in 156 of the 2,323 proteins, no residues were predicted to be DNA-binding by our classifier; 264 had fewer than 5 predicted DNA-binding residues; 502 had fewer than 10 predicted DNA-binding residues, and 999 with fewer than 20 DNA-binding residues. Exploring the implications of these observations would require experimentally testing some of the proteins on which our Naïve Bayes classifier predicts putative DNA-binding sites for DNA-binding activity. Another potentially interesting direction would be to train classifiers to distinguish proteins that are DNA-binding (without necessarily identifying the DNA-binding residues) from those that are not.

**DISCUSSION**

**Effectiveness of local amino acid sequence based approach to prediction of putative DNA-binding sites**

In this paper, we have described a computationally efficient approach to identifying putative DNA-binding residues of DNA-binding proteins using Naïve Bayes classifiers trained to predict DNA-binding residues using amino acid identities of the target residue and
its sequence neighbors. The resulting classifier achieves 71% overall accuracy with a correlation coefficient of 0.24, 35% specificity and 53% sensitivity in identifying interface residues as evaluated by leave-one-out cross-validation. Our results indicate the feasibility of identifying interface residues based on local sequence information alone.

We found that the performance of the classifier is improved by using sequence entropy of the target residue (the entropy of the corresponding column in multiple alignment obtained by aligning the target sequence with its sequence homologs) as additional input. This observation is consistent with the suggestion that DNA-binding residues are likely to be conserved (because of their function). The resulting classifier achieves 78% overall accuracy with a correlation coefficient of 0.28, 44% specificity and 41% sensitivity in identifying interface residues.

Incorporating additional structure-derived information such as solvent accessibility, electrostatic potential, hydrophobicity or secondary structure of the target residue as additional input, however, did not improve the performance in this study. This should not be taken to mean that these features are not useful predictors of a residue's functionality. In particular, electrostatic potential has been shown to be useful in identification of protein-DNA interface residues [10,11]. The fact that this information does not improve performance of our Naïve Bayes classifiers might have to do with the properties of input encoding or the classification method. Specifically, the additional features were simply added as additional input. The underlying assumption of the Naïve Bayes classifier that the inputs are independent given the class almost certainly does not hold in the case of protein sequences. Hence, more systematic analysis is needed to identify features that are useful for identification of interface residues and develop methods of representing them in input to a broad range of classifiers. Jones and Thornton [18] analyzed six features of surface patches in protein-protein interaction sites and developed an approach to identify protein-protein interfaces based on the scores combining the six features. Sen et al. [19] developed an
ensemble method to identify protease-inhibitor binding sites based on sequence, structure and evolution information. It would be interesting to explore such methods for computational prediction of protein-DNA interfaces.

Comparison of Naïve Bayes classifier with a PSSM-based neural network classifier

Ahmad and Sarai [15] used a PSSM-based neural network classifier to identify interface residues in protein-DNA interactions. Our comparison of the PSSM-based classifier with the Naïve Bayes classifier shows that the Naïve Bayes classifier achieves higher hit rate than the PSSM-based classifier for any given choice of the false alarm rate.

We note that the PSSM-based classifier's ROC originally reported by Ahmad and Sarai [15] is better than the PSSM-based classifier's ROC achieved by their online server [16] on the data set used in our comparison. A few factors may have contributed to this difference: (1) the data set used by Ahmad and Sarai in their original study is different from the data set of 86 proteins used here. It is possible that the current implementation of the PSSM-based method is well optimized for their original data set, but not for the 86 proteins used here; (2) the ROC reported by Ahmad and Sarai includes predictions on proteins of all lengths, whereas the online server only makes predictions for proteins with a length in the range of 40–200. We chose to compare the Naïve Bayes classifier with the online server because the server is publicly available and it provides the raw probabilities of the predictions making it possible to compare the ROC curves of the two classifiers on the same data set. However, it should be noted that in the case of Naïve Bayes classifier, our use of leave-one-out cross-validation ensures that the training and test data do not overlap. We have no control over the training data used by the PSSM-based classifier. Nevertheless, a comparison of the two ROC curves suggests that the Naïve Bayes classifier achieves higher
hit rate than the current implementation of the PSSM-based neural network classifier for any
given choice of the false alarm rate.

A thorough assessment of the performance of the Naïve Bayes classifier relative to
the PSSM-based classifier requires systematic comparisons using leave-one-out cross-
validation on identical data sets – which is at present, not feasible without access to an
implementation of the algorithm and the precise parameter settings used to train the PSSM-
based classifier. Plans are underway to perform such a comparison using identical data sets
and evaluation procedures, in collaboration with Ahmad and Sarai.

