Structure-based Computational Retargeting of RNA Binding Proteins

Daniel Bjerre

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2012

Reading Committee:

Gabriele Varani, Chair

Stephen D. Hauschka

Philip H. Bradley

Program Authorized to Offer Degree:

Department of Biochemistry

University of Washington

Abstract

Structure-based Computational Retargeting of RNA Binding Proteins

Daniel Bjerre

Chair of the Supervisory Committee
Professor Gabriele Varani
Department of Biochemistry

Protein-RNA interactions play a central role in post-transcriptional regulation. By interacting with precursor and mature mRNA transcripts, RNA binding proteins (RBP) regulate the expression level and isoform of proteins within the cell in an often spatially and temporally dependent manner. I know of no existing computational method that infers base binding probabilities from structural models. Here I describe the development of a tool to infer the specificity of RNA binding interactions within the Rosetta framework. I use the well-established knowledge-based methods trained on existing x-ray models of RBP in complex or on small molecules as well as a mixture of statistical and physical parameters used in Rosetta prediction of DNA binding proteins.

My computational approach infers local base and residue specificity by performing substitutions on models from x-ray crystallography or NMR. The approach explores limited local structure space through sampling of residue side chains. The structure exploration improves realism of the approach by physically accommodating base and residue substitutions. With a representative set of RBP interactions with single stranded RNA, the scoring functions are able to recover many of the interface side-chain dihedral angles and recapitulate the contacts involved in specific base recognition. I benchmark the scoring functions ability to predict the magnitude and order of base preference.

I explore the application of the specificity prediction tools to design applications selected to illustrate a rational understanding of protein-RNA interactions and with potential therapeutic applications. The scoring function is able to largely recapitulate the results of experimentally investigated mutations of the pumilio-1 domain being investigated as a universal platform for binding arbitrary RNA sequences specifically. I also apply my technique to suggest changes to a RNA recognition motif aimed at re-targeting the domain to specifically bind a target involved in dysregulation in certain cancers.

The results suggest that the Rosetta scoring function may be coupled with small changes in protein sequence and structure to design specificity switches on RBP domains. The computational approach to specificity promises to improve our understanding of sequence specific binding of RNA and aid the development of protein-based approaches to target RNAs involved in disease.

Table of Contents

Page

List of Figures iii

List of Tables v

Glossary vi

Acknowledgements x

Chapter 1. Introduction 1
- A. An Information Theory View of RNA Binding Protein Biology 2
- B. Experimental Approaches to Specificity 17
- C. Computational Specificity Prediction and Design 29
- D. Challenges in RBP Specificity Prediction and Design 43
- E. Thesis Outline 48
- Figures 49

Chapter 2. An Empirical Scoring Function for the Structure-Based Prediction of Specificity in RNA Binding Proteins 51
- A. Introduction 51
- B. Background 54
- C. All-Atom Distance-Dependent Scoring Function: Parameterization and Training 57
- D. Implementation 75
- E. Challenges of Applications to RBP Specificity 78
- F. Statistical RNA Intra-molecular Term 86
- G. Summary 90
- Figures and Tables 92

Chapter 3. Scoring Functions in RNA Binding Protein Specificity Prediction 107
- A. Introduction 107
- B. Background 109
- C. Development of Tools for Specificity Prediction 111
- D. Searching Structure and Sequence Space 119
- E. Comparing Specificity Predictions with Experimental Results 122
- F. Structure Benchmarks 125
- G. Specificity Calculations 135
- H. Design Test with Pumilio1 Protein 150
- I. Summary 155
- Figures and Tables 158

Chapter 4. Application of Specificity Prediction Tools to Protein Design 182
- A. Introduction 182

B. Detailed Comparison of Binding Predictions with Microarray Data 186
C. Retargeting of RNA Recognition Motif to Bind Micro-RNA Precursors 202
D. Summary 226
E. Conclusions 228
Figures 231

Bibliography 243

Appendix 1. Code 268
A. Sequence Logo Code 268
B. Sample Scoring Function in Rosetta 276

Appendix 2. Math 280
A. Lennard-Jones Equation 280
B. Hamming Distance 280

Vita 281

List of Figures

Figure Number Page

Figure 1.1: The interactions between biological macromolecules by molecule type. 49
Figure 1.2: Roles played *in vivo* by RBPs that recognize RNA in a sequence specific manner..... 50
Figure 2.1: Solved structures of protein-nucleic acid complexes deposited in the PDB by year..... 93
Figure 2.2: Number of side-chain nucleobase contacts in training set by base type. 99
Figure 2.3: Number of side-chain nucleobase contacts in training set by amino acid residue and base. 100
Figure 2.4: Density plot of the number of base contacts by residue type in the all-atom training set. 101
Figure 2.5: Score profile for contacts between arginine NH atoms and all the base oxygen acceptors. 102
Figure 2.6: Difference between sequence recovery using scores from complete and fair matrix. 103
Figure 2.7: Discrimination of a correct RNA tertiary structure from near native decoys. 104
Figure 2.8: Test of the empirical scoring function to identify correct RNA tertiary structure..... 106
Figure 3.1: Virtual competition approach to calculating base probabilities at recognized RNA sites. 159
Figure 3.2: Density plots summarize fraction of side-chain dihedral (χ) angles recovered by structure. 162
Figure 3.3: Recovery of correct side chain conformations at protein-RNA interfaces by scoring function. 163
Figure 3.4: Summary of side-chain recovery using Rosetta and the all atom statistical potential..... 164
Figure 3.5: Comparison of base specificity recovery for RBP recognizing ssRNA with all scoring functions. 166
Figure 3.6: Probability of recovering the correct base at each rank or better with each scoring function..... 167
Figure 3.7: Fraction of bases correctly recovered for positions with the strongest predicted preferences. 168
Figure 3.8: Comparison of amino acid recovery at binding interface of RBP recognizing ssRNA. 169
Figure 3.9: Probability of recovering the correct RBP RNA binding residue at each rank or better..... 170
Figure 3.10: Preferred mispredictions of interface residues substitute for correct chemical properties. 171
Figure 3.11: Fraction of residues correctly recovered at positions with the strongest predicted preference.. 172
Figure 3.12: Representative predicted RNA binding motifs for the U1A protein. 174
Figure 3.13: Relative performance of scoring functions in RNA binding motif recovery test. 175
Figure 3.14: Representative predicted binding RNA motifs for the structures of A2BP1 and MBNL1. 176
Figure 3.15: Representative predicted binding RNA motifs for the structures of NOVA2 and PABPC1..... 177
Figure 3.16: Representative predicted binding RNA motifs for the structures of the four subunits of PTB. 178
Figure 3.17: Representative predicted binding RNA motifs for the structures of RBMY1A1 and ZRANB2. 179
Figure 3.18: Structure illustrating Pumilio1 repeat 6 residues used in retargeting experiment. 180

Figure 3.19: The effect of Pumilio1 double mutants on the binding preference at a RNA position. 181
Figure 4.1: Schematic of direct evaluation of predicted and microarray confirmed binding sequences. 231
Figure 4.2: Direct comparison of predicted binding sequences for U1A with microarray intensities. 232
Figure 4.3: Correlation between microarray intensity and mismatches with U1A consensus sequence. 233
Figure 4.4: Summary of microarray intensities for U1A and SLM2 with mismatch number. 234
Figure 4.5: Sorted mismatch intensity plot for SLM2 using the consensus sequence AUAAA. 235
Figure 4.6: Schematic step-wise approach to retarget a RBP to recognize a desired nucleobase. 236
Figure 4.7: Comparison of target sequence with Fox1 binding preference and structure. 237
Figure 4.8: Structure of Fox1 design site B199 with contacting residues labeled. 238
Figure 4.9: Residue preferences at positions contacting for Fox1 position B199 illustrated in logo format. ... 239
Figure 4.10: Change in base specificity with single mutations at five Fox1 residue positions. 240
Figure 4.11: Change in base preference at B199 with a second mutation to a TRP mutated Fox1. 241
Figure 4.12: Specificity calculations with Fox1 RRM double mutants. 242

List of Tables

Table Number Page

Table 0.1: Abbreviations and acronyms used in the dissertation. viii
Table 2.1: Most prevalent RNA binding proteins from Pfam annotation of known sequences. 92
Table 2.2: Structure of origin for chains included in the training set for the all-atom scoring function. 94
Table 2.3: List of protein chains included in the data set with their associated Pfam family. 96
Table 2.4: List of structurally diverse RNAs included in the test of RNA self-scores. 105
Table 3.1: Component weights used for scoring applications with the Rosetta scoring function. 158
Table 3.2: Test set of thirty representative RBP structures binding to single stranded RNA. 160
Table 3.3: Recovery of different types of intermolecular contacts observed in the native structure. 165
Table 3.4: Structures of RBPs in complex with RNA with independently determined specificity data. 173

Glossary

affinity – The relative energy of binding between a protein and partner biological macromolecule.

decoy – A candidate structure built to be a reasonable molecular structure but which likely differs from a physically correct natural structure in subtle ways. Since scoring functions containing empirical terms do not define a complete, differentiable potential, they cannot be used to follow a minimization path. Decoy discrimination tests allow decoupling of the structure building and evaluation components.

information content – The Shannon information (Shannon, 1948; Jaynes, 1957) of the binding preference at a position defined as $H = -\sum_i p_i \log_2 p_i$. The definition is equivalent to the physics definition of entropy because they represent the same concept (Jaynes, 1957).

logo – see **sequence logo**.

position weight matrix (PWM) – A matrix of dimensions $n \times l$ representing the binding probabilities of each of n symbols in the base or residue alphabet for a site of length l (Schneider, Stormo, Haemer, & Gold, 1982; Stormo, 2000). The probability of binding a sequence can be read from the matrix. A PWM may be visually represented as a **sequence logo**.

ribonome - The entire collection of RNA molecules in the cell and organism at any one moment, along with the diverse proteins that associate with them.

sequence logo – a graphical representation of a site-independent PWM where each possible residue or base is represented by a letter (Schneider & Stephens, 1990). All possible residues at a position are stacked with the most probable residue on top. The probability of finding a specific base is indicated by the ratio of the residue letter height to the letter stack height. The height of the stack is either 1.0 or the information content of the position (see **information content**). When information content is used, the scale of the *y*-axis is $\log_2 4$ or $\log_2 20$ for bases and residues, respectively.

specificity – In the context of this dissertation, specificity is the preference of a protein for a binding partner exhibiting a particular base or residue sequence at the interface.

state information – A unique state defined by an linear arrangement of residues or bases or by a given arrangement of particles that represents a particular role or state for a cell. The amount of information is determined by the entropy of the digital message or of arrangement of particles.

transcriptome – The set of all RNA molecules, including mRNA, rRNA, tRNA, and other non-coding RNAs produced in one or a population of cells. The transcriptome refers to the genes transcribed and the rate of their transcription. Quantifying the transcriptome in cells is difficult since techniques such as gene chips quantify the RNA at a later time following changes in the RNA content due to processes such as alternative splicing and degradation.

transition – A switch between nucleobases with aromatic rings of like size: purine to purine or pyrimidine to pyrimidine. For RNA, the transitions are $\mathrm{A} \leftrightarrow \mathrm{G}$ and $\mathrm{C} \leftrightarrow \mathrm{U}$.

transversion – A switch between nucleobase with aromatic rings of different sizes: purine to pyrimidine and *vice versa*. For RNA, the transversions are $\{\mathrm{A},\mathrm{G}\} \leftrightarrow \{\mathrm{C},\mathrm{U}\}$.

Z-score or (**standard score**) **–** Indicates how many standard deviations a datum is above or below the mean. The significance of a prediction may be indicated by how it relates to a distribution of measurements of randomly simulated data or decoys. The Z-score assumes that the non-significant data follows a normal, Gaussian distribution.

Table 0.1: Abbreviations and acronyms used in the dissertation.

abbreviation	**expansion**
AUC	area under curve
CAPRI	critical assessment of predition of interactions
CASP	critical assessment of structure prediction
CDS	protein coding sequences
CDF	cumulative distribution function
CLIP	cross-linking and immunoprecipitation
CSD	Cambridge Stucuture Database
DBD	DNA binding domains
DNA	deoxyribonucleic acid
ED	Euclidian distance
EMSA	electrophoretic mobility shift assay
H-bond	hydrogen bond
HE	homing endonuclease
hnRNP	heteronuclear ribonucleoprotein complex
IC	information content
KLD	Kullback-Leibler divergence
LJ	Lennard-Jones
MD	molecular dynamics
miRNA	microRNA
mRNA	messenger RNA
mRNP	messenger ribonucleoprotein complex
ncRNA	non-coding RNA
NMR	nuclear magnetic resonance
NMR-SIA	NMR scaffold independent analysis

Table 0.1 continued.

PBM	protein binding microarray
PDB	protein data bank
PMF	potential of mean force
PNA	peptide nucleic acids
pre-mRNA	precursor messenger RNA
PWM	position weight matrix
RBD	RNA binding domain
RBP	RNA binding protein
RBPDB	RNA binding protein database
ROC	receiver operator characteristic
RNA	ribonucleic acid
RNP	ribonucleoprotein complex
RRM	RNA recognition domain
SELEX	systematic evolution of ligands by exponential enrichment
SUMOylation	post-translational modification attaching small ubiquitin-like modifier (SUMO)
TF	transcription factors

Acknowledgements

I need to extend a great deal of thanks to my advisor, Dr. Gabriele Varani, for all his advice, mentoring, grant writing and friendship during my tenure at the University of Washington in his lab. I greatly appreciate his willingness to support a difficult project and, especially, for the freedom he gives his students to explore their own ideas.

I also owe a great deal of thanks to Dr. Philip Bradley for providing resources and a way when my project was facing a difficult path forward. The office and computational resources at the Fred Hutchinson Cancer Research Center greatly helped me in pursuing my work with Rosetta and with a great environment in which to learn and think.

I thank Dr. Tim Robertson for pioneered the project on which I worked. His work got me started and introduced me to generic programming in C++. His project also allowed me to pursue computational work in an experimental laboratory with a better work environment than the traditional computational laboratory. I also thank Dr. Suxin Zheng for the computational work he did in the Varani laboratory and for many discussions about our work.

I'd like to thank my colleagues in the RNA binding protein design work Dr. Yu Chen and Jana Mandic. Both worked tirelessly on the experimental aspects of the design project. Their insight and the experimental systems they worked greatly aided my experimental work.

I need to thank Dr. Mike Bardaro, Dr. Ravi Barnwal, Alisha "Jonesy" Jones, Deanna Clem, Dr. Darren Begley, Dr. Brad Lunde and Dr. Matt Shortridge for helping to make working in the Varani lab enjoyable and for keeping perspective on life beyond our immediate projects.

I wish to thank my parents for their support and encouragement through my apparently endless quest for an advanced degree. I'd also like to thank my Seattle friends for helping me enjoy living in this exciting city.

Chapter 1. Introduction

The phenotype of a cell is determined to a significant extent by the interplay between RNA transcripts and RNA binding proteins. RNA binding proteins (RBPs) play a central role in determining the content and localization of mRNAs as well as in determining whether mRNAs are translated into functional proteins. Many RBPs function by recognizing specific RNA sequences and by recruiting other proteins to the recognized sequence (Moore, Schwartzfarb, Silver, & Yu, 2006; Stamm et al., 2005; Xin Wang et al., 2009). Thus, understanding the sequence specificity of RBPs is essential for predicting the biological program of cells. In this project, I used a computational approach to predict the specificity for RNA sequences of RBPs using structure information.

I applied both a machine learning approach and the successful Rosetta molecular modeling tools to predict the sequence specificity of RNA binding proteins from their structures. Both tools were used to predict the interface structure of RBPs binding to candidate sequences and to predict the energy of binding. The machine learning approach used a Bayesian formalism to learn the interaction energies from known structures. The Rosetta approach used both physical and empirical terms that had successfully been used in similar problems. Specificity was expressed as a profile of the binding site that would predict the relative probabilities of binding possible RNA

sequences. I compared the predicted binding motifs to those from previous experimental measurements.