It should be noted that the Naïve Bayes classifier described in this paper offers
several advantages over the PSSM-based neural network classifier: (a) The Naïve Bayes
classifier can be trained in a single pass through the training data whereas training a neural
network classifier requires many, often hundreds of passes through the training data. (b)
Training the Naïve Bayes classifier, unlike the neural network classifier, requires no time-
consuming and computationally expensive exploration of many possible choices of network
architecture (e.g., number of hidden neurons) and parameter settings (e.g., learning rate). (c)
The Naïve Bayes classifier, as well as predictions generated by it is amenable to a
straightforward probabilistic interpretation whereas the neural network classifier is more of a
"black box".

These advantages, together with the superior performance of the Naïve Bayes
classifier relative to the current implementation of the PSSM-based neural network classifier,
make it an attractive alternative to the latter in identifying DNA-binding residues from a
protein sequence. However, the neural network classifier is not limited by the strong
independence assumption of the Naïve Bayes classifier. Hence, it would be interesting to
explore whether a neural network classifier or a variant of it could be optimized to yield
results that are better than that of the simple Naïve Bayes classifier.
Use of Naïve Bayes classifiers to identify putative novel DNA-binding motifs

Protein sequence motifs (defined here as sequence segments associated with specific protein functions or structural families) are often used to identify putative DNA-binding domains. Discovery of such motifs requires alignment of protein sequences that are known to have the same or similar functions. Generating multiple sequence alignments that reveal useful sequence motifs requires significant human expertise to identify a suitable set of sequences to be aligned and to manually refine, through an iterative process of trial and error, the multiple sequence alignment. Against this background, it is interesting to note that in 118 out of 171 DNA-binding proteins used in this study, we found no PROSITE motifs whose annotations suggest a possible DNA-binding role. In the remaining proteins, 61 PROSITE motifs were found to overlap with protein-DNA binding sites. The DNA-binding sites predicted by the Naïve Bayes classifier significantly overlapped with 56 of the 61 PROSITE motifs that overlapped with DNA-binding sites. PROSITE motifs cover at least 20% of the DNA-binding residues in only 20% (34 out of 171) of the proteins. In contrast, the Naïve Bayes classifier identifies at least 20% of the interface residues in 87% (149 out of 171) of the DNA-binding proteins used in this study. This raises the possibility of identifying novel sequence motifs that correspond to protein-DNA interfaces by using a Naïve Bayes classifier trained to identify protein-DNA binding sites. More systematic comparison of this approach with alternative approaches to identification of putative DNA-binding motifs using other motif libraries and different motif finding methods is needed to evaluate its efficacy relative to other approaches.

CONCLUSION

In previous work, we have used similar approaches to identify interface residues involved in protein-protein interactions [20,21] and protein-RNA interactions [22]. Here we show that it is also feasible to identify interface residues involved in protein-DNA interaction
using sequence information. With the level of success achieved in this study, putative DNA-binding sites predicted by the classifiers trained using a machine-learning approach should be useful for guiding experimental investigations into the role of specific residues of a protein in its interaction with DNA, e.g., by localizing candidate residues for alanine-scanning mutagenesis [7,8]. Moreover, analysis of the binding site "rules" generated by classifiers may provide valuable insight into the protein-DNA recognition code responsible for the specificity and affinity of protein-DNA interactions in living cells.

METHODS

Data sets

DNA-binding proteins: A data set of DNA-binding proteins was extracted from structures of known protein-DNA complexes in the Protein Data Bank [23]. The dataset was culled using PISCES [24]. The resulting dataset consists of 171 proteins with mutual sequence identity <= 30% and each protein has at least 40 amino acid residues. All the structures have resolution better than 3.0 Å and R factor less than 0.3.

Proteins that do not have evidence of a DNA-binding role: A non-redundant set of proteins with mutual identity less than 30% was extracted from the PDB using the cluster file from the Protein Data Bank [25]. Structures with resolution worse than 2.5 Å were removed. The annotations for each protein were retrieved from the Gene Ontology Annotation (GOA) [26]. Proteins with annotations indicative of a DNA-binding role were eliminated, leaving a data set of 2,313 proteins with no evidence of a DNA-binding role.