The clearest way to demonstrate the biological relevance of the computational predictions is to use it to perform a novel design application. I developed tools to re-target RBPs to specifically recognize an alternate sequence of single stranded RNA. I demonstrated that I could reproduce the specificity switch for a protein for which the target switch had been designed using selection experiments. I also apply the approach to preliminary design steps for a protein domain that could form the basis for a new approach to targeting micro-RNAs implicated in an aggressive form of cancer.

In this introduction, I discuss the role RBPs play in cells from a biological and information theory perspective. The sparseness of the experimental data for the RNA target specificity of RBPs suggests that a computational approach is necessary. Yet computational approaches have not previously been widely used for RNA specificity prediction and for designing specificity switches. I thus survey previous applications of computational methods for similar molecular recognition problems. Finally, I provide an overview of the organization of this dissertation.

A. An Information Theory View of RNA Binding Protein Biology

Life is fundamentally an information system that learns from and responds to its environment. The cell is the basic unit of the biological program that interacts with its surroundings. What defines life is the interaction of the cell with information from its environment, the representation and processing of that information internally and the propagation of captured information through the reproduction of fit organisms (Loewenstein, 1999). The information theory view of biology can be extended beyond

the cell to inform our understanding of organisms, evolution and even ecosystems, but the cell is the basic unit of life and the information cycle that it exploits (Avery, 2003, pp. 95–108). The information that expresses the state of the cell consists of the digital information stored in the DNA as well as all of the state information stored in expressed RNAs and proteins (van Driel, Fransz, & Verschure, 2003).

Molecules that have been considered as simple intermediaries or which serve a functional role such as RNA and proteins are a critical part of the information state of the cell. A cell's program can largely be replicated by copying its DNA. However, the state of the cell is determined by other factors including chromatin structure, RNA content, and protein modifications.

Relevant to my thesis work, RNA binding proteins play a critical role in defining the state of the cell responding to internal changes in the cell and processing sensor information from outside the cell. RNA and the proteins that bind it are central to the flow and processing of information in the cell (Sharp, 2009). We can describe the types of molecules involved in defining the cell state and in processing information (Figure 1.1). The overview of interactions between molecule types in the cell could be seen as analogous to a map of roads and highways in a city. A city map shows the conduits between locations. From the map, we can imagine how objects and information could be physically transported in the cell. However, the map does not show the movement of objects or information through the city. Figure 1.1 maps the connectivity between classes of macromolecules in the cell to show that proteins can interact with other proteins, with RNA and with DNA. However, the meanings of these interactions can only

be understood if we also understand the function of the proteins and the details of their interaction with other molecules.

The recognition of specific sequences of DNA and RNA by transcription factors and RBPs, respectively, bridges two forms of information: digital information in nucleic acid sequences and physical state information. Cell state information and information about the cell's environment may be represented by the physical arrangements of molecules and in post-transcriptional modifications of proteins (Goldberg, Allis, & Bernstein, 2007). Proteins that recognize specific nucleic acid sequences can regulate the transcription of DNA (Pabo & Sauer, 1992) or the translation of RNA (Keene, 2007). In effect, proteins that recognize specific sequences form decision nodes that control the use of digital information in the cell. Thus, a cell can perform metabolic activities and interact with the external world by calling new functional proteins and regulatory proteins and RNA from its digital storage (Junhyong Kim & Eberwine, 2010).

RBPs regulate the physical and temporal expression of mRNA by forming ribonucleoprotein (RNP) complexes at key locations and times in the mRNA metabolic cycle. Specific recognition of RNA sequences is important in regulating splicing, transport, localization, stability and translation of RNA transcripts (Figure 1.2). Sets of RNAs for genes involved in a cell function are similarly regulated at each of the steps of RNA metabolism (Keene & Tenenbaum, 2002). This suggests that RNAs coding for proteins involved in similar functions contain a common signal in the structure or sequence of RNA (Hieronymus & Silver, 2003). Many of the RBPs involved in each step contain domains implicated in sequence specific recognition. Protein-binding regions of RNA are found mostly in an unpaired state (A. Gupta & Gribskov, 2011). The structure

of the RNP binding interface supports an important role for sequence recognition. Predicting the specificity of RBPs could allow a better understanding of these information rich interactions.

Furthermore, the ability to alter RBP specificity could be exploited for pharmaceutical applications. RBPs are some of the most tightly regulated proteins (Schwanhausser et al., 2011). mRNA encoding RBPs and transcripts processed to micro-RNA are potentially good drug targets (Costa, 2009). Peptides and protein domains are gaining acceptance as molecules that can be adapted as drugs (Sato, Viswanathan, Kent, & Wood, 2006; Mason, 2010). RBPs that may be modified to recognize regulatory RNAs could be good starting points for the development of biologic drugs.

The understanding gained from predicting sequence specific interactions of RBPs and the practical advancement achieved by designing RBPs to target coding and non-coding RNAs (ncRNAs) can be understood in the context of information theory. In the rest of this section, I elaborate on the roles of RNA and RBPs in defining the information state of a cell and in increasing biological complexity. I also discuss non-coding RNA (ncRNA) as pharmaceutical targets and RBPs as potential starting points for developing new regulators of non-coding RNAs.

1. Information theory and biology

Thinking about biological information helps us identify holes in our knowledge of biological mechanisms. Remarkable strides have been made in understanding the role and flow of information in cells during the half-century since the first description of the structure of DNA (Watson & Crick, 1953). The central dogma describes how the code of life results in the protein sequences that form the enzymatic and structural

foundations of life (Crick, 1970). The discovery of alternative splicing forms the foundation of an expanded genetic repertoire in eukaryotes (Berget, Moore, & Sharp, 1977; Gilbert, 1978). Exploring this forward flow of information from genetic code to protein product and elucidating the function of the protein product represents much of the work in experimental and computational molecular biology and biochemistry. Yet the forward flow of information fails to explain how the cell uses its library of structural information to respond to and interact with its environment.

Many of the key features of cells, especially those of cells in multicellular organisms, cannot be explained directly in terms of the digital information that describes the components of life or by the functions of those components. A cell must understand and respond to its environment by drawing from its library of structures in response to needs and stimuli on a variety of time scales. The standard view is insufficient for understanding important aspects of multicellular life such as cell differentiation, the establishment of connections between neurons, and the establishment, maintenance and transfer of epigenetic markers. We may understand what has been missed in terms of 'cell state' and 'information processing' by applying information theory to what is currently understood.

In looking at the cell as an information system, one's attention may be drawn to the digital information that is encoded in the DNA. The central dogma of biology is the model for how living organisms maintain and propagate the information required to build the machinery that perform the chemical and structural functions of life (Crick, 1970). The primary flow of information in cells and in organisms is from a digital storage mechanism (DNA storing and replicating quaternary information) to active,

structural molecules (proteins and ribozymes). The storage of digital information was first quantified in terms of the bits necessary to transfer a message (Shannon, 1948). The quantification of information takes the same form as the equations of entropy. The equations of statistical mechanics can be re-derived from an information theory perspective (Jaynes, 1957). This perspective helps us not only in understanding how and why life exploits a digital message in the DNA but also helps us understand how arrangements, partitions of ions across membranes and even macromolecular structure are information rich.

The information stored in DNA contains the blueprints for protein structures that perform most of molecular transformations and for structural components of cells. However, the information on the DNA is about as useful as data structures and algorithms stored on a computer hard drive (Junhyong Kim & Eberwine, 2010). Even including the entire eukaryotic transcription and translation apparatus would result in a cell of limited utility (Bonasio, Tu, & Reinberg, 2010). Without feedback mechanisms, a cell would serve only to alter the equilibrium conditions and a system that achieves total equilibrium is dead.

The forward flow of genetic information to the active protein molecules consists of transcription, RNA processing, and translation (Figure 1.1, horizontal blue arrows). A cell with only those functions would be similar to a machine that could build its own peripherals such as display and printers (Junhyong Kim & Eberwine, 2010). Having just the digital information that encodes only the physical structures of life results in something that does not resemble life or even have any capabilities of a basic computer.

Much of the information in a cell defines the components that implement the algorithms of life and that store state information. The components that store state information and implement state transitions functions are specialized and function on specific timescales. At the shortest timescales, information transmitted from the environment might be rapidly received through modifications to an ion gradient by channel proteins or may directly initiate signal transduction through complex formation and protein modification. On a longer time scale, a needed enzyme or structural protein may be called up by activating transcription of a gene. RNA may provide a place for responses on an intermediate time scale and may be the locus for processing more complicated signals (Junhyong Kim & Eberwine, 2010).

While most of the state information and algorithms are stored and implemented in arrangements of proteins, RNAs, and DNA packing and modification, the components that implement these functions must be encoded in the DNA. Since the complete genomes of many organisms including humans have been sequenced (International Human Genome Sequencing Consortium, 2004), scientists have tried to explain the relative complexity of organisms in terms of genes and their function. Interestingly the number of bases in the protein coding sequences (CDS) does not differ by a large factor between organisms as different in complexity as *D. discoideum*, *C. elegans,* and *H. sapiens* (Taft, Pheasant, & Mattick, 2007). All these organisms have between 21 and 32 megabases of CDS. However, the percentage of the genome corresponding to CDS drops from 62% for *D. discoideum* to around 1% for *H. sapiens* (Taft et al., 2007). This suggests that most of the complexity is attributable to the non-coding regions of DNA (Costa,

2008). Many if not most of the additional genome sequences must account for regulatory signals that increase complexity.

2. Cellular macromolecules define cell state

The additional non-coding regions of DNA likely encode regulatory signals and RNAs that help define the cell state. Regulatory signals involving protein modifications and proteins that bind and epigenetically mark DNA sequences are better understood than those involving the sequence and structure properties of those involving RNA sequence recognition sites and ncRNAs. The importance of regulatory proteins in controlling gene expression (Jacob & Monod, 1961) and cell differentiation (Rudel & Sommer, 2003) was recognized early in molecular biology. Investigation of regulatory mechanisms focused on proteins that regulate transcription, such as transcription factors and chromatin remodeling factors (van Driel et al., 2003). Protein response cascades and DNA epigenetic markers have been extensively studied as well (Williamson & Whetton, 2011; Esteller, 2007).

Proteins form signaling cascades in response to environmental signals. The activity of proteins is modulated through interactions with other proteins and through protein modifications. Modifications that affect protein function include phosphorylation, methylation, ubiquitination and SUMOylation (Williamson & Whetton, 2011). The phosphorylation of rhodopsin in response to light was one of the earliest and best understood examples of protein modification as part of a signaling pathway (Wilden & Kuehn, 1982; Maeda, Imanishi, & Palczewski, 2003). Proteins that modify other proteins in response to environmental factors can be seen to be participating in simple state machines (Fisher & Henzinger, 2007). These post-translational

modifications only partially store the state information of the cell. The effects of protein modifications may feed back and affect transcription and translation.

State information is stored on DNA in modifications to the DNA molecule such as methylation (Esteller, 2007; Wade, 2001) and in a histone code (Scharf & Imhof, 2011). Additional state information in the form of transcription factors involved in operons allows genes to be activated under specific cellular conditions. Relatively few operons have been described in eukaryotes and the time from activation to change in protein concentration is relatively long (Osbourn & Field, 2009). Changes to the expression of genes encoding transcription factors have been shown to be critical to cell differentiation (Boyer et al., 2005). Myelin formation involves coordination of gene expression by a complex network of transcription factors (Fulton, Denarier, Friedman, Wasserman, & Peterson, 2011). However, transcription factors and epigenetic information associated with DNA only partially explains changes in cell protein composition.

The protein-based response mechanisms and response and epigenetic mediated retrieval of genes from the DNA only partially accounts for the range of response and information processing required for cellular function. Many of the mechanisms establishing and maintaining cell differentiation can probably be largely explained in terms of transcription factors and epigenetic markers associated with DNA (Jonghwan Kim, Chu, Shen, Wang, & Orkin, 2008). However, these transcriptional regulators do not explain the organismal complexity that has arisen without significant expansion of the genome (Taft et al., 2007). They also do not provide an obvious mechanism for temporally and spatially localized gene expression whose necessity is most obvious in

highly specialized cells and temporally evolving cells such as neurons as well as in early embryotic development.

The missing information flow in cells involves protein and RNA feedback loops. The feedback mechanisms involving proteins and ncRNA molecules determine the information state of the cell (Keene, 2001). The concentration of cytoplasmic protein is significantly controlled at the point of translation (Schwanhausser et al., 2011; Waldman, Tuller, Shlomi, Sharan, & Ruppin, 2010; Zeisel et al., 2011). But the gene transcripts and ncRNAs constituting the cellular transcriptome play a much bigger role, as evidenced by experiments that demonstrate that RNA is the primary locus of cell state information (Junhyong Kim & Eberwine, 2010) and can even be used to pass on the state information (Nowacki, Shetty, & Landweber, 2011). Additionally, the RNA binding proteins provide a critical bridge between information in digital and structural form.

3. RNA, RNA binding proteins and cellular information

The RNA in cells may function like state memory for cellular information processing. Much attention has been paid to DNA as the basic information molecule in the cell. Indeed the DNA does contain all the information about the sequence of the ribosome, tRNAs, and proteins that are the active agents in the cell. However, the state information about the cell and what it knows about its environment is only partially stored in interactions between proteins and between proteins and DNA. A great deal of the state information is stored in RNA and in protein interacting with RNA (Junhyong Kim & Eberwine, 2010).

A clever experiment demonstrated that cellular information largely defines the state of the cell. Sul et al. showed that transferring RNA isolated from the cytoplasm consistently transferred much of phenotype of differentiated astrocytes. This suggests that the RNA largely defines the state of the cell (Junhyong Kim & Eberwine, 2010). That the RNA contains the information for causing a predictable change in cell phenotype does not imply that RNA is the molecule that executes the change. Many cytoplasmic mRNAs encode RBPs (Glisovic, Bachorik, Yong, & Dreyfuss, 2008; Hinman & Lou, 2008). The RNA binding proteins encoded by some mRNA binds to regulatory sequence in related mRNA transcripts of related genes to regulate protein expression levels (Hogan, Riordan, Gerber, Herschlag, & Brown, 2008). The co-regulation of related genes is characteristic of a regulatory network (Keene, 2007).

A key indicator of the centrality of RBPs is their tendencies to regulate the stability and translation of their own transcripts (Keene & Tenenbaum, 2002; Boutz et al., 2007; Kishore, Luber, & Zavolan, 2010). The regulation of its own expression level indicates that the RBP is a node in the post-transcriptional regulatory network (Janga & Mittal, 2011). This ability to tightly control its own expression allows post-transcriptional control to be a dominant regulator of cellular protein composition.

Measurements of mRNA and protein half-lives allow insight into whether the expression levels of cellular protein are mostly determined by transcriptional or translational control mechanisms. Using pulse-labeling techniques coupled with sequencing or mass spectroscopy, the half-lives of transcripts and the proteins they encode could be determined (Mittal, Roy, Babu, & Janga, 2009; Schwanhausser et al., 2011). Proteins with related functions were often encoded by mRNA transcripts with

similar half-lives (Mittal et al., 2009). Genes encoding RBPs were found to have stable mRNAs but unstable proteins (Schwanhausser et al., 2011). A stable mRNA that generates a short-lived protein is consistent with a translational regulator function. Similarly, Schwanhausser et al. (2011) found that cytoplasmic protein concentrations were best explained by rates of translation.

The sequence-specific binding of RNA by RBPs provides the link between a response in terms of physical and structural elements and the sequence space of the cell. The binding of RBPs to the transcripts of genes with related functions suggests that these RNAs expose similar sequences in their regulatory binding regions (Hogan et al., 2008). The presence of regulatory binding sequences is likely a feature of all aspects of post-transcriptional RNA processing that are regulated by the formation of RNP complexes. A whole-genome microarray was used to reveal tissue dependent alternative splicing patterns which may be regulated by the recognition of specific *cis* regulatory sequences by differentially expressed RBPs (Castle et al., 2008). The inclusion and exclusion of exons can be tied to combinatorial binding of several RBPs (Xin Wang et al., 2009). Many of the interactions regulating mRNA processing and transport involve one or more sequence specific recognition events. The mechanism of RNA regulation will reveal how RBPs integrate digital and structural information that determines the cell state.