Definition of interface residues

Interface residues are defined as described in Jones et al. [10]. Accessible surface area (ASA) was computed for each residue in the unbound protein (in absence of DNA) and in the protein-DNA complex using NACCESS [27]. A residue is defined to be an interface residue
if its ASA in the protein-DNA complex is less than its ASA in the unbound protein by at least 1Å². The 171 proteins have 38,649 residues in total and 5,050 of them are interface residues.

**Naïve Bayes classifier**

We used the Naïve Bayes implementation in the Weka package from the University of Waikato, New Zealand [28,29]. For each input target residue, the classifier produces a Boolean output (with 1 denoting an interface residue and 0 denoting a non-interface residue). The Naïve Bayes classifier assumes independence of the attributes given the class. The Naïve Bayes classifier performs as well as more sophisticated methods on many classification tasks [30]. For an input $X = x_1 x_2 ,...,x_n$ , a Naïve Bayes classifier assigns it a class label $c$ by optimizing the posterior:

$$c = \arg \max_c P(c \mid X = x_1 x_2 ... x_n) = \arg \max_c P(c) \prod_{i=1}^{n} P(x_i \mid c)$$

In the case of two class classification ($c \in \{0, 1\}$), this is equivalent to determining $c$ by comparing the ratio likelihood with a parameter $\theta$ as in equation (1).

$$\frac{P(c = 1 \mid X = x_1 x_2 ... x_n)}{P(c = 0 \mid X = x_1 x_2 ... x_n)} = \frac{P(c = 1) \prod_{i=1}^{n} P(x_i \mid c = 1)}{P(c = 0) \prod_{i=1}^{n} P(x_i \mid c = 0)} > \theta \quad (1)$$

$c$ is predicted to be 1 if the ratio likelihood is greater than $\theta$, and 0 otherwise. When a local sequence around the target residue was encoded using numeric features such as hydrophobicity, the numerical values were discretized using the discretization filter of Weka.
In a standard Naïve Bayes classifier, $\theta$ takes the value of 1. The predictions of Naïve Bayes classifier are biased in favor of the majority class when the dataset consists of unequal numbers of examples for the two classes. Hence, we trained $\theta$ to optimize classification performance on training data. We used leave-one-out cross-validation to train and test the classifier. In each round of experiment, all proteins except one were used as the training set and the remaining protein was used to test the classifier. In the training stage, the conditional probability table $P(x_i | c)$ and prior probability $p(c)$ were estimated using the training set. To determine $\theta$, the classifier was applied to the training set and different values of $\theta$ ranging from 0.01 to 1 were tested, in increments of 0.01. The value of $\theta$ for which the classifier yields the highest correlation coefficient was used to make predictions on the test set.

**Naïve Bayes classifier using only local sequence identity as input**

The input to the Naïve Bayes classifier contains the identities of $2n+1$ residues in the form of $X = (x_{t-n}, x_{t-n+1}, ..., x_t, x_{t-1}, ..., x_{t+n-1}, x_{t+n})$, where $x_t$ is the identity of target residue, $x_{t-n}, x_{t-n+1}, ..., x_{t-1}$ and $x_{t+1}, x_{t+n-1}, x_{t+n}$ are the identities of $n$ residues on each side of the target residue. Different values of $n$ from 1 to 10 were tried and the best performance was obtained when $n = 4$ (corresponding to a window size of 9). A training example is an ordered pair $(X, c)$, where $c \in \{0, 1\}$. 1 indicates that the target residue (the residue in the center of the input window) is an interface residue and 0 indicates that target residue is not an interface residue. For a test example $X$, the classifier outputs 1 (i.e., $X$ is predicted to be an interface residue) or 0 (i.e., $X$ is predicted to be a non-interface residue) as the class label of $X$.

**Naïve Bayes classifier using additional inputs**

Relative solvent accessibility (rASA), sequence entropy, secondary structure, electrostatic potential and hydrophobicity were considered. When a feature of the target residue is added into the input of amino acid identities of residues in a 9-residue window, the
input to the classifier is encoded as $X = (x_{t-n}, x_{t-n+1}, ..., x_{t-1}, x_t, x_{t+1}, ..., x_{t+n-1}, x_{t+n}, f_t)$, with $f_t$ standing for the corresponding feature of the target residue (e.g., sequence entropy, hydrophobicity, etc.), and $x_i$ denotes the amino acid identity of the corresponding position within the sequence window. When a feature other than residue identity of the input window (i.e., the target residue and its sequence neighbors) is used to encode the local sequence around the target residue, the input to the classifier has the form of $X = (f_{t-n}, f_{t-n+1}, ..., f_{t-1}, f_t, f_{t+1}, ..., f_{t+n-1}, f_{t+n})$, where $f_i$ is the corresponding feature (e.g., hydrophobicity) of the residue $i$.

The relative solvent accessible surface area (rASA) of each residue (in the absence of DNA) was computed using NACCESS [27]. Entropy of each sequence position (the sequence entropy for the corresponding column in multiple of the multiple sequence alignment) was extracted from the HSSP database [31]. The sequence entropy is normalized to the range of 0–100, with lower entropy values corresponding to more conserved sequence positions. Secondary structure for each residue was extracted from the PDB database [25]. Electrostatic potential for each atom was calculated using Delphi [32,33], using parameters based on the study of Jones et al. [10]. The electrostatic potential for each residue was calculated in a similar way as the study of Jones et al. [10]: the electrostatic potential of an atom is set to 0 if its solvent accessibility is less than 1Å$^2$ and the electrostatic potential of a residue is the average over all its atoms. Hydrophobicity of each residue is obtained from the consensus normalized hydrophobicity scale derived by Eisenberg et al. [34].

**Performance measures**

Because no single performance measure provides a complete picture of performance of the classifier [35], we used a combination of accuracy, correlation coefficient (CC), specificity and sensitivity. These measures are defined as described in Baldi et al. [35].
where TP = the number of true positives (residues predicted to be DNA-binding residues that are in fact interface residues); FP = the number of false positives (residues predicted to be DNA-binding residues that are in fact not interface residues); TN = the number of true negatives (residues predicted to be non DNA-binding residues that are in fact not DNA-binding residues); FN = the number of false negatives (residues predicted to be non DNA-binding residues that are in fact DNA-binding residues); N = TP + TN + FP + FN (the total number of examples).

Sensitivity is the fraction of positive examples (DNA-binding residues) that are predicted as such by the classifier. Specificity is the fraction of positive predictions (residues predicted to be DNA-binding residues) that are actually interface residues. Accuracy is the fraction of overall predictions that are correct. Correlation coefficient measures the correlation between predictions and actual class labels.

The Receiver Operating Characteristic curve (ROC curve) is a plot of the "hit rate" (TP/(TP+FN)) versus the "false alarm rate" (FP/(TN+FP)) [35]. It shows the tradeoff between hit rate and false alarm rate when different threshold values are used for the classifier.

**Identifying PROSITE motifs in protein sequences**

The PROSITE motif database was downloaded from the PROSITE [36]. Protein sequences were scanned using the ps-scan program [37] to identify motifs. Frequently matching (unspecific) patterns and profiles were omitted by setting the "-s" and "-r" options of ps-scan.
ACKNOWLEDGEMENTS

This Research was supported in part by a grant from the National Institutes of Health (GM 066387) to VH, DD, and RLJ. We thank O. Yakhnenko and D. Caragea for providing comments on the manuscript. We thank Dr. S. Ahmad and Dr. A. Sarai for sharing the details of their PSSM-based neural network classifier.

REFERENCES


7. Radlinska M, Kondrzycka-Dada A, Piekarowicz A, Bujnicki JM: Identification of amino acids important for target recognition by the DNA:m5C methyltransferase
M.NgoPII by alanine-scanning mutagenesis of residues at the protein-DNA interface. 

*Proteins* 2005, **58**:263-270.


15. Ahmad S, Sarai A: **PSSM-based prediction of DNA binding sites in proteins.** *BMC Bioinformatics* 2005, **6**:33.

16. **Prediction of DNA-binding residues by PSSM and sequence homology.**

http://wwwnetasaorg/dbs-pssm/.


22. Terobilini M, Lee JH, Yan C, Jernigan RL, Honavar V, Dobbs D: **Prediction of RNA-binding sites in proteins based on amino acid sequence.** Submitted


27. Hubbard SJ: NACCESS. Department of Biochemistry and Molecular Biology,

28. Witten IH, Frank E: Data mining: practical machine learning tools and


30. Buntine W: Theory refinement on Bayesian networks: ; Los Angeles, CA. ;
1991:52-60.

31. Sander C, Schneider R: Database of homology derived protein structures and

32. Rocchia W, Alexov E, Honig B: Extending the applicability of the nonlinear
Poisson-Boltzmann equation: multiple dielectric constants and multivalent ions. *Journal

33. Rocchia W, Sridharan S, Nicholls A, Alexov E, Chiabrera A, Honig B: Rapid
grid-based construction of the molecular surface for both molecules and geometric
objects: applications to the finite difference Poisson-Boltzmann method. *Journal of
34. Eisenberg D, Weiss RM, Terwilliger TC: **The hydrophobicity moment detects periodicity in protein hydrophobicity.** *Proc Natl Acad Sci USA* 1984, **81**:.