4. RNA binding proteins in post-transcriptional regulation

RNA binding proteins are now recognized for their role in increasing protein diversity through alternative splicing and regulating the temporal and spatial expression of proteins through various mechanisms. Specifically, RBPs involved in

alternative splicing, RNA transport and localization, RNA stability and translational control employ specific recognition of RNA in all cell types (Mittal, Scherrer, Gerber, & Janga, 2011). RNA binding proteins containing representative examples of most of the domains known to bind single stranded RNA can be traced back to the last common ancestor of all animals (Kerner, Degnan, Marchand, Degnan, & Vervoort, 2011). Looking at their roles in specific cell types and the diseases associated with them elucidates the contributions of RBP binding to essential cell functions.

The importance of post-transcriptional regulation at the RNA level is clearly demonstrated in neurons. Dysregulation of RBPs with regulatory roles is a common characteristic of many cancers and neuronal diseases (Lukong, Chang, Khandjian, & Richard, 2008). The causes of these diseases highlight the many critical functions that RBPs have in post-transcriptional regulation. These functions include specific editing and methylation of the mRNA (Mehler & Mattick, 2007), diversification of proteins libraries through alternative splicing (Q. Li, Lee, & Black, 2007) and regulation of mRNA translation (Mittal et al., 2011). The RBPs that perform these functions that diversify and control mRNA often recognize RNA based on sequence.

In neurons, sequence specific binding of RBP is involved in alternative splicing, mRNA localization and translational control. The importance of alternative splicing in neurons is demonstrated by the variety in RBPs dedicated to this function. Hu proteins are involved in alternative splicing in neurons that is important for establishing neuronal junctions (Hinman & Lou, 2008). Serine-arginine rich proteins (SR proteins) recognize specific RNA sequences and are critical components of the splicing machinery in neurons (Graveley, 2000). Nova proteins regulate alternative splicing sites in motor

neurons by directly binding to pre-mRNAs (Ule et al., 2006). Mouse Nova knockout models exhibit motor neuron degeneration similar to a human neurodegenerative disease (Ule et al., 2003). Genome-wide screens have linked the binding of RBPs with exon splicing patterns (Licatalosi et al., 2008). Alternative splicing is one mechanism neurons use for establishing a particular phenotypic state.

Additionally, alternative splicing is exploited to establish combinatorial unique states. Alternative splicing allows for a dramatic expansion of the proteome with a smaller expansion of the genome (Nilsen & Graveley, 2010). The mechanism may be particularly important in neurons where highly alternatively spliced genes such as Dscam allow for much of the evolved complexity in brain tissues (B. E. Chen et al., 2006). Specific isoforms of many genes are specifically expressed in many tissue types using the alternative splicing mechanism (E. T. Wang et al., 2008). Alternative splicing represents a post-transcriptional expansion of genetic information regulated by feedback in the form of specifically binding RBPs.

The phenotype of cells such as neurons is further regulated by RBPs at the point of translation. The extended shape of neuronal cells requires tight translational control with specific localization and timing (Sossin & DesGroseillers, 2006). Loss of translational control in synapses is found in many neurodegenerative diseases (Liu-Yesucevitz et al., 2011). The flexibility and control over protein expression afforded by RBPs that regulate pre-mRNA processing and mRNA translation is most dramatic in neurons, but the same processes are useful to cells in all tissues.

The transcriptome includes many non-coding RNAs that do not function as transcriptional intermediates in protein synthesis. Many of these RNAs have functions

in regulating transcription, RNA processing and translation. Many of the RNA binding domains that bind single stranded RNA and are involved in mRNA processing, transport, and transcriptional regulation also bind these other regulatory ncRNAs (Fabian, Sonenberg, & Filipowicz, 2010). Understanding sequence specific binding of RNA binding domains would help in understanding RNPs involving ncRNAs. Additionally, RNA binding domains could be used to design protein-based biologic pharmaceuticals that target these ncRNAs.

5. Protein interactions with non-coding RNAs

Non-coding RNAs include a variety of small RNAs that effect transcription, RNA processing and localization (Prasanth & Spector, 2007). Specifically, microRNAs (miRNAs) have regulatory roles in post-transcriptional gene regulation (Bushati & Cohen, 2007) that can be traced back to the dawn of multicellular life and provide robustness to transcriptional programs (Berezikov, 2011). RBPs involved in forming RNP complexes with regulatory ncRNAs employ the domains known to bind single stranded RNA in a sequence specific manner (Fabian et al., 2010). The roles of some of the RBPs that bind miRNA were discovered by studying their expression in diseased cells (van Kouwenhove, Kedde, & Agami, 2011). Being able to predict the sequence specificity of these proteins would help elucidate their roles in the processing of small regulatory RNAs such as miRNAs. However, miRNAs that bind mRNA transcripts play important roles in neurodegenerative diseases and cancer.

Correlation between the expression of miRNAs and disease states has been found using new high throughput sequencing techniques. One of the primary roles of miRNAs is to regulate mRNA function by directly binding the transcript usually at

untranslated regions (Fabian et al., 2010). miRNA binding may alter the rate of mRNA decay or inhibit protein translation. miRNAs have emerged as potential drug targets for treating a variety of human diseases (Mack, 2007). Cancer is a disease of genetic dysregulation. Specific miRNAs are often found to be over-expressed in cancer (Croce, 2009; Melo & Esteller, 2011). Molecules that specifically inhibit the expression of oncogenic miRNAs are attractive drug candidates.

Existing RBPs may provide good starting points for interfering with miRNA regulatory activity by specifically inhibiting its biogenesis or through competitive binding. Several strategies have been explored for inhibiting miRNA formation or function (Reichel, Li, & Millar, 2011). Some approaches to altering miRNA expression levels have made it to clinical trial (Wahid, Shehzad, Khan, & Kim, 2010). The most straightforward approach to inhibiting miRNAs is through the use of complementary anti-miR sequences using nucleic acid analogues that resist degradation (Elmen et al., 2008). Nucleic acid binding protein fragments and peptides that target specific nucleic acid would theoretically perform a similar function with improved deliverability and less toxicity (Cooper & Waters, 2005). Regulating miRNAs associated with disease is a powerful way to alter the information state of cells.

B． Experimental Approaches to Specificity

A lesson of genome sequencing projects and of structural genomics projects is that experimental methods may be scaled up to widen our knowledge of important biological information. There is increasing interest with the aspect of cellular information contained in ‘interactomes’ (Vidal, Cusick, & Barabási, 2011). Systems biologists infer the interactions in cellular information networks from associations in

large-scale experiments. Protein-RNA binding data should be an essential part of understanding the cellular information network. However, experimental approaches that have been applied to understanding RBP binding face significant challenges and are not likely to scale up sufficiently.

Experimental approaches to specificity have generally either tried to (1) characterize the physical interactions of RBPs or (2) to directly characterize specificity without structural details. The structural characterizations inform biochemical intuition about binding interactions, but insight has not coalesced into rules with predictive applications. The sequence specificity information from biochemical assays have illuminated some aspects of how RBPs help in executing post-transcriptional regulatory programs, but the current data are insufficient for the discovery of novel co-regulatory programs based on sequence specific recognition.

Computational methods could be useful tool in filling-in our knowledge of the structure and specificity of uncharacterized RBPs. The emerging data provide an opportunity for constructing test sets that can be used to objectively assess the computational approaches.

1. Structural characterization

Structure provides a rich source of information for attributing binding or enzymatic activity to a chemical or energetic mechanism. RBP structures in complex with their RNA target have been carefully dissected. Investigators have generated links between residue types and binding pockets and specific residue interactions that predict the RNA-binding sequence. However, the predictive value of structural observations has been much more limited.

a. characterization by general properties

Several studies have examined the types of chemical interactions characteristic of RNA-protein interfaces (Treger & Westhof, 2001; Jones, Daley, Luscombe, Berman, & Thornton, 2001). Unsurprisingly, the binding surfaces are mostly characterized by a preponderance of positively charged amino acids such as LYS and ARG that bind the electronegative phosphate backbone with high affinity. The aromatic residues PHE and TYR are more enriched at the binding interface of proteins binding to RNA than those binding to double-stranded DNA (Jones et al., 2001). The significant difference between RNA and DNA structure is most evidenced by the dramatically higher percentage of contacts with the sugar and a slight increase in the percentage of direct contacts with the bases.

Contacts are more common with the base edge for RBPs than for proteins binding DNA (Lejeune, Delsaux, Charloteaux, Thomas, & Brasseur, 2005). The amino acids most likely to interact with a base edge are ASN, HIS, ASP, GLN and TYR. The propensities of these amino acids to engage in interactions such as hydrogen bonds with groups on the base edges leads to their interaction with specific bases (Lejeune et al., 2005). While an analysis of the contacts can help form a heuristic understanding of the influence of residue on specificity, the structural diversity of the bound RNA precludes predictions based on amino acid composition alone.

The RBP recognizes functional categories of RNA that present distinct binding surfaces. The analysis by Bahadur, Zacharias and Janin (2008) breaks down the types of bonds and calculable properties of protein-RNA interactions by RNA structure properties. RBP interfaces with single stranded RNA regions have a greater per residue

buried surface area and form less hydrogen bonds with proteins than interfaces with double stranded regions. There are not enough examples of each type of bound RNA structure to analyze differences in residue composition for the RNA in the assigned structural classes. Statistics of base edge contacts, regardless of the categorization of the interaction mode, provides evidence for the involvement of chemical interactions determining the probability of binding a RNA with a specific nucleobase.

b. examples of making binding predictions from structure

Structure superposition plots provide insight into specific recognition by revealing the chemical properties of binding pockets. The amino acid nucleotide interaction database superimposed all examples of direct interactions between bases and side-chains observed in a sample of protein-RNA and protein-DNA complexes (Hoffman et al., 2004). Interestingly, an accounting of the numbers of interactions between the standard amino acids and canonical bases revealed more contacts with all bases by polar and positively charged residues. However, this analysis did not reveal significant differences between the numbers of contacts with each the four residues by each amino acid type.

In another study, the base environment in the protein-binding pocket was visualized in terms of chemical groups (Morozova, Allers, Myers, & Shamoo, 2006). After superimposing the bases, the positions of acceptor and donor nitrogen and oxygen atoms were plotted. They showed the locations of atoms that could participate in important contacts such as hydrogen bonds with each base. The analysis showed that the protein binding sites were optimized to provide the preferred hydrogen bond and van der Waals interactions with base edges (Morozova et al., 2006). However, the

amino acid composition of the binding pocket does not uniquely determine the preferred nucleobase.

RBP binding affinity is dependent on the geometry and the strength of the interaction between functional groups on protein side-chains and on the base edge. Purine binding in a RNA recognition motif (RRM) was studied in human heterogeneous ribonucleoprotein A1 (hnRNP A1) (Myers & Shamoo, 2004). Because the part of hnRNP A1 referred to as UP1 also binds DNA, Myers and Shamoo (2004) were able to build DNA targets replacing a single adenine with a variety of purines displaying different donor and acceptor functional groups. Correct hydrogen bond contacts proved most important to binding affinity. A LYS residue that participates in a stacking interaction also appeared to confer specificity for the native base (Myers & Shamoo, 2004). The study highlights how the contacts between side-chains and the base edge are central to specific recognition of a base. However, we cannot distill adenine recognition in one particular protein to a set of predictive rules for binding.

c. lessons from structural characterization

Groups performing analyses of structural aspects of interactions have speculated on which interactions may confer specificity. With the increasing number of available structures and the application of computational tools to analyze the interface contacts, the types of interactions that confer specificity and the strength of the contributions of the interface features is now partially understood. The degeneracy of functional groups across amino acids and bases makes binding specificity highly dependent on structure (Perutz, 1983). Protein sequence mostly determines structure (Anfinsen, 1973), and structure mostly determines binding specificity (Pabo & Sauer, 1984). The relationships

between sequence and structure and between structure and specific binding are none-the-less extremely complex (Redfern, Dessailly, & Orengo, 2008).

The structural origins of specificity explain why a first-order residue code for RBP binding specificity is unlikely. Many of the reasons for the difficulty in discovering a code for DNA recognition by proteins apply to the RNA case. Analysis from the frame of the nucelotide reveals common binding patterns. Small differences in protein sequence and structure have a large effect on local geometry and structural changes that occur during binding (Pabo & Nekludova, 2000; Fuxreiter, Simon, & Bondos, 2011).

The complexity of local geometry also applies to RNA and is complicated by the variety of RNA structure (Auweter, Oberstrass, & Allain, 2006). However, the extensive use of the RRM and K-homology (KH) domains where structure of the domain is conserved and protein contacts are mostly with the RNA bases, strongly suggests a discoverable code may exist. The structural studies demonstrate that a sequence code for RNA recognition is non-trivial, but extensive measurements of sequence specificity for closely related proteins may still lead to a partial code.

The type of code that may be expected for RNA binding proteins depends on the consistency of the binding surface. For cases where the RNA base is consistently bound at a specific position using aromatic stacking, there could be a relationship between amino acid side-chains at positions consistently contacting base edges and the recognized base. For more subtle cases the specificity of a binding pocket relies on the specific chemical properties of the pocket (Morozova et al., 2006). In the latter case the binding preference cannot be decoupled from knowledge of structure.

2. Experimental measurements of specificity

Several experimental approaches have been explored for measuring the preference of RBPs for binding specific RNA sequences. Most approaches that have been applied to measuring the specificity of DNA binding proteins have been used for measuring RNA binding specificity. The methods that have been adapted to RNA include a microarray method for exploring the affinity of a binding protein for the entire sequence space. However, the lower stability of RNA and its varied tertiary structure have precluded scaling up these specificity measurements. Thus, the small number of known binding motifs is insufficient to search for a recognition code based on protein sequence without the use of information from solved structures and computational techniques.

a. specificity measurements of DNA binding domains

Knowledge of the binding specificity for DNA binding domains (DBDs) is increasing rapidly. Older techniques provided a fragmented view of sequence specificity and binding constants, but high-throughput methods have recently yielded a large amount of information about the sequence motifs recognized by DBDs.

Measurements using the basic techniques developed for measuring sequence specific binding still constitute much of the data for DBDs and RBPs. The most accurate binding affinities were assayed using the electrophoretic mobility shift assay (EMSA) (Garner & Revzin, 1981). EMSA has the advantage of allowing measurement of binding constants under diverse conditions (Lane, Prentki, & Chandler, 1992). However, EMSA measures affinity for only a single sequence. Methods such as the systematic enrichment of ligands by exponential enhancement (SELEX) coupled an *in vitro* pull

down assay with improving sequencing technology to infer sequence specificity (Tuerk & Gold, 1990; Djordjevic, 2007). These techniques yield affinity and specificity data for single domains and proteins.

With DBDs, a considerable effort has been made to discover how sequence specific interactions participate in the broader regulatory network. The JASPAR database of transcription factor (TF) regulatory sequences contains more than 457 binding motifs from different sources (Portales-Casamar et al., 2010). The TRANSFAC database attempts to integrate information about known transcription binding sites, their specificity and interactions with other regulatory functions from multiple sources (Wingender, Dietze, Karas, & Knüppel, 1996). The TRANSFAC database couples knowledge about sequence and gene specific binding with information about regulatory networks (Wingender, 2008). The TRANSFAC approach is the essence of a systems biology approach where data from heterogeneous sources are integrated into a single database.

More recently, a variety of high-throughput measurements of binding specificity of DBDs have been developed. Methods that can be scaled up include *in vivo* assays such as chromatin immunoprecipitation with microarray (ChIP-chip) or with sequencing (ChIP-seq) or *in vitro* methods such as protein binding microarrays (PBM) or protein microarrays (Xie, Hu, Qian, Blackshaw, & Zhu, 2011). The universal protein binding microarray (PBM) method is one of the most scalable methods for determining DBD binding motives (Philippakis, Qureshi, Berger, & Bulyk, 2008). The UniPROBE database of PBM data reports more than 393 binding motifs (Newburger & Bulyk, 2009). The stability of DNA and the greater amount of effort directed at understanding

trascriptional regulation has led to an explosion of binding motif data in the last four years.

The experimental binding profiles provides a basis for understanding higher-order binding interactions and provides a basis for the development of computational methods. Inference of the affinity and specificity of sequence specific binding of DBDs from experiment allows the prediction of binding sites and for a mathematical understanding of expression patterns that result from the binding of multiple TFs (Wasserman & Sandelin, 2004). The experimental data have made it possible to objectively test the structure-based approaches to computationally predicting binding preferences described in section C.2.c below. While most of the methods used for DNA have been applied to quantifying RBP specificity, substantially less data have been obtained for those domains.

b. specificity measurements of RNA binding proteins

As described in section A.4, regulation at the level of the RNA transcript substantially determines the level of protein expression. Applications of the RNA equivalents of the specificity measurements developed for DNA have yielded only a sparse set of binding profiles. Additionally, efforts to store RBP binding motifs in databases and to integrate this knowledge with other aspects of post-transcriptional control lag far behind those efforts with DBDs. The sparse data set and the lack of organization are large obstacles to understanding post-transcriptional regulation and to the development of computational methods for predicting RBP binding.

methods – Some methods for measuring RBP binding affinity and specificity could be directly ported to the RNA case while others needed substantial modification.

Experiments have mostly been aimed at discovering the parts lists of functionally important RNPs and at solving the structures of the component RBPs (Godin & Varani, 2007). The importance of sequence specific recognition for the modular coordination of RNA transcripts was not recognized until recently (Mansfield & Keene, 2009).

Experiments aimed at understanding single proteins or domains constitute much of the current literature on RBPs. Most of what is currently known about RBP affinity and specificity still comes from RNA adaptations of EMSA and SELEX experiments (Garner & Revzin, 1981; Tuerk & Gold, 1990; Djordjevic, 2007). NMR site independent analysis (NMR-SIA) was recently developed to comprehensively assay the specificity of a single RNA recognition motif (García-Mayoral, Díaz-Moreno, Hollingworth, & Ramos, 2008). However, these experiments are not scalable and the results of these experiments escaped being centrally catalogued until recently

High-throughput techniques have recently been applied to RNA binding proteins but the amount of data they have generated is still limited. The protein binding microarray (PBM) cannot be used to directly assay RNA binding proteins. While some RBPs have been known to bind single stranded DNA, most are specific for RNA. The adaptation of PBMs to the RNA case, RNAcompete, used multistep process including a normal microarray to assay RBPs (Ray et al., 2009). The approach cleverly transcribed a cDNA microarray to RNA, allowed labeled RBPs to bind the RNA, and then labeled and re-annealed the RNA to the microarray. The RNAcompete approach found the binding profile of nine proteins binding to unstructured single-stranded RNA of up to seven nucleotides. The approach has not yet been applied on a larger scale.

The *in vivo* binding of RBPs can be assayed using cross-linking and immuno-precipitation (CLIP). CLIP can be used to identify high affinity binding sites recognized by RBPs (Hafner et al., 2010; Kishore et al., 2011). The results can be used to predict binding motifs in many cases, although increased C to U base transversion does complicate the analysis (Hafner et al., 2010). CLIP results were shown to agree with binding preferences from the RNAcompete approach (Kishore et al., 2011). CLIP confirms that binding preferences are physiologically relevant. However, CLIP has not been used to determine a large set of binding motifs.

databases – Only recently has the sequence specificity of RBPs been understood to be relevant beyond the specific RNPs in which they have been studied. Groups studying RBP's have discovered many consensus target sequences for these proteins over the years. SELEX has also yielded the information needed to infer preferred binding motifs. The SELEX results were often reported as figures in articles making extracting binding motif information tedious.

Recently, a few databases have sought to collect consensus sequence and binding motif information. The SpliceAid database collected reported consensus RNA binding sequences from literature (Piva, Giulietti, Nocchi, & Principato, 2009). The database allows the search of user-provided sequences for possible RBP targets. The database tables may be downloaded and used as a reference for the papers reporting RBP binding information. More interestingly, the RNA Binding Protein Database (RBPDB) provides the first searchable reference to reported binding preferences by gene (Cook, Kazan, Zuberi, Morris, & Hughes, 2010). The database reports the binding preference in

terms of a position weight matrix (PWM) when it is available. The RBPDB dataset includes PWMs built from older SELEX data in a format useful for derivative work.

Only a few databases make use RNA binding motifs to help find and visualize biologically important interactions of RBPs. ESEFinder uses a small set of known PWMs to scan genes for exonic splicing enhancers (ESEs) (Cartegni, Wang, Zhu, Zhang, & Krainer, 2003). ESEs help regulate the inclusion or exclusion of exons in gene transcripts where alternative splicing is tissue or time dependent. The non-coding RNAs and protein related biomacromolecules interaction database (NPInter) catalogues biologically relevant interactions with non-coding RNAs (Wu et al., 2006). NPInter does not inform our knowledge about sequence specific interactions between proteins and RNAs, but does represent a way to use protein-RNA interactions to understand the state of the cell.

3. Limitations of an Experimental Approach

The experimental determination of binding profiles for proteins binding nucleic acids in a sequence specific manner is essential for understanding the feedback mechanisms the regulate cell phenotype. High-throughput methods have begun to yield binding preferences for hundreds of DBDs. Even with the increasing dataset for DBD binding, there is still a need for computational techniques that can help fill in the regulatory networks. Computational techniques are needed for the discovery of binding motifs for transcription factors and other DNA binding proteins that have not been expressed or whose functions have not been determined experimentally.

In the case of RBPs, the experimental approach has proven much more difficult. As a consequence, very few binding motifs are known to date and the rate of motif

discovery is likely to remain low. High-throughput microarray methods cannot yet be directly applied to assaying RNA binding proteins. The ability to computationally discover the preferred binding motifs is even more important for RBPs where sequence specific binding codes appear to play a greater role in modulating the cell state and response to environmental signals. Since much is known about the structures of RBPs and about the physical properties of macromolecular interactions, structure may provide a means for predicting RBP binding.

C. Computational Specificity Prediction and Design

RBPs form the bridge between digital information stored as a nucleic acid sequence and information stored as arrangements of protein structures and other molecules. Structures provide crucial information on how proteins interact with other molecules. Under the paradigm of structural biology, function follows form. Structure provides a rich source of experimentally supported information that informs predictions about function. Structure-based computational approaches use existing and predicted structures in the context of physical and statistical scoring functions to assist in evaluating the interaction between biological macromolecules such as proteins in complex with RNA. In this dissertation, I investigate whether a structure-based computational approach can be used to map how the amino acid sequence of RNA binding proteins determines its preference for recognizing specific bases on RNA targets.

Structure-based approaches assume that the details of protein function can be uncovered by applying calculations and chemical intuition to the interpretation of the structures. Ideally the structure of RBPs would reveal their specificity and their

functional role in a cell. In theory, the encoded sequence of amino acids and the environment including chaperonins define a unique structure for a protein (Y. Zhang, 2009). Projects such as the protein structure initiative aim to use computation and known structures to model all protein structures from sequence (Liu, Montelione, & Rost, 2007). Additionally, the physics of a protein of known structure in the context of its environment should precisely define its function (Whisstock & Lesk, 2003; Shenoy & Jayaram, 2010). Solving the general problem of predicting protein structure from sequence, the protein-folding problem, will require enormous insight and a monumental amount of work. The highest goal of *ab initio* prediction of function remains elusive (Redfern et al., 2008). However, progress on scoring functions allows homology modeling and interpretation of function.

Since proteins and their function have evolved, the structures of most domains can be modeled from known homologues. An early estimate was that 1000 structural folds explain most proteins (Chothia & Finkelstein, 1990; Chothia, 1992). Regardless of the classification of the representative folds, calculations suggest that the universe of compact single domain folds has been completely covered (Y. Zhang, Hubner, Arakaki, Shakhnovich, & Skolnick, 2006). The critical assessment of structure prediction (CASP) is a benchmark of progress on computational protein structure prediction that started in 1994 and is now in its eleventh round (Moult, Fidelis, Kryshtafovych, & Tramontano, 2011). Methods that use current structure in combination with physical and statistical scoring functions are now able to produce reasonably accurate structures (< 3 Å RMSD) where sequence similarity is greater than 60% (Read & Chavali, 2007). Even great accuracy in structure prediction may not be sufficient for function prediction

(Whisstock & Lesk, 2003). However, RBPs that recognize single stranded RNA employ a small number of domains with high structure similarity (Y. Chen & Varani, 2005). Thus, it should be possible to model the structures of RBP domains involved in binding.

General docking predictions have been used to assess whether scoring functions properly predict macromolecular interfaces. In the case of protein-protein complexes, calculations suggest that structure space is close to being fully explored (Gao & Skolnick, 2010). The CAPRI benchmark has been used to push forward computational approaches to protein folding and docking between proteins and other macromolecules (Lensink & Wodak, 2010a). The binding of proteins to RNA has hardly been benchmarked. The recent CAPRI round contained two instances of protein-RNA complexes for blind docking prediction. The results of docking the two RNA targets were disappointing in that no group succeeded in binding one of the models (Pons, Solernou, Perez-Cano, Grosdidier, & Fernandez-Recio, 2010). However, solving the docking problem is not necessary for gaining insight into the function of a few domains repeatedly employed in the recognition of single stranded regions of RNA.

Predicting sequence specific binding between two biological macromolecules such as proteins and nucleic acids should be a tractable problem where the structures are known or where homology modeling can be used. I am specifically interested in mapping how small variations in protein sequence influence the preference for RNA base sequences within the target. The specificity problem consists of understanding the sequence preferences of a small number of domains used by RBPs to recognize single stranded RNA sequences binding along a moderately conserved interface. Scoring

functions like those used in protein folding and the docking of molecules to proteins can be applied, but the sequence space that needs to be explored is substantially reduced.

The scoring functions that have been applied to predicting target sequence specificity of proteins binding other biological molecules generally fall into the broad categories of purely physical approaches and statistically sampled approaches. The purely physical approaches such as MD seek to fully model the forces in a complex and simulate the energy of a complex from first principles. In contrast, the statistically sampled approach seeks to sample structure space and minimize physical and empirical scoring terms. The statistically sampled methods are usually computationally less intensive. However, the computational cost of either method could be justified in an appropriate and successful application.

Both the purely physical and statistical approaches have been applied to modeling protein complexes with partner proteins and nucleic acid. Efforts to model docking, binding, and binding site predictions have been reported for protein complexes with protein, DNA and RNA. Works predicting target sequence specificity have been reported with binding to protein and DNA partners. To my knowledge, treatment of target specificity for RNA binding proteins has not been reported.

It is worth reviewing what has been done with physics-based and statistical approaches to predicting the binding specificity of proteins to other macromolecules. Much can be adapted from the work with DNA binding proteins, but unique properties of the more conformationally diverse RNA molecule provides additional challenges some of which I discuss in this dissertation.

1. Physics-based approaches

Molecular dynamics (MD) simulations usually seek to realize a complete representation of the physical forces acting on atoms in a molecular system. Inferences additionally rely on constructing a model with sufficient completeness that the minimization process converges toward correct structure or physically relevant local minima. MD has been employed in short simulations aimed at understanding protein-RNA interfaces and as part of docking protocols used in predicting protein-RNA complexes.

Starting from established structures, a number of studies have employed molecular dynamics (MD) to investigate the molecular basis for affinity and specificity of proteins binding nucleic acids (MacKerell & Nilsson, 2008; H. Wang & Laughton, 2009). These simulations have demonstrated that interface interactions between protein and RNA often involve short-lived and water mediated contacts (Castrignanò, Chillemi, Varani, & Desideri, 2002). As an example, MD was able to elucidate the role of specific positively charged residues at the U1A interface (Tang & Nilsson, 1999). Yet, while MD can yield insights into binding mechanisms, it is too computationally intensive to be used in systematically docking multiple RNA sequences with a protein domain.

Recently, a protein-RNA complex was included in round 15 of CAPRI (Lensink & Wodak, 2010b). While only a small subset of CAPRI participants submitted predictions for the protein bound to a double-stranded RNA, the approaches of the participants provide a sample of the hybrid methods used to attempt this task, often ported from modeling protein-protein interactions (Andrusier, Mashiach, Nussinov, & Wolfson,

2008). HADDOCK (de Vries et al., 2010) and SwarmDock (Xiaofan Li, Moal, & Bates, 2010) rely on physical potentials for initial docking and on MD for model refinement. Other approaches rely on mixtures of statistical and physical terms and a variety of simulated annealing algorithms (Pons et al., 2010; S.-Y. Huang & Zou, 2008) for side-chain optimization. These simulated annealing approaches provide good results while not depending on explicit solvation models and remain less computationally intensive than MD.

The main criticisms of MD are that it is too computationally intensive and that the imperfect force fields do not allow for robust solutions to many of the problems to which it is applied (Piana, Lindorff-Larsen, & Shaw, 2011). Approaches have been developed to improve the performance of MD using customized hardware (Shaw et al., 2010). A good choice of force parameterization allows for important structure based insight into local molecular structure (Schaeffer, Fersht, & Daggett, 2008). However RNA structure in general has posed significant challenges to methods of computational structure prediction (Laing & Schlick, 2011). Complete physical simulations of molecules containing RNA poses significant challenges that are beyond the scope of specificity problems I wish to address.

Statistical sampling of local structure and sequence is a more direct way to ask questions about specificity when a complete molecular structure is difficult to define or is modeled. Methods other than MD have thus been more widely employed by groups trying to solve the sequence specificity problem.

2. Statistical approaches

Currently, practical computation approaches to identifying correct protein-RNA complexes rely on a mixture of statistical and physical terms for scoring interaction decoys and for optimizing interface contacts. Heuristic terms that are not required in MD are required in these alternative approaches because conformational transition states are not explored during modeling and conformational sampling is performed at low resolution (Boas & Harbury, 2007).

Structure-based statistical approaches have not previously been applied (1) to predicting the specificity for a target RNA sequence for RBPs or (2) for altering the sequence specificity of RBPs. Examples of pure statistical approaches and mixed physical and statistical approaches with protein-DNA and protein-protein complexes have been reported. We may apply much of what has been learned in these other systems to the problem of proteins that bind to single stranded RNA in a sequence specific manner, but the RNA binding problem does present some unique challenges.

a. protein structure prediction

The development of terms used in scoring functions used for predicting the structure of biological macromolecules was largely driven by the challenge of predicting protein structure. A recent review discusses the performance of the methods currently used for prediction (Y. Zhang, 2009). The advances in protein structure prediction have come both from improved scoring functions and from improved methods of sampling structure space. The most successful scoring functions for protein structure prediction are generally those that use empirical terms learned from existing protein structures. Generally, these are either purely empirical scoring functions such as those employing

Bayesian statistics or they are weighted linear combinations of empirical and physics based terms such as that in Rosetta.

Bayesian scoring functions in protein structure prediction – Purely empirical scoring functions continue to be developed with strong theoretical justifications (Hamelryck et al., 2010). These potentials of mean force (PMFs) were one of the earliest approaches tried. Due to its roots in statistical physics and exponential growth in the number of solved protein structures (Y. Zhang & Skolnick, 2005), the empirical scoring functions continue to be among the most successful approaches (Rykunov & Fiser, 2010). Many types of machine learning can be applied to discoving the relationships in native structures needed to build a scoring or descrimination function. Bayes theorem is one of the most basic expressions of scientific logic (Jaynes, 2003) and has proven to be a good basis for these functions.