APPENDIX B. IDENTIFYING INTERACTION SITES IN "RECALCITRANT" PROTEINS: PREDICTED PROTEIN AND RNA BINDING SITES IN REV PROTEINS OF HIV-1 AND EIAV AGREE WITH EXPERIMENTAL DATA


Michael Terribilini, Jae-Hyung Lee, Changhui Yan, Robert L. Jernigan, Susan Carpenter, Vasant Honavar, and Drena Dobbs

INTRODUCTION

The human AIDS virus, Human immunodeficiency virus Type 1 (HIV-1), is closely related to a number of lentiviruses that cause persistent, insidious infections in other primates and domestic animals. Recent advances in molecular virology have resulted in novel antiviral therapies that inhibit specific proteins required for the replication of lentiviruses and other important retroviruses. Rev is a multifunctional regulatory protein that plays an essential role in the production of infectious virus (1, 2) and, as such, is an attractive target for new antiviral therapies. To date, however, no Rev-targeted drugs for AIDS therapy are available.

Rev is known to participate in protein-protein interactions with several cellular proteins as well as in RNA-protein interactions with lentiviral RNAs (3, 4). It is required for the transition to the late stage of viral replication and facilitates export of incompletely spliced viral RNAs from the nucleus to the cytoplasm. After its import into the nucleus, HIV-1 Rev binds a structure in the viral pre-mRNA called the Rev-responsive element (RRE) (5, 6), multimerizes (6, 7), then utilizes the CRM1 nuclear export pathway to redirect movement of incompletely spliced viral RNA out of the nucleus (8). As shown in Figure B.1, functional domains within HIV-1 Rev are known to mediate interactions with viral RNA and with host
cell proteins that are required for nuclear localization, RNA binding, multimerization, and nuclear export (3).

Figure B.1 Functional domains of HIV-1 and EIAV Rev proteins. The linear organization of functional domains within the two Rev proteins differs significantly, but both have been shown to contain specific sequences involved in Rev interactions with proteins (MUL, NLS, NES) or RNA (RBD, ARMs).

Efforts to develop inhibitors of Rev activity have been hampered by a lack of information regarding Rev protein structure. A major stumbling block for structural analysis is the tendency of Rev to aggregate at concentrations needed for crystallization or solution NMR studies (9). The only high resolution information available is for short peptide fragments of HIV-1 Rev. In an NMR solution structure of a 23 amino acid fragment of Rev bound to a 34 nucleotide RRE RNA fragment, the Rev peptide adopts an $\alpha$-helical
conformation and is bound in the major groove of the RNA (10). Structures of other critical functional domains of Rev (e.g., nuclear localization, multimerization, export) have not been reported. Furthermore, it has not been possible to apply homology modeling approaches to gain insight into Rev structure because Rev has no detectable sequence similarity to any protein of known structure. Indeed, despite their apparently conserved functions, Rev protein sequences are highly variable between species, with < 10% sequence identity between HIV-1 and one of the most divergent Rev proteins, equine infectious anemia virus, (EIAV) Rev (11).

When protein structures cannot be solved using experimental approaches, computational analyses can provide valuable insight into protein structure-function relationships and aid in identification of key functional residues that may offer tractable targets for therapeutic intervention in disease (12). Here we describe the identification of critical residues that mediate protein-protein and protein-RNA interactions in Rev, using machine learning approaches that rely on the primary amino acid sequence of Rev, but do not require any information regarding its structure or the sequence or structure of its interaction partners. Our predictions are in good agreement with previously published biochemical, biophysical and genetic data for HIV-1 and EIAV Rev as well as with our recent experimental mapping of RNA binding sites in EIAV Rev (13). Taken together, these results demonstrate the utility of sequence-based approaches for identifying putative binding sites of proteins with potential therapeutic value that are, at present, recalcitrant to experimental structure determination.

Datasets, Materials and Methods

Datasets

Protein-protein binding site dataset (PBS). We extracted individual proteins from a set of 70 protein–protein heterocomplexes used in the study of Chakrabarti and Janin (14).
After removal of redundant proteins and molecules with fewer than 10 residues, we obtained a dataset of 77 individual proteins with sequence identity <30%. The dataset contains a total of 12,719 amino acids, of which 2340 (18.4%) are interface residues (positive examples).

RNA-protein binding dataset (RBS). A dataset of protein-RNA interactions was extracted from structures of known protein-RNA complexes in the Protein Data Bank (PDB) (15). Proteins with >30% sequence identity or structures with resolution worse than 3.5Å were removed using PISCES (16). This resulted in a set of 109 non-redundant protein chains containing a total of 25,118 amino acids. Amino acids in the protein-RNA interface were identified using ENTANGLE (17). Using default parameters, 3518 (14%) of the amino acids in the dataset are defined as interface residues (positive examples).

**Protein-protein interface residue prediction**

We have previously developed a two-stage classifier for predicting interface residues in protein-protein complexes (18). In the first stage, a Support Vector Machine (SVM), trained on the PBS dataset, is used to classify each residue as interface or non-interface. Input to the SVM is a window of nine amino acid identities. Because interface residues tend to be clustered in primary sequence, a second stage was introduced to take advantage of this to improve predictions. In the second stage, a Bayesian Network classifier is trained based on the predictions of the target residue and its neighbors from the first stage SVM. Let $C \in \{0,1\}$ denote the actual class label of a residue; $X \in \{0,1\}$ be the prediction of the SVM classifier; $Y$ denote the number of predicted interface residues within 4 amino acids of the target residue. For each residue, the likelihood that it is an interface residue given the SVM predictions for itself and its neighbors is calculated and compared to a chosen threshold $\theta$ as formula 1.
The residue is predicted to be an interface residue if the likelihood is larger than \( \theta \) and non-interface otherwise. The conditional probability table \( P(C|X,Y) \) is derived from training datasets. To determine \( \theta \), the classifier was applied to the training set and different values of \( \theta \) ranging from 0.01 to 1 were tested, in increments of 0.01. The value of \( \theta \) for which the classifier yields the highest correlation coefficient was used to make predictions on the Rev proteins.

**Protein-RNA interface residue prediction**

We have previously developed a Naïve Bayes (NB) classifier for predicting which amino acids in a given protein are likely to be found in protein-RNA interfaces (19), using the NB classifier from the Weka package (20). The input is a window of 25 contiguous amino acid identities. The output is an instance where + indicates that the target residue is an interface residue and – indicates a non-interface residue. A training example is an ordered pair \( (x, c) \) where \( x = (x_{-n}, x_{n-1}, ..., x_{-1}, x_0, x_1, ..., x_n) \) and \( c \) is the corresponding class label (interface or non-interface). A training data set \( D \) is a collection of labeled training examples.

Let \( X = (X_{-n}, ..., X_0, ..., X_n) \) 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 Naïve Bayes classifier assigns input \( x \) the class label + (interface) if:

\[
\frac{P(C = 1 | X, Y)}{P(C = 0 | X, Y)} > \theta
\]  

(1)

The value of \( \theta \) for which the classifier yields the highest correlation coefficient was used to make predictions on the Rev proteins.
and the class label – (non interface) otherwise. $\theta$ was set to the value that optimized the correlation coefficient (21) on the training set in each leave-one-out cross validation experiment.

**Experimental mapping of RNA binding sites**

Details of our experimental mapping of RNA binding sites are provided in Lee et al., (13). Briefly, Maltose Binding Protein-EIAV Rev (MBP-ERev) constructs containing deletions or point mutations in the EIAV Rev coding region were cloned in pHMTc, based on the pMal-c2x expression vector, which enhances solubility of Rev fusion proteins. MBP-ERev fusion proteins were expressed in E. coli, purified prior to use in RNA binding experiments. UV cross linking experiments were used to quantitate the effects of mutations on Rev RNA binding activity (13).

**RESULTS**

**Binding site predictions on datasets of known protein-protein and protein-RNA complexes**

In previous work, we have developed classifiers for predicting interface residues in protein-protein, protein-DNA and protein RNA complexes (18, 19, 22), typically using a combination of sequence and structure-derived information as input. In choosing classifiers for the task of predicting protein-protein and protein-RNA interface residues in Rev proteins, we compared several types of classifiers for predicting each type of interface residue (data not shown). Table B.1 shows an example of the classification performance values obtained for protein binding site prediction using the PBS dataset, which contains 77 proteins used in our previous study (18) and for RNA binding site prediction using the RBS dataset, which contains 109 RNA-binding proteins (19).
Table B.1 Classification performance in predicting protein-protein and RNAprotein binding site residues, using leave-one-out experiments.