The Bayes theorem can be applied to learn scores for the distances between atoms in the PDB atomic coordinate files or similarly for relative orientations and distances between groups of atoms. The details of the distance dependent methods method are described in chapter Chapter 2.C. One of the first implementations was a scoring function based on the distances between the C_β of residues in protein folds (Hendlich et al., 1990). Several distance-dependent functions for protein structure decoy discrimination considering all atom positions have been very successful (Samudrala & Moult, 1998; E. S. Huang, Samudrala, & Park, 2000; H. Lu & Skolnick, 2001). Additionally, a block orientation approach was successful especially when it was applied to correctly positioning amino acid side-chains (M. Lu, Dousis, & Ma, 2008a). Scoring functions based entirely on empirical scoring functions and employing Bayesian

logic continue to perform well (Rykunov & Fiser, 2010). These scoring functions perform comperably to those that use physics-based terms (Ferrada, Vergara, & Melo, 2007).

Rosetta in protein structure prediction – Rosetta was developed as a way to tackle the protein structure prediction problem, but the applicability of the scoring terms and the algorithms for building and exploring macromolecular structure make it applicable to many problems of predicting biological structure. An overview of the new Rosetta 3 describing what it can do, starting with how it implements its original purpose of protein structure prediction, can be found in (Leaver-Fay et al., 2011). Rosetta describes a range of terms for assigning scores to residues with respect to other molecular structure and its environment. Many of the most important scoring terms are empirical. Molecules are scored with linear combinations of the weighted terms optimized for performance in scoring structure or complexes of each molecule type.

The Rosetta scoring function is among the most successful of scoring functions using physical and empirical terms. It continues to work competitively in recent CASP tests (Bradley et al., 2005; Raman et al., 2009). More impressively, Rosetta has been used in the design of proteins with novel folds and functions (Pantazes, Grisewood, & Maranas, 2011). The core scoring terms and the molecular modeling algorithms that are used in the protein structure prediction and design are applicable to the problems of predicting the affinity and the specificity of molecular interactions.

b. protein-protein recognition

The problem of protein-protein interactions is a molecular recognition problem that has the advantage of being chemically identical to protein folding. However, if all

bimolecular docking conformations are allowed, the potential sequence space becomes large. Adding the possibility of sequence changes to a model where all bimolecular interactions are allowed would result in an enormous search space. Thus, most work has concentrated on exploring either structure or sequence space.

Most scoring functions developed for proteins have been used in combination with docking algorithms or decoy discrimination tests (Ritchie, 2008). A few recent applications illustrate docking and sequence space exploration using Bayesian scoring functions and Rosetta. Both approaches apply developments from docking calculations to solve specificity problems involving short polypeptides that bind at a well-defined interface.

Bayesian scoring functions for protein-protein interactions – Empirical scoring functions are computationally efficient and thus continue to be developed for docking applications and sequence space exploration. Scoring functions based on observed distances between atoms have been used to predict protein-protein binding affinity (C. Zhang, Liu, Zhu, & Zhou, 2005; Su, Zhou, Xia, Li, & Sun, 2009). In addition to exploring various reference states, these scoring functions introduce the concept of volume correction to specifically account for the volume of space that could contain atoms belonging to the bound protein.

A near complete representation of the structure of protein-protein interfaces facilitates the development of novel empirical scoring approaches (Gao & Skolnick, 2010). A Bayesian scoring function implementing residue level terms for pair contacts and context with the protein showed impressive results (Feliu, Aloy, & Oliva, 2011). Another impressive use of a large dataset of peptide binding PDZ domains used a

Bayesian treatment of residue interactions to identify PDZ domain residues responsible for the specific binding of peptide sequences (J. R. Chen, Chang, Allen, Stiffler, & MacBeath, 2008). These implementations demonstrate clever use of large datasets to capture important features of structure and sequence space for proteins recognizing other proteins and linear epitopes. The case where a huge set of sequence interaction information is available for one system is unusual. Atomic level statistics are designed to address the more usual case where specificity information is available for a variety of different interfaces.

Rosetta for protein-protein interactions – Rosetta has been successfully applied to the challenges of protein-protein docking and to the design of interacting proteins. Das et al. (2009) demonstrated an application of simultaneous folding and docking of two proteins. More impressively, Rosetta was used in the de-novo design of a protein-protein pair with a K_d of 180 nM (Karanicolas et al., 2011). Directed evolution improved the affinity to physiologically relevant values.

Rosetta has also been used in predicting binding affinities and specificities of proteins for the sequence of small, unstructured peptides. Rosetta was used to recover the structure and sequence of peptides bound by proteins (Sood & Baker, 2006). Rosetta and homology modeling were employed to predict sequence specific binding to PDZ domains and the information content of recognized positions (King & Bradley, 2010; C. A. Smith & Kortemme, 2010). Smith and Kortemme (2011) report a method for finding protein positions that can be mutated and the mutations that can be tolerated at those positions. The authors also discuss applications of filtering designs based on tolerance to a use in designing proteins to target alternate PDZ sequences.

The structural aspects of the problem of designing proteins that bind short peptides are similar to those required for RBP design. The degree of flexibility of short peptides is similar to that seen in single stranded RNA regions. The methods developed by Smith and Kortemme (2011) could be used to develop ensembles of protein structure for redesigning RBPs.

c. approaches to protein-DNA recognition

Approaches used in evaluating the specificity of proteins that bind to DNA are a good starting point for developing methods for RBPs. The chemical properties involved in protein-DNA interactions are nearly identical to those required at protein-RNA interfaces. The chemical differences include the additional 2'-OH group in the ribose backbone and a methyl group on thymine. The differences in the structure of the protein interface with double-helical DNA significantly change the scoring matrices that are learned in the machine learning approaches to interface structure prediction.

Bayesian approaches to protein-DNA interactions – Distance-dependent Bayesian scoring functions developed for protein complexes with DNA were the basis for the statistical method I apply to RNA in this dissertation. The specifics of Bayesian scoring functions and the concepts such as reference state and volume fraction that arise are discussed extensively in Chapter 2.D.

Distance-dependent Bayesian scoring functions have been used for docking decoy discrimination and predicting binding target sequence specificity. Zhang et al. (2005) report a Bayesian scoring function with a reference state based on the ideal gas formula and atom types defined by chemical properties. Robertson and Varani (2007) compared several reference states and atom type definitions for use in docking decoy

descrimination. The distance-dependent Bayesian scoring functions have been used for specificity applications for proteins specifically binding DNA sequences (Xu, Yang, Liang, & Zhou, 2009). One successful approach to increase the amount of training data is to use an unrelated dataset such as protein interactions with small molecules (Bernard & Samudrala, 2009). The reference states, atom type definitions and training sets are discussed in Chapter 2.D, where I also discuss developing a Bayesian scoring function for evaluating the specificity of RBPs.

Rosetta in protein-DNA interactions – Rosetta has proven to be one of the most successful approaches to the prediction of the specificity and affinity of proteins that bind to DNA and has been used in several applications for transcription factor design. Rosetta successfully reproduced the experimentally determined binding motifs DNA binding proteins (Morozov, Havranek, Baker, & Siggia, 2005). Morozov et al. (2005) also found agreement between the reported changes in binding energy due to changes in the bound DNA sequence and calculated values. More recently, Yanover and Bradley (2011) included local exploration of protein backbone and some flexibility about the nucleic acid glycosidic bond with concordant adjustments to the deoxyribose structure. The more flexible approach allowed accurate prediction of zinc finger motifs.

Work investigating the retargeting of homing endonucleases provides a dramatic, experimentally verifiable test of the use of the Rosetta scoring function for predicting affinity and specificity. An altered endonuclease can be confirmed by visualizing the correct cleavage products on a gel (Chevalier et al., 2002). Successful retargeting of the binding sequence recognized by homing endonucleases is described for a single and for multiple adjacent recognized base-pairs (Ashworth et al., 2006, 2010). The designed

homing endonucleases bind to DNA with nanomolar affinities. The design application confirms that the altered specificities represent functional changes that can be experimentally verified.

Other approaches to protein-DNA interactions – The DNA specificity prediction and design problem is largely analogous to what I am attempting with RPBs. Other approaches attempt to solve the problem of understanding and altering the specificity of proteins binding to DNA. Other approaches such as molecular dynamics inform our understanding of specific binding of nucleic acids.

A complete molecular dynamics simulation is difficult to apply to molecular complexes in a way that allows exploration of sequence space. However, physics-based scoring functions can be used to evaluate the energy of candidate structures generated through rethreading and that allow only local minimization of structure (Paillard & Lavery, 2004). The physics-based approach may be more sensitive to small structure variations than approaches that include statistical terms.

Machine learning approaches may also be used to incorporate non-structural information. Andrabi, Mizuguchi, Sarai and Ahmad (2009) used neural networks to find the most predictive relationships between structure, amino acid compostion and the recognized base pair. The inputs to the neural network included information from evolutionarily related sequences. Specific recognition of base pairs could be predicted based on amino acid sequence, evolutionary profile and basic structure, but the atomic contacts in the major groove could not be predicted (Andrabi et al., 2009). This suggests that while there is still no simple recognition code there are strong relationship between amino acid sequence and the recognized sequence, when the basic structure is

known. However, these results may more strongly apply to DNA than RNA recongition interfaces because there is larger conservation of DNA structure and the location of residues that interact with DNA.

Methods for designing proteins are also being codified using scoring functions other than Rosetta. The idea of using predicted position weight matrices to guide the retargeting of proteins that bind to DNA was described for the FoldX scoring function (Alibés, Serrano, & Nadra, 2010; Nadra, Serrano, & Alibés, 2011). Redesign may be achieved by creating libraries of mutants for DNA-contacting positions and selecting amino acid positions that approach a desired binding motif. The approach is supported by an example where many changes in binding specificity for PAX6 SNPs are predicted (Alibés, Nadra, et al., 2010). A similar approach to design may be applicable to RBP design.

D. Challenges in RBP Specificity Prediction and Design

The structure-based approach to RBP specificity and design requires the accurate prediction of the structure and energy of protein-RNA interfaces. Solutions to predicting RBP binding energies from structure may be borrowed from previous work with other biological macromolecules that engage in sequence specific recognition. The RBP target specificity shares a strong chemical identity with the problem of DNA binding and a similarity with protein-peptide recognition in terms of the size of the structure space.

1. Transferability of protein-DNA scoring functions

Scoring functions successful in predicting DNA binding protein specificity are natural starting points for tackling the RBP specificity problem. Physics-based energy functions and many empirical scoring functions for molecular interactions between protein and DNA should work with only small changes for complexes with RNA. Scoring functions containing empirical terms may need to be retrained and term weights may require adjustment.

Scores for empirical atomic distance-dependent scoring functions are usually formulated in such a way that the residue and base identity will influence the Bayesian inference. Previous work in the Varani laboratory, demonstrated the effectiveness of the all-atom distance-dependent scoring function for protein-DNA docking decoy discrimination (Robertson & Varani, 2007). Later work demonstrated that an identical scoring function could be trained on protein-RNA structure and used for docking decoy discrimination and complete rethreading tests with RBP complexes (Zheng, Robertson, & Varani, 2007). The DFIRE distance dependent scoring function was shown to correctly predict protein-DNA binding energy and to predict binding motifs (Xu et al., 2009). Distance-dependent scoring functions have not been used for binding motif prediction, but it has been applied to a rigourous task of predicting RNA binding sties (H. Zhao, Yang, & Zhou, 2011). We expected that distance-dependent scoring functions could be applied to specificity prediction problems.

In scoring functions such as Rosetta, the energy terms are mostly defined in terms of energies from physical properties or from empirical properties that are residue independent. Thus, the components of the score should transfer across

applications especially to systems as chemically similar as RNA bases. However, the relative weights of terms in scoring functions are generally optimized for a particular scoring application.

A scoring function that employs multiple energy terms may be weighted preferentially for DNA specificity applications over their RNA counterpart. Bound DNA is found almost exclusively in a B-form double-helical conformation. In contrast, RNA's more complex structure presents aspects of nucleotide chemistry such as the base edge that are rarely accessible in DNA. Specific contacts with DNA bases are heavily biased in number toward the base edge exposed in the major groove of the double helix. The scoring function optimization may not properly weight terms important for contacts between protein atoms and nucleic acid base edges. The relative weight of a term that is less consequential to DNA design applications may reduce the effectiveness of the complete scoring function in the RNA application.

Protein-DNA scoring functions should have addressed most conceptual challenges in scoring protein complexes with nucleic acid. Empirical scoring functions trained exclusively on protein-DNA complexes need to be retrained. The essential difference between the double-helical DNA and single stranded RNA targets is in the accessible structure space.

2. Transferability of protein-peptide scoring applications

The problem of predicting the RNA target sequence specificity of RBPs that bind single-stranded RNA is conceptually similar to the problem of peptide-binding proteins in terms of the accessible structure space of the target. The contacts on the protein

participating in the sequence specific binding interface are significantly less conserved in the peptide binding case.

Protein-peptide methods suggest that in spite of the flexibility of the peptide predictions work with near native sequences. The prediction of binding motifs for PDZ domains was succesfully accomplished using the Rosetta scoring function (King & Bradley, 2010; C. A. Smith & Kortemme, 2010). Both groups used a small amount of PDZ domain flexibilility and some local protein flexibility in the specificity predicitions. An interesting extension to this work shows that there are intrinsic constraints to the sequence of the protein side of the peptide binding interface (C. A. Smith & Kortemme, 2011). These constraints should be applicable to proteins that bind RNA and should inform future design work.

3. RNA specific advantages and challenges

The RBP specificity and design problem shares many of the chemical properties of the DNA binding problem and resembles the peptide-binding problem in terms of structure space. However, the RBP binding problem is in a few respects more tractable than the analogous problems. RBPs that bind single stranded sequences have more opportunities for specific contacts with the base edges and the sequence space of RNA is significantly smaller than peptide sequence space.

The problems we address computationally can be formulated so that useful results may be obtained by exploring limited structure and sequence space. In regions of single stranded RNA from disordered regions or from the loop region of stem-loop structures, base positions are likely to be independent to a good approximation. As long as base positions are mutated one at a time and protein sequence is only mutated at a

few positions, the number of residues that form the binding interface is not large. Thus, a limited exploration of sequence space is possible without generally solving the complete RBP binding problem. Possible sequence space is also significantly smaller if base position independence is considered. While structure and sequence space remains large, these limited approaches seem computationally tractable.

The development of a computational approach to RBP binding specificity presents additional challenges due to limited experimental data (section B.2.b above). Constructing challenging and informative test sets is difficult due to the relative paucity of solved structures and experimental measurements of specificity.

4. Reasons for computational specificity prediction and design

As discussed in section A.2 (above), understanding protein-RNA binding specificity is critical to understanding the post-transcriptional state of the cell. Demonstrating that, even in the absence of a binding code, RBP specificity preferences can be inferred computationally from structure is important to understanding post-transcriptional regulation in cells. Experimental techniques are not likely to scale up sufficiently for this task.

RBPs binding single stranded regions of RNA are a reasonable starting point for a number of useful design applications (Mackay, Font, & Segal, 2011). That nature has repeatedly employed a small set of domains for sequence specific recognition (Kerner et al., 2011) implies that these domains are highly versatile. These domains should be good starting points for designing RBPs to target just about any sequence. Furthermore, the domains may be linked together to build proteins that recognize unique regions of RNA (Mackay et al., 2011). The efficiency of binding, the number of protein residues

needed to recognize each base, is less than zinc finger domains being successfully used for genome engineering. However, domains such as the RRM and KH domains still represent good starting points for the design of RNA-targeting proteins.

E. Thesis Outline

My work extends previous work in the Varani laboratory building computational approaches to nucleic acid binding and specificity (Robertson, 2007). I concentrate specifically on the problem of sequence specific recognition of RNA by RBPs and approach the problem of redesigning RBPs to target alternative sequences. My solution to these problems is broken down into smaller pieces. I approach each problem conservatively, starting with the least parameterized scoring functions for specificity prediction and in finding the smallest change in protein sequence that will allow for retargetting domain specificity.

In Chapter 2, I explore a machine-learning approach to creating a scoring function for use in scoring the structure of RBP-RNA interfaces and to scoring base and residue positions for inferring specificity.

In Chapter 3, I perform benchmarks of structure-based specificity prediction using computational tools and various scoring functions. I compare two pure statistical scores and the Rosetta mixed physical and statistical scoring functions

In Chapter 4, I perform proof of principle calculations for a biologically relevant RBP retargeting application. I also discuss the use of data from a microarray approach to comprehensively measure RBP specificity for improving computational approaches to specificity prediction.

Figures

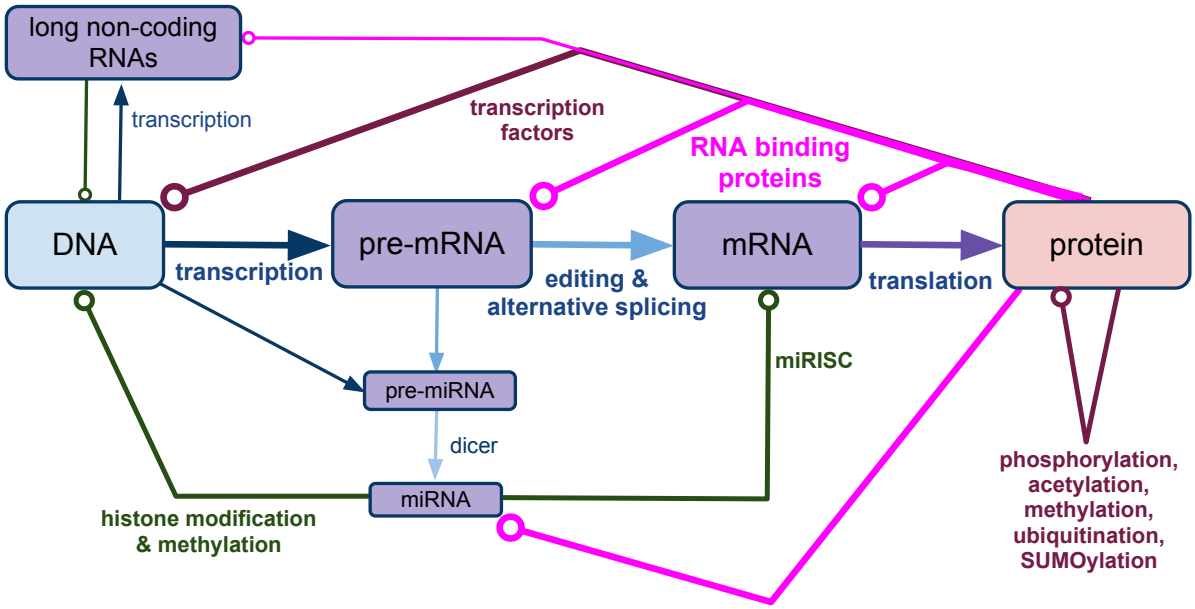


Figure 1.1: The interactions between biological macromolecules by molecule type.
Proteins and RNAs serve regulatory functions promoting and inhibiting protein expression by binding to DNA and RNA in specific and non-specific interactions. Until recently, the importance of proteins that regulate RNA in specifying cell state and in coordinating cellular responses to the environment was not fully recognized. The diagram indicates the molecular classes that participate in regulatory interactions. The specifics of molecular composition and their interactions determine the information carried along the interaction conduits. Open circles represent regulatory interactions. Regulatory interactions of proteins binding RNA are highlighted in bright magenta.

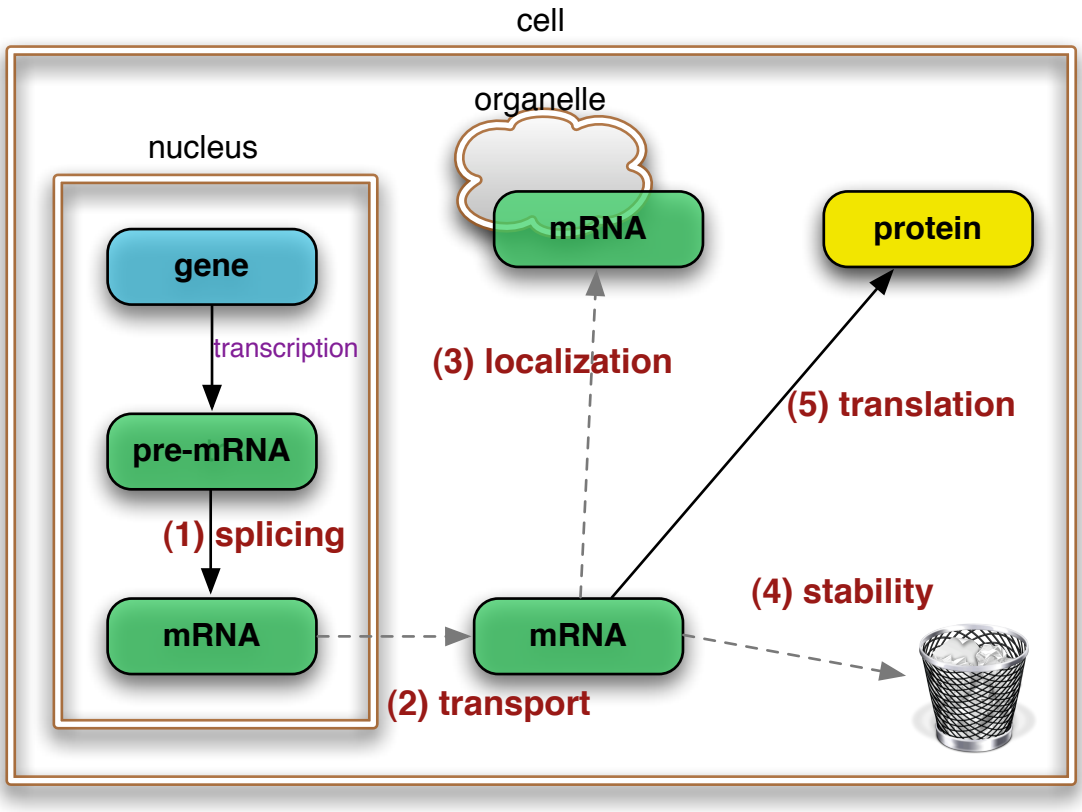


Figure 1.2: Roles played *in vivo* by RBPs that recognize RNA in a sequence specific manner.
The diagram highlights RNA binding proteins (RBPs) specifically interacting with pre-mRNA or mRNA in eukaryotic cells. Sequence specific recognition plays roles in (1) splicing, (2) transport, (3) localization, (4) stability and (5) translational control. RBPs bound to RNA are referred to as ribonucleoprotein complexes (RNPs) RBPs bound to pre-mRNA in the nucleus are referred to as heteronuclear RNP complexes (hnRNPs). RBPs binding to processed mRNA are messenger RNP complexes (mRNPs). The composition of RNP complexes holds part of the information state of the cell.

Chapter 2. An Empirical Scoring Function for the Structure-Based Prediction of Specificity in RNA Binding Proteins

A. Introduction

RNA binding proteins (RBPs) are key players in the post-transcriptional regulation of gene expression. RBPs bind to RNA transcripts often in a sequence specific manner and effect phenotype by determining transcript transport and localization, modulating alternative splicing, the translation of mRNA transcripts and RNA metabolism (Lukong et al., 2008; Mansfield & Keene, 2009). Understanding each of these processes necessitates a complete understanding of the RBPs that participate in post-transcriptional regulation. The function of RBPs is dependent upon their expression, post-translational modifications such as phosphorylation, and the specificity and affinity with which they recognize and bind to specific regions of single stranded RNA (Glisovic et al., 2008; Mittal et al., 2011). Since it is essential to understanding post-transcriptional gene regulation, I seek to predict the specificity of binding to RNA targets from known structures of RBPs.

RPBs employ a handful of binding domains for the sequence specific recognition of RNA. The small number of protein domains employed by proteins that bind to RNA in a sequence specific manner suggests that it should be possible to infer the preferred RNA binding sequence from the protein sequence and structure (Y. Chen & Varani, 2005; Ellis, Broom, & Jones, 2007). However, no simple code has been found to predict which base may be bound by a given set of residues or interaction types (Auweter et al.,

2006). However, a machine learning approach may still be able to discover a relationship between protein sequence and RNA target indirectly through structure. Thus, I employed a knowledge-based structural approach to predicting target RNA from protein structure and sequence. Knowledge-based approaches are well suited to situations where the problem is well defined, but where patterns cannot be recognized by inspection.

The factors that make it impractical to experimentally determine the preferred RNA binding sequences of a large set of RBPs make the discovery of RBP binding targets from structure and sequence a particularly attractive application for a knowledge-based structural approach. A recent database summarized the current knowledge in literature about RBP target RNA sequences. The RNA binding protein database (RBPDB) v1.2.1 catalogues 1456 experiments for 1171 proteins including 424 human proteins (Cook et al., 2010). However, most of these binding sequence predictions only report a single preferred binding sequence. In most cases, the literature reports multiple distinct binding sequences for a protein. The reported sequences do not correspond in length to the recognition region of a given RBD. Full specificity information is available for less than a few dozen proteins. Even with the development of a new micro-array technique for fully characterizing binding specificity for proteins binding up to a heptamer of RNA (Ray et al., 2009), experimentally mapping RBPs to bound sequences would be impractical due to the difficulty of expressing and assaying each protein. Thus, I sought to infer RNA binding targets for proteins of similar structure to those whose target preferences are known.

Structures provide a rich source of information from which RNA target specificity should be discoverable. Structures have been determined for many important RBPs in complex with their preferred RNA target sequence. As of 2011, over 250 structures of RBPs have been determined in complex with RNA. These existing structures provide valuable insight into the mechanism of binding by RBPs, but the problem of discovering the RNA sequence specificity has not been solved.

Since the known structures contain several examples of each of the common RNA binding domains, I expected that the corresponding coordinates can be used as a basis for learning the relationship between protein sequence and preferred target RNA sequence. Table 2.1 lists the protein domains that account for most sequence specific binding of proteins to single stranded RNA. Pfam predicts the number of these domains encoded in most known genomes and in the human genome. The known structures represent only a small fraction of the total number of existing complexes. However, the relationship between local protein sequence and structure and preferred bound RNA bases should be discoverable from the complete set of bound RNA structures, including those that do not contain the RNA binding domains listed in Table 2.1. Given an accurate scoring function and a comprehensive approach to exploring local structure, the existing structures of the domains of interest should provide a basis for inferring target sequence and for the subsequent redesign of the specific binding activity of theses protein domains.

My work extends previous work in the Varani laboratory building tools for the computational retargeting of RNA binding proteins. I applied knowledge-based scoring functions to the problem of sequence specificity at protein-RNA interfaces. Knowledge-

based scoring functions, alternatively referred to as an empirical potentials or potentials of mean force (PMF), have been thoroughly investigated for the structural studies of proteins (Hamelryck et al., 2010). Work in the Varani laboratory validated the use of these scoring functions for identifying correct complexes of protein and DNA (Robertson & Varani, 2007) and of protein and RNA (H. Zhao et al., 2011; Zheng et al., 2007) from sets of near native decoys of the complexes. These docking decoy results suggested that statistical scoring functions could be used for important applications beyond docking, for example for inferring the preferred RNA sequence that will be bound by a RNA binding protein domain. My work has explored the application of this empirical scoring function to predict the specificity of RNA binding proteins.

In this chapter, I describe my approach to parameterizing and training a statistical scoring function. I implemented the statistical approach as part of an application developed in the Varani laboratory and subsequently within the Rosetta molecular modeling framework. Additionally, I present the challenges in applying the statistical approach to specificity predictions and discuss specific examples of insufficiency in the training data sets.

B. Background

Knowledge-based scoring functions are a common approach to molecular prediction applications (Hamelryck et al., 2010). These purely empirical approaches to protein structure prediction have performed as well as more complex approaches involving physical terms. The empirical approaches have fewer adjustable parameters than statistical approaches consisting of weighted physics-based and empirical terms.

Statistical approaches are less demanding in terms of the completeness of the molecular environment than purely physics-based approaches. Previous members of the group chose to first attempt a purely empirical approach because it is simpler to implement and was likely to sufficiently capture the specificity of binding interactions. The predictions are expected to be robust even if the scoring function does not fully determine the energy landscape with the detail needed for perfect prediction of the protein-RNA interface.

Computational approaches to the prediction of protein binding interactions employing purely physical terms have also been applied to the protein-RNA binding problem. However, physical approaches such as molecular dynamics have not been more successful than approaches that include statistical terms. If it were possible to construct a perfect force field for every atom in a protein, a molecular dynamics approach to predicting protein structure would be preferable despite of the computational expense of the approach (Shaw et al., 2010). However, the present state of the art does not allow perfect MD simulations of bound proteins, in part due to limits on the thermodynamic properties that affect the exploration of structure space at the interface (Wereszczynski & McCammon, 2011). Thus, the most successful approaches to protein structure prediction have employed some statistical treatments of existing protein structures.

When knowledge based terms are included in a scoring function, the scoring function itself is not assumed to be a perfect, differentiable representation of forces in the proteins of the type that could be used in a MD simulation. Given a good selection of a statistical model, enough training data and carefully chosen weighting factors in cases

where multiple independent statistical or physical approaches are used, statistical scores will be nonetheless predictive of interaction energies (Ferrada & Melo, 2009). Since the terms have imperfections and limited training data may necessitate discretization of a variable, structure space is not searched by relaxation down a force-field gradient. Instead, structure-space is sampled from re-arrangements of dihedral angles guided by the frequencies with which interactions and conformations are observed in known structures.

Studies of statistical approaches have demonstrated that a well-parameterized statistical scoring function trained on known structures can faithfully reproduce the energy landscapes expected from physical treatments of the score (Hamelryck et al., 2010). The most common approach to knowledge-based scoring functions is to use Bayesian statistics on a measurable feature of existing structures. Measurable features that are completely defined by structures include the distances between atoms and the values of dihedral angles (Ferrada & Melo, 2009). If a reference state can be established, then the scoring function will increasingly approximate a complete physical potential as the training set grows and becomes more comprehensive (Hamelryck et al., 2010). In the course of my thesis work, I have implemented and tested an empirical scoring function for evaluating the energetic consequences of mutations to protein sequence upon a protein preference for specific RNA sequences.

The all-atom distance-dependent scoring function correlated with change in energy in large scale docking and rethreading experiments, but was less successful in repacking tests and single position specificity tests. It was shown that scores of RNA sequence mutants in some RNPs using the all-atom scoring function correlated well

with experimental change in free energy of binding (Zheng et al., 2007). However, I have established through my work that the scoring function is not sufficiently trained to accurately enough to reproduce some features of the structure at protein-RNA interfaces or to perform well in more stringent tests of specificity.

C. All-Atom Distance-Dependent Scoring Function: Parameterization and Training

We tested a statistical scoring function based on pairwise distances between typed atoms in known protein-RNA complexes. The model was chosen based on insight from knowledge-based approaches to scoring protein-DNA interactions (Robertson & Varani, 2007; Xu et al., 2009). I discussed Bayesian approaches to protein-DNA structure and specificity in Chapter 1.C.2.c.

The choice of the parameters of the model and of the definition of the set of structures used as training data were complicated by the limited number of protein-RNA complexes for which structures have been solved experimentally. Here, I introduce the Bayesian model chosen for predicting correct protein-RNA structure and extended to predict RNA sequence specificity (Zheng et al., 2007). I discuss considerations given to choosing and adjusting the training sets used for specific scoring problems.

1. Model selection

Bayesian inference is one of the simplest and most powerful approaches to testing predictions based on known prior information (Jaynes, 2003). Sippl (1990) first proposed the Bayesian approach to structure using pairwise distance measurements for use in the protein-folding problem. These potentials of mean force have continued to be developed over the intervening decades because of their ability to approximate

the energetics of proteins (Hamelryck et al., 2010). The use of a pairwise distance potential has been shown to work in practice with protein structure prediction (Samudrala & Moult, 1998), with transcription factor binding and specificity (Robertson & Varani, 2007; Xu et al., 2009), and with protein-RNA docking and complete rethreading applications (H. Zhao et al., 2011; Zheng et al., 2007). Here, I reintroduce and update the model for application to predicting the specificity of RBPs for given RNA sequences.