<table>
<thead>
<tr>
<th>Classification Performance Measure</th>
<th>Protein Interface Residues (2-stage classifier)</th>
<th>RNA Interface Residues (NB classifier)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>72%</td>
<td>85%</td>
</tr>
<tr>
<td>Specificity</td>
<td>58%</td>
<td>51%</td>
</tr>
<tr>
<td>Sensitivity</td>
<td>39%</td>
<td>38%</td>
</tr>
<tr>
<td>Correlation Coefficient</td>
<td>0.30</td>
<td>0.35</td>
</tr>
</tbody>
</table>

These results were obtained using a modified 2-stage classifier developed in this work to predict protein interface residues (see Methods) and a Naive Bayes classifier published previously (19) to predict RNA interface residues. The results of the latter study are reproduced here for comparison.

**Predicted binding sites in wildtype HIV-1 and EIAV Rev proteins**

Using classifiers trained on the datasets described above, we predicted protein-protein and protein-RNA interface residues in Rev proteins from HIV-1 and EIAV. As shown in Figure B.2A, the 2-stage protein classifier predicted a total of 56 protein-protein interface residues (indicated by "p") within the 116 amino acid HIV-1 Rev sequence. These are primarily located in 5 clusters consisting of 6-15 amino acids. The Naive Bayes classifier predicted a total of 26 RNA-protein interface residues (indicated by "r"), located in a single large cluster near the N-terminus of the protein. The predicted RNA binding site sequence is PPNPEGTRQARRRRRWRERQRIHSIG, corresponding to amino acids 28-56. Ile26 and Ile29 are the only two residues within this sequence that are predicted to be non-interface residues.
Figure B.2 Predicted interface residues in Rev proteins. The protein sequences (SEQ) for A) HIV-1 Rev & B) EIAV Rev are shown on top line, with binding site residues for protein (PRO) and RNA shown by "p" or "r" on the lines below. Important functional domains boxed in the sequence are: **NES, NLS/ARM, RBD, MULTIMERIZATION, MULTIFUNCTIONAL, ARM, UNKNOWN.**

The prediction results for EIAV Rev, using the same classifiers, are shown in Figure B.2B. A total of 79 protein-protein interface residues were predicted in the 165 amino acid protein. In EIAV Rev, most of these predicted protein-binding residues are also located in 5 clusters that are somewhat larger (8-24 amino acids) than those predicted in HIV-1. There are two predicted clusters of RNA-protein interface residues, one consisting of 15 contiguous amino acids, located in the central region and a second consisting of 19 contiguous residues at the C-terminus of the protein. The predicted RNA binding site sequences are RHLGPGPTQHTPSRR, (aa 63-77) and QSSPRVLPGDSKRRKHL (aa 147-165. The only other predicted interface residues are 5 scattered amino acids in the region of aa 113-133.
Comparison of predicted Rev binding sites with experimental data

Functional domains in HIV-1 Rev have been extensively interrogated through the analysis of sequence variants and mutants generated both in vivo and in vitro (4). These experimental results are summarized in Figure B.1 and mapped onto amino acid sequence of HIV-1 Rev for comparison with our predicted RNA and protein interface residues in Figure B.2A. Notably, the single cluster of RNA interface residues predicted by the Naive Bayes classifier closely matches the experimentally mapped RNA binding domain (RBD), which in HIV-1 also includes an Arginine Rich Motif (ARM) that also functions as a nuclear localization signal (NLS). Three predicted clusters of protein interface residues also characterized protein binding sites: one cluster (aa 22-32) maps to Rev multimerization domain, and two clusters are located within a large C-terminal domain (aa 87-116) that has been shown to play multiple roles in nuclear export, dimerization and transactivation activities of HIV-1 Rev (23). One of these clusters (aa75-93) also overlaps with the modular nuclear export signal (NES), which is interchangeable between various lentiviruses, including HIV-1 and EIAV (24).