We chose to model protein-RNA complexes based on Euclidean distance between atom pairs. The pairwise distance is a simple metric that may be calculated from coordinate structure files. The distance metric requires no additional information about the structure except what is implicit in choosing atom types.

The Bayesian scores become meaningful when atoms are defined or grouped or typed such that the scoring function can learn atom pair types and distances that are characteristic of correct structures. The most basic atom typing strategy is to assume that each named atom in each standard residue or base is chemically unique (Samudrala & Moult, 1998). We refer to our model based on pairwise distances and greedy typing as the *all-atom distance-dependent scoring function*. This scoring function is a simple yet powerful machine-learning method that can be trained on known experimental structures.

2. Bayesian pairwise distance model

A former member of the Varani group, Dr. Tim Robertson, implemented the all-atom distance-dependent scoring function for use in docking applications involving protein complexes with DNA and RNA (Robertson, 2007). The all-atom distance-

dependent scoring function has a number of model parameters that must be defined in order to practically apply this model. These were tested in the context of protein-DNA docking decoy discrimination tests (Robertson & Varani, 2007). Aspects of the model must be defined to optimize its performance with respect to data available for training. However, once the model is established, it contains no adjustable parameters and the model may only be improved by adding additional training data.

model definition: The naïve Bayes classifier assumes independence among the features of a system. In the all atom model, the features are the atom types defined by residue name and atom name. Since the positions of atoms within residues are dependent on each other, the independence assumption is an approximation. However, we retained this assumption since atom types have the appearance of independence within the context of the whole molecule or complex. We began with the Bayes theorem:

$$P(C \mid D) = \frac{P(D \mid C)P(C)}{P(D)}, \qquad (2.1)$$

where *D* is the set of all the measured distances annotated with atom types and *C* represents a physically valid and correct structure. P(C), the marginal probability of finding a correct structure, can be seen as an arbitrary constant and ignored, because we are interested in ranking candidate structures for which P(C) would have a similar value. If types and distances are chosen well, the Bayesian formalism learns the atom-type pairs and distances that, when observed, are most likely to be present in a correctly bound structure.

Taking the log of the Bayes equation, we can estimate the free energy of the complex from the sum of log odds scores derived from inferred distance probabilities:

$$G \approx \log P(C \mid D) = -\sum_{t_i^p} \sum_{t_j^r} \frac{P(d_{ij}, t_i^p, t_j^r \mid C)}{P(d_{ij}, t_i^p, t_j^r)} \quad , \tag{2.2}$$

where d_{ij} is the distance between the i^{th} protein atom of type t_i^{p} and the j^{th} RNA atom of type t_j^{r}. Using these log scores also makes the problem more computationally tractable by avoiding overruns in floating point calculations. The connection between the Bayesian scores and energy has been argued by analogy to the reversible work theorem (Sippl, Ortner, Jaritz, Lackner, & Flockner, 1996).

The practical application of the Bayes theorem is dependent on the assignment of atom types. In the limit where the number of counts is very large, the prior distribution for the atom pair with distance d_{ij} and atom types t_i^{p} and t_j^{r} is the frequency of observations of atoms of the same type at that distance to atoms of the same types observed at any distance in the training set.

$$P(d_{ij}, t_i^{\text{p}}, t_j^{\text{r}}) \approx f(d_{ij}, t_i^{\text{p}}, t_j^{\text{r}}) = \frac{N_{\text{obs}}(d_{ij}, t_i^{\text{p}}, t_j^{\text{r}})}{\sum_{d_{ij}} N_{\text{obs}}(d_{ij}, t_i^{\text{p}}, t_j^{\text{r}})} \tag{2.3}$$

Due to the finite amount of training data, we need to discretize the distance measurements in order to obtain accurate frequency estimates. Additionally, infrequent combinations of distance and type pairs may appear sparsely in the training and therefore may be erroneously taken to be improbable. Since we want the classifier to accurately discriminate structures based on evidence in a finite training set, pseudo-counts are assigned by one of several methods. Given an accurate reference state

discussed below, a value may be obtained for the probability of a given structure based on the known structures.

The components of the model which must be defined include the atom type scheme, the reference state $P(D)$ to be used, the bin size for discretizing distances, and the low count correction. These components must be optimized to make best use of an available training set. However, the performance of the scoring function is most strongly dependent on the quantity and quality of the training data.

selection of the reference state: The selection of a reference state affects the ability of the learning function to distinguish meaningful features from background. The construction of reference states for the pairwise distance formalism has been described in several manuscripts (Xu et al., 2009; C. Zhang et al., 2005; Zhou & Zhou, 2002). The definition of the reference state is harder to establish *ab initio* in the case of an interface scoring function because atom types do not mix and thus a theoretical approach must approximate or justify ignoring the edge effect (H. Lu, Lu, & Skolnick, 2003). Cognizant of this added difficulty, we tried several reference states before settling on the empirical approach. I discuss these here because they may help to explain why the scoring function performs better in docking than with specificity calculations.

A correct reference state cannot be easily defined for pairwise distance measurements even in the case of large globular proteins (Ferrada & Melo, 2009). In the naïve Bayes approach, we have already assumed the independent assortment of atoms of the various types, but this is not strictly true since the structures are composed from a finite library of residues. The reference state is thus subtly dependent on sequence in a way that is hard to specify mathematically and difficult to capture empirically from a

finite training set (Solis & Rackovsky, 2009). Additionally, the variable sizes of macromolecular structures and their inherent structure make it difficult to theoretically construct a reference state that would correspond to unordered assortment of those atoms. However, Bayesian scoring functions that have not addressed all of these concerns have remained successful.

Theoretical reference states have been proposed based on statistical physics principles. Many groups have used variations of what would be expected in an ideal gas as a reference state (C. Zhang et al., 2005; Zhou & Zhou, 2002). The ideal gas assumption makes sense in large proteins where, in spite of the obvious distance constrains for atoms within residues and the accompanying sequence dependence, the atoms of neighboring residues appear to be randomly distributed. In the case of protein-DNA interactions, the ideal gas approximation yielded similar results to an empirical reference state constructed from an atom-type agnostic count of pairwise contacts in the training set (Robertson & Varani, 2007). Small corrections to the ideal gas approach have been tried in order to make it more applicable to a molecular interface. Xu et al. (2009) corrected for the volume fraction using an empirical term for the average fraction of space occupied by atoms from each type of molecule. However, even with these corrections the ideal gas reference state does not provide a useful approximation of a disordered protein-nucleic acid complex.

The alternative to an ideal gas reference state is a purely empirical approach, where the background distribution of atoms is inferred from the training-set. We assume, as in the ideal-gas reference set, that all atom-types share a background distribution (Sippl, 1990). The underlying assumption of the pairwise distance

approach to the Bayes classifier was the independent assortment of atoms. Combining those assumptions, we may approximate the background empirically from our training set by setting the reference state to the distance-dependent distribution of atoms when type is ignored. While this approach makes the reference state dependent on the training set, the empirical reference state does ensure that scores are based on significant deviation from the training set (Solis & Rackovsky, 2006). The volume fraction correction is also implicitly included in this approach.

The lack of a significantly better docking decoy discrimination for the ideal gas reference state over the empirical reference state in protein-DNA docking interactions lead us to prefer the empirical approach. For predictions based on the independent assortment of atoms, the volume density of the atoms is likely to be the most important factor in the reference state (Ferrada & Melo, 2009). We expect the raw numbers of atom contacts, the number of pairs of protein-nucleic acid pairs separated by less than the cutoff distance, to be fairly stable for even small training sets. However, the density of the interfaces will be dependent on molecular size and structure.

The approximations used in constructing the pairwise potential and the reference state becomes more significant when we use the potential for evaluating the atomic structure of the molecular interface. While assuming the independent assortment of atoms makes sense for the overall structure of large, globular proteins, it is clear that at the protein-RNA interface, protein side chain atoms are more likely to be found than backbone contacts (Morozova et al., 2006). A correct potential of mean force would contain an additional probability term accounting for residue structure and protein fold (Hamelryck et al., 2010). However, we prefer to start with a model that

assumes that an independent assortment is able to capture key features, as models with fewer features are less likely to be over-trained.

3. Training the pairwise distance scoring function

The all-atom approach assigned atom types to correspond to the set of all PDB residue name and atom name combinations found in the complexes being evaluated. The score is an interface score, so that only distances between pairs of atoms where one atom pertains to protein and the other to nucleic acid are considered. Applying this formalism to protein-RNA problems requires that the classifier be trained on a representative subset of the available protein-RNA structures.

In performing benchmark tests with the limited amount of data about protein-RNA structures that is available, we were concerned about under-training for the general problem or in situations where the available data concentrates on a specific set of structurally redundant proteins for overtraining for a specific problem (Solis & Rackovsky, 2008). In order to address these concerns, I carefully selected the training set to maximize diversity of structure and to include a similar number of examples of interactions between all residue types. Construction of the training set and the concept of 'fair training sets' are described later.

The data for protein-RNA complexes in the PDB are obtained from x-ray crystallography and NMR. The number of protein-RNA structures in the database is small relative to those available of protein-DNA. Figure 2.1B shows that more than a thousand protein-DNA structures were deposited in the PDB prior to 2005 with the number of solved structures continuing to grow exponentially. The number of structures of complexes of protein-DNA has grown at an exponential rate with the

number of reported complexes doubling every 4.5 years. The available protein-DNA training data is now large especially given the minimal structural diversity of DNA structure.

The size of a potential protein-DNA training set is an order of magnitude greater than that for the protein-RNA case. Figure 2.1A shows the cumulative number of protein-RNA structures deposited in the PDB by year. When the all-atom scoring function was trained and used for discriminating native protein-RNA complexes from docking decoys (Zheng et al., 2007), there were less than 100 deposited x-ray crystal structure. Anticipating that tests inferring sequence specificity from point mutations would be significantly more demanding than docking applications, I constructed a new training set to take advantage of the 50% increase in structure data during the intervening three years.

Protein-RNA complexes were chosen for the training set in a manner intended to represent the structural diversity in the existing structures. Given the limited number of structures, no attempt was made to select for specific structural properties in the bound RNA. We assumed that there would be sufficient structures in the database to be a representative sample of the complete set of protein-RNA complexes. While there is no guarantee that the existing PDB would satisfy the conditions of diversity and sufficiency, the history of protein structure discovery coupled with the natural tendency of structural biologists to pursue the most novel structures, suggests that the database represents most important protein folds and interaction features (Chothia, 1992; Gao & Skolnick, 2010).

The increased and less well understood structural diversity of RNA might not be sufficiently represented. In contrast to DNA, which is nearly always found in B-form double helices where the backbone structure is highly constrained, varying only a few degrees in a base pair composition dependent manner (Farwer, Packer, & Hunter, 2006), RNA adopts a variety of secondary and tertiary structures. The most common forms of RNA structure at a protein recognition site are the A-form double helix and single stranded regions. However, many RNAs, including tRNA and ribosomal components, have higher order structures that can serve a functional role similar to protein (Bahadur et al., 2008). Many of the single stranded regions constitute part of larger RNA structures such as the loop in a stem loop structure. This added structural diversity increases the number of residue atom and base atom pairs available to interact in a chemically meaningful way.

Chemically important interactions between RNA and protein residues are not likely to segregate by the evolutionary relationship among the RBPs. Instead, general features of the binding pocket are likely to confer specificity for a particular nucleobase (Morozova et al., 2006). Positively charged amino acids make preferential contacts with RNA backbone phosphates, the single stranded regions expose base edges to the protein interface (Lejeune et al., 2005). With sufficient training data, maintaining a representation of RNA structures proportional to those bound by the complete set of RNA binding proteins would be prudent. Alternatively, since my interest is in predicting the binding sequence of single stranded RNA regions bound to proteins, I would have liked to be able to construct a training set from just proteins bound to single stranded RNA. However, with the limited amount of RNA structure data available,

I opted for the greedy approach of including all unique structures likely to represent a different fold.

I selected structures for the training set to maximize the structural diversity of the protein chains. Domain classification databases demonstrate that the secondary and tertiary structure of protein domains is robust with respect to sequence. Domains may contain nearly identical folds with as little as 20% sequence identity (Chothia & Lesk, 1986; Flores, Orengo, Moss, & Thornton, 1993). While the PDB is now annotated with metadata about structure classification, quantifying the structural relationship between two protein chains is difficult and often tied up with fold such as in the structural classification of proteins (SCOP) (Andreeva et al., 2008). Therefore, I chose to use sequence similarity as a metric for structural similarity.

I used well-developed tools to choose structurally diverse proteins for the training set based on sequence. In early 2009, I selected all PDB deposited structures from x-ray crystallography experiments solved to a resolution better than 2.5 Å and which contained both RNA and protein chains. I subdivided the atomic data in the structures such that I had a structure representing each chain of protein from the original structures and all atoms belonging to nucleic acid residues. I used the Dunbrack PISCES server (G. Wang & Dunbrack, 2005) to select the set of highest resolution chains with less than 30% sequence identity to any other chain in the set. The set of chains selected using the sequence identity criteria are listed in Table 2.2. The structures chosen were not filtered based on the structure or sequence of the RNA.

4. Domain composition of the training set

I included RBP structures without respect to annotation with a particular role or binding mode. I included the protein components from the ribosome in the set of candidate sequences. Many of the chains from ribosomal proteins survived the sequence filter because the evolutionary distances of these proteins from most RNA binding proteins are large (T. F. Smith, Lee, Gutell, & Hartman, 2008; Caetano-Anollés, Kim, Mittenthal, & Caetano-Anollés, 2010). Table 2.3 lists the PFAM annotation of the included chains. The most common RNA binding families are the RNA recognition motif (RRM), the K homology domain (KH) and the pumilio-FBF (PUF) domains (Messias & Sattler, 2004). I note that each of these key domains is only represented by one or two examples in the training set. From an information theory perspective, we expect that contacts relevant to high affinity binding of base edges will be represented in the dataset regardless of the domain classification of the proteins in the set.

The structure diversity criteria may be too strict given that small changes in protein sequence are known to switch the target sequence specificity of RBPs. The RRM domain is known to recognize nearly the entire library of tetramers of single stranded RNA (Auweter et al., 2006). A brief survey of the PFAM database reveals 1173 unique sequences for RRM in the human genome. This suggests that subtle variation in structure and sequence is particularly important for specific interactions. We have structural information representing only a few of the biologically relevant sequence variations for each domain. Ensuring that the database contains sufficient representations of all contacts between each amino acid chain and all four canonical RNA bases is important yet probably impossible with the currently available structures.

5. Validating contact composition of the training set

I verified that residue and base composition of the training set did not exhibit sequence composition or structure biases that would not be expected for complexes of RBP with RNA. Since the training set was composed from a diverse set of structures and selected for non-redundant protein sequences, I did not expect biases in the sequence composition of the bound RNA molecules or deviation from the expected sequence bias of an RNA binding protein favoring electro-positive residues.

In order to check for bias in residue to base contacts, I checked the residue-base contact composition of the training set by summing contact counts where a residue C atom was proximal (within 12 Å) to a ribose C1' atom. Figure 2.2 shows contacts of each residue with each of the four base types. Each residue type makes approximately the same number of contacts with each base (Figure 2.3). With lysines, the training set contains more contacts with guanosine than with the other nucleosides. Significantly fewer contacts are made with uridine by arginine, glycine and aspartic acid.

The overall contact count by each residue is summarized in a density plot (Figure 2.4). As might be expected, arginine and lysine make the most base-contacts in the training set. Glycine is surprisingly highly represented. The most under represented residues, cysteine, tryptophan and methionine are likely to provide too few examples to result in a meaningful score value. For these cases less than 25 residue to base contacts are recorded. Molecular contacts not well represented in the training set are likely to be scored incorrectly.