Although the functional domains in EIAV Rev have been studied in less detail than those in HIV-1 Rev, previous biochemical and genetic studies had localized the NLS and NES domains and implicated two motifs in the central region in RNA binding, RRDRW and ERLE (Figure B.1) (13, 25-28). In predictions generated before we initiated our experimental mapping of EIAV RNA-binding domains, the Naive Bayes classifier identified one potential RNA-binding region overlapping the RRDRW motif and another overlapping a KRRRK motif within the mapped C-terminal NLS domain, but did not predict any interface residues near the ERLE motif. Our recent direct mapping of the RNA binding domain of EIAV Rev by UV cross linking showed that two separate regions of Rev are necessary for RNA binding: a central region encompassing aa 75-127 and a region comprising the 20 C-terminal residues of EIAV Rev (13). These experiments also demonstrated critical roles for both the
central RRDRW motif and the KRRRK motif within the NLS in RNA-binding (13). Interestingly, however, the ERLE motif was not required for RNA-binding, in agreement with our predictions. Thus, our biochemical RNA-binding site mapping studies for EIAV Rev have provided direct experimental validation of the RNA interface residue predictions of the Naive Bayes classifier.

Of the five clusters of predicted protein binding residues in EIAV Rev, two overlap with known or putative protein interaction domains (the NES and the NLS, respectively), one is located in the non-essential "hypervariable" region (13), one is located near the N-terminus of the protein, and one overlaps within the central RNA binding domain (Figure B.3B). There is no available biochemical data regarding the possibility that the central region of EIAV Rev binds both RNA and protein, but it is interesting that the classifier predicted binding of the NLS region to both protein and RNA. The same residues could directly interact with both the nuclear import machinery and RNA because these interactions occur at different times and in different cellular compartments. Also, by analogy with HIV-1 Rev, it is likely that some of the protein interactions that occur when EIAV Rev multimerizes after binding RNA, involve additional residues located near the RNA binding region that initiates the specific interaction between Rev and the RRE in unspliced EIAV RNA.
Figure B.3 RNA binding site predictions differ for "wildtype" and mutant EIAV Rev sequences. Predicted protein (PRO) and RNA binding sites are indicated along the sequence (SEQ). A. Wildtype, B. & C. Mutant EIAV Rev sequences. RNA binding activity is reduced by >80% in both mutants (see text for details).

Comparison of predicted and biochemically mapped RNA binding sites in EIAV mutant Rev proteins

Site-specific mutagenesis, coupled with functional assays has identified functional domains of EIAV Rev (13, 25, 26). As mentioned above, an NLS/ARM at the C-terminus was identified at the EIAV Rev C-terminus and our cross-linking analyses of the RRDRW and KRRRK motifs indicated that both are likely to contact RNA. To investigate whether our classifiers are capable of detecting mutations that give rise to differences in RNA binding, we performed predictions on several mutant EIAV Rev sequences. As shown in Figure B.3, changes in RNA interface predictions are seen in sequences in which Ala residues are substituted for positively charged residues in the RRDRW and KRRRK motifs (to AADAA and KAAAK). These mutations result in >80% reduction in RNA binding activity (13). The
predicted RNA binding sites no longer overlap these motifs. In contrast, predicted protein interface residues are remain unchanged, consistent with the experimental results.

SUMMARY AND DISCUSSION

Many effective antiviral drugs are directed at blocking the interaction between regulatory proteins and their binding partners or small effector ligands. HIV-1 Rev is one of many clinically important proteins for which there is no high resolution structural information. Identifying critical functional residues in Rev proteins is further complicated by the fact that Rev has no significant sequence similarity to any protein with known structure, and that Rev sequences from different species have very little similarity to one another.

Our comparison of predictions with experimental data on the Rev proteins from HIV-1 and EIAV demonstrates that sequence-based computational methods can identify protein residues that interact with other proteins or nucleic acids. Enhanced prediction accuracy can be achieved if structural information is also available (18, 29). Developing improved methods for predicting binding sites will contribute to our understanding of how proteins recognize their targets in cells and may significantly decrease the time needed to precisely identify binding sites in the laboratory. The level of accuracy obtained using the sequence-based methods presented here suggests that they could expedite the design of experiments to explore the function of key regulatory proteins, even when no structural information is available, with obvious implications for developing new therapies for both genetic and infectious diseases.

ACKNOWLEDGEMENTS

This Research was supported in part by grants NIH, GM 066387 (VH, DD, & RLJ) and CA97936 (SC), by an ISU Center for Integrated Animal Genomics grant (DD, VH &
RLJ), and by USDA Formula Funds (SC & DD). We thank Sijun Liu for technical assistance and Jeffrey Sander for useful comments.

REFERENCES