Scores for underrepresented atoms will deviate from the expected profile of score with distance in unexpected ways. The training data is smoothed out by evenly

meting out a small number of pseudo-counts (artificial counts added to the observed counts to correct for rarely observed cases). When the number of pseudo-counts meets or exceeds the number of counts from observations, application of the training function will result in flattened out scoring profile and is likely to be uninformative. The underrepresented residues are not likely to influence base-specificity results, since these residues will be similarly underrepresented in the test set and do not interact preferably with any of the bases. However, with respect to selecting the most favorable residue, these residues may receive artificially optimistic scores that I will describe later as a challenge to applying the all-atom distance-dependent scoring function to RNA specificity calculations.

6. Optimization of model for scoring applications

The features of the Bayesian pairwise distant dependent function that can be altered are the definition of the reference state, the chemical diversity allowed by selecting an atom-typing scheme and the granularity with which the statistical data may be learned. The reference state and the typing scheme are selected considering quantity of training data available. The reference state and typing scheme are not easily optimized. Since we must learn the priors from a limited amount of training data, we must also discretize the distances in order to obtain accurate estimates frequency with respect to distance. The size of these distance bins and the range over which distances may be optimized to maximize the performance of the potential with respect to an external metric.

Bayesian priors from discretized Euclidian distances – When training a Bayesian classifier with a limited set of example data, we need to bin the data such that the relative fraction of counts within each distance interval approximate the underlying statistics. Bayesian inference assumes that data from a small sample of priors may be binned such that they approximate the relative probabilities that would be found if a significantly larger training set were available (Sippl, 1995). While the binned priors help approximate the underlying statistic, the learned probabilities are necessarily discretized. With an infinite data set, we would be able to infer a score that is continuous with respect to distance.

I chose distance bin parameters based on what was effective with specificity tests with the protein-DNA set and on what was consistent with the literature at the time. The size of the bins was chosen to maximize the detail that could be discriminated by the scoring function. Of course I wanted to choose as small of a bin as possible so that I could repack interfaces with the scoring function and accurately place amino-acid side-chains. On the other hand, the bin size is limited by the distance interval over which accurate frequencies could be obtained.

The applications of Bayesian statistics to other biomacromolecules illustrate the relationship between amount and quality of available training data and the optimal distance bin size. In the case of pair-wise distance score for protein structure, there was sufficient protein data to choose a bin size as small as 0.25 Å (Samudrala & Moult, 1998). When the same Bayesian model discussed here was constructed with atom types based on atomic valency and trained on small molecule interactions, there was sufficient training data to choose bin sizes as low as 0.1 Å (Bernard & Samudrala, 2009). The

small bin for the protein structure was allowed by the large amount of training data. The smaller bin size in the small-molecule binding set was also aided by smaller uncertainties in the x-ray crystal structures composing the training set.

In the case of protein-DNA and protein-RNA interfaces, frequencies were only stable with bin sizes between 0.5 and 1.0 Å. Since we expected packing to be a primary use for this scoring function, I chose to error on the side of the smaller 0.5 Å bin size. The selection of the bin size is justified indirectly through the quality of structure predictions.

interaction range – The range of contact distances is chosen to focus on those interaction distances that are likely to be physically meaningful. I selected the high and low cutoffs based on characteristics of the training structures and on ranges over which electrostatic forces are observed in proteins.

The Bayesian classifier learns atom pair probabilities for distances greater than the low cutoff and less than the high cutoff. While the learned scoring entries do not directly correspond with energy, the sum over the interacting pairs roughly corresponds to the free energy (Sippl et al., 1996). The dominant force at large distances would be electrostatics. But the learned probabilities at large-distances will strongly depend on assigned atom type and will depend on typical protein structure.

At short distances the learned atom pair probabilities should resemble electrostatic interactions. The interactions should reflect Coulomb force and van der Waals interactions. But at small distances a repulsive term described by the Lennard-Jones term should dominate (Lennard-Jones & Devonshire, 1937). The correct distance dependent probabilities cannot be learned from the limited training data available for

protein-RNA complexes and with the large bin sizes required for meaningful statistics. However, the intended use of the scoring function for interface packing necessitates learning the repulsive interactions, because steric effects may contribute to specificity.

low cutoff – The low cutoff was chosen to allow for learning the low probability of close interactions and for maintaining correct overall statistics. A typical low cutoff places counts less than 3 Å into a single bin (Samudrala & Moult, 1998). This cutoff is reasonable since we are concentrating on non-covalent interactions with distances exceeding 2 Å and only searching interaction distances between 'heavy' atoms (all atoms excluding hydrogen) that are likely to be further apart due to an interceding hydrogen atom. However, when I was investigating the use of a scoring function with a minimum bin size for side-chain packing, I found that the predicted positions were incompatible with what would be expected for a typical Lennard-Jones repulsive term.

I tried lower cutoffs in order to allow for an empirically learned close energy function that would capture in the scores the effect of the Lennard-Jones repulsive force. By lowering the low cut-off distance below 3 Å, the scoring function is allowed to learn the improbability of atoms at close distances. Eliminating the minimum cutoff improved packing in the case of protein-DNA interactions; thus, I maintained this implementation when training on protein-RNA data. The resulting log-odd scores for sub 3 Å distances were unfavorable, but, given the limited amount of available training data, they could not learn the exponential increase in energy observed at these distances. The learned scores were nonetheless sufficient to prevent clashes between protein and nucleic acid in the repacking tests.

high cutoff **–** The choice of the high cutoff loosely corresponds to the maximum distance over which atoms within a protein-RNA complex could interact through higher order electrostatic effects. Previous work in the Varani laboratory investigated high cutoffs between 10 and 20 Å with protein-DNA interactions (Robertson & Varani, 2007) and 6, 10 and 12 Å cutoffs with protein-RNA interactions (Zheng et al., 2007). In both cases, a cutoff of 10 Å demonstrated the best performance in the recovery of native docking structures and the best correlation between scores and binding energies free energies.

A recent study of a similar distance-dependent force field for protein interactions suggests that the first level of unshielded contacts are most important for reproducing features of physical potentials with a distance dependent potential (Ferrada & Melo, 2009). The observation that direct contacts are most important helps explain why the performance of the scoring functions varies only slightly as the cutoff in increased beyond 6 Å. The importance of the molecular interface also suggests that a quality training set is one that well represents direct contacts.

summary of Bayesian model parameters – The basic parameters chosen for building the RBP scoring matrices that were used in specificity tests can be summarized as follows. I chose to use the empirical reference state based on atom type agnostic counts in the training set. The bins were allocated from 0.5 to 10 Å in steps of 0.5 Å. The training set consisted of the protein chains listed in Table 2.2 with all accompanying RNA contained in the PDB structure from which the chain was selected. The performance of the scoring function would likely improve if these model parameters were further optimized based on more benchmarks of known structure and interaction

energy. The performance of the scoring function will likely be significantly improved when the size of the protein-RNA training set approaches the size of that which is available for training protein-DNA interactions.

D. Implementation

Initial work with docking decoy discrimination was implemented as an independent software package built on top of an open source molecular structure library for reading and modifying protein and RNA structures that was developed in-house. I extended this program for performing specificity tests involving the systematic exploration of all point mutations at the interfaces of bound protein and RNA complexes. When it became necessary to explore the inclusion of additional statistical terms and to perform direct comparisons with a typical Rosetta scoring function, I re-implemented the statistical scoring function within the Rosetta molecular modeling suite (Leaver-Fay et al., 2011) and performed the remainder of my calculations within that environment.

Our original, in-house approach to molecular specificity involved the development of a library for reading, writing and editing molecular structure and the development of applications that performed the specific computational tests. The code comprising the open-source package, the Biological Toolkit (BTk), is available online as source code (https://sourceforge.net/projects/btk/). BTk is a C++ library that reads PDB version 2 flat files and places them in C++ standard library compliant iterable containers. The smallest container pertains to a protein residue and nucleic acid base by extending a common base class that defines a container of atoms referred to as a monomer. Residue and nucleic acid classes extend the monomer class to add

appropriate access mechanisms for modifying any of the dihedral angles required for modifying amino acid or nucleic acid structures.

The RAINIER applications are implemented in C++ to perform scoring tasks with molecules internally represented using the BTk library. The source code for RAINIER is distributed as a separate open source project with source code available online (https://sourceforge.net/projects/rainier/). The original RAINIER application used for scoring docking decoys was a simple scorer application for the pairwise distance potentials. This original program used in the docking decoy manuscripts (Robertson & Varani, 2007; Zheng et al., 2007) simply employed the BTk library to iterate over pairs of protein-RNA interactions. My subsequent work involved performing point mutations and rethreading experiments at the protein-RNA interface. I extended the RAINIER package to perform rethreading replacing any specified protein residue or RNA-base with a desired replacement sequence. Additionally, since amino acid side-chains vary substantially in size, chemical properties and are highly flexible about their side-chain dihedral angles, it was necessary to implement side-chain repacking as well.

The RAINIER applications perform side-chain repacking of interface protein residues using a combined statistical potential with terms for protein intra-molecular energy and the intermolecular protein-RNA term described above. The protein intra-molecular term was implemented using score tables from mathematically equivalent work with intra-molecular protein energies performed by Samudrala and Moult (1998). For all residues near a substituted base or residue, possible side-chains dihedral angles were drawn either from the Dunbrack backbone-dependent or backbone-independent rotamer libraries (Dunbrack, 2002; Dunbrack & Karplus, 1993). Since in our application

no backbone moves were allowed and many residue side-chains remained unmoved, all interaction-energies could be pre-computed and cached in order to save computational time. Repacking was performed using a Monte Carlo approach with the temperature parameter chosen by pre-sampling the changes in score over a few thousand side-chain move steps. The side-chain dihedral angles chosen as a result of this procedure resembled those obtained using more complex scoring functions such as Rosetta (Havranek, Duarte, & Baker, 2004). The side-chain repacking procedure allows for a more physically realistic approximation of interaction energies between proteins and their bound RNA molecules.

In order to improve the performance of the statistical potential described in the previous section for the specific problem of specificity of binding at the protein-RNA interface, we considered augmenting the potential by adding additional statistical or physical terms such as a statistical term for protein-RNA hydrogen bonds (Y. Chen, Kortemme, Robertson, Baker, & Varani, 2004) or exploring the use of statistical terms derived for larger, reusable fragments of protein and RNA residues similar to that described by Lu, Dousis and Ma . However, a statistical scoring function augmented by linearly combining the distance-dependent score and additional physical or statistical terms begins to resemble the scoring function implemented in the Rosetta molecular modeling package (Leaver-Fay et al., 2011). Thus, in order to perform direct comparisons with Rosetta and to facilitate the exploration of mixed scoring, I re-implemented the statistical scoring function within Rosetta.

The Rosetta molecular modeling package includes dozens of statistical and physical scoring terms, routines for modifying protein structures, and algorithms for

searching structure-space (Leaver-Fay et al., 2011). I implemented RNA-base mutations within Rosetta and wrote programs using the Rosetta library to recreate my scoring and repacking tests using Rosetta modules. Additionally, in order to perform direct comparisons between specificity tests using the pure statistical approach, a common set of weighted scoring terms from Rosetta and additional combinations of terms, I added a scoring module that would implement my statistical scores using the tables of statistical scores from our original scoring function. The approach of implementing the exploration of structure within Rosetta had the additional benefit of allowing the rapid testing of new features introduced in Rosetta for transcription factor design with the RBP test sets.

E. Challenges of Applications to RBP Specificity

The prediction of the preference for a native base or residue at an interface location is much more difficult than the problem of selecting a native complex from a set of docking decoys. Structure-based specificity prediction is premised on the ability to search local physical space and to reproduce the intermolecular energies for contacts with the bases or residues being mutated. With the previous docking-decoy discrimination tests and rethreading tests (Robertson & Varani, 2007; Zheng et al., 2007), accurately reproducing each intermolecular interaction was less important. In this section, I look at several indicators of the sufficiency of the training set and suggest how insufficient training data affects predictions. I also discuss how fair tests are performed when the test and training sets are structurally similar.

1. Performance on previous RBP-RNA computational tests

The requirements for successfully discriminating a native structure from docking decoys and for full interface rethreading are different from those required for point substitution tests. The all atom distance-dependent scoring function was previously tested with a few decoy discrimination tests with extensive but near native changes to protein-RNA complexes (Zheng et al., 2007). These included docking decoy tests, MD generated decoys and rethreading calculations. What these calculations have in common is that the decoys form a spectrum of non-native structures where the median structure is significantly different from the native. In these tests, success is as obtained when the native structure is among the small fraction of structures very close to the native.

While the good z-scores for docking decoy discrimination from near native structures and good ranking received by the correct sequence suggest that the scoring function would be sensitive to mutations of a single base (Zheng et al., 2007), the results do not guarantee the precise scoring of monomers. Complete rethreading tests are significantly less challenging than position substitution tests. In complete rethreading, all 4^n possible sequences (where *n* is the number of interface bases) were substituted onto the RNA backbone preserving the original glycosidic bond angle. The distance-dependent potential successfully ranked the native structure in the top decile of decoy structures for more than half of the structures tested (Zheng et al., 2007). If the training data has learned favorable scores for key residue-base interactions, sequences recapitulating those interactions will be selected. Correctly predicting just a few base positions ensures that the correct sequence will score significantly better than average.

In the alternative approach described in the next chapter, where each base is substituted independently, each structure is near native.

Obtaining an accurate score or relative energy for a set of base substitutions at a single base position requires that the scoring function be correct for the majority of the atom pairs for all candidate base substitutions. The all-atom scores are dominated by the scores of directly contacting atoms (Ferrada & Melo, 2009). This observation suggests that only a fraction of training set was selected for the uniqueness of protein sequence consisted of protein-RNA structures where the base edges make direct contact with the protein. This further suggests that if we are training a Bayesian scoring function for these interactions, using training data where direct interactions between the amino acid side-chains and nucleic acid bases are well represented is important. Furthermore, we observe that score versus distance profiles of interactions such as hydrogen bonds where direct side-chain to base interactions may confer specificity are particularly informative. We may speculate on the effect of undertraining on specificity calculations and look for evidence of this insufficiency in the scoring functions.

Under the conditions where most base contacts in the training set are shielded by one or more layers of atoms and one of the bases is underrepresented in the training set, we may expect a better score for atoms in the underrepresented base. The training data is necessarily a subset of the theoretical complete set where the frequencies of specific contacts deviate from their true frequencies. If the learned priors over-represent the frequency, the scores will be better, more 'optimistic', than they should be. We need to consider how the training set composition influences contacts between

proteins and base edges in general and we need to consider what happens when contacts between specific atom pairs are sparsely represented in the training set.

2. Indicators of training set sufficiency

The training set was purposefully selected to represent all solved RBP non-redundantly. Since much of the training data is from double-stranded regions of RNA or from regions containing higher order tertiary contact sequences, the ratio of contacts in the 5 to 10 Å bins to those in bins <5 Å will be relatively high. I expect the scores for protein side-chain contacts with the base edge to be overly pessimistic. The number of close contacts between side-chains and bases are under-represented with respect to what they would be if we knew the structures of all RBPS in complex with RNA. Thus, close contacts will receive a worse score than they should.

However, since we are using this potential for finding a relative score for a position where all bases are tried, we would expect contacts to any base edge to be similarly penalized by the prevalence of more structured RNAs in the training set. An imbalance in the types of RNA structures represented may affect binding preference in a docking application. For discriminating between different bases binding to the protein in similar orientations, the prevalence of RNA with high tertiary structure in the training set should not affect specificity predictions significantly.

Insufficient data for resolving the features of the pairwise interaction over the desired range would affect specificity. If the amount of training data for a specific atom-pair is small, the contact counts will be dominated by the pseudo-counts and will result in a Bayesian score that is close to zero across all distance bins. When all interactions

are sparsely represented in a distance region, the score of that region will change more as more training data is added.

Training data interaction with base edges may not be sufficiently represented and this may affect specificity calculations as well. Since uracil is less represented in the training set than the other bases (Figure 2.2), we may expect that the predicted score landscape with respect to distance may be less well determine for uracil. Indeed the scoring function is flatter overall for uracil with respect to distance than for those involving the other three bases (Figure 2.5). If near contacts are under-represented in general with respect to more distant contact, the score for close contacts will appear less favorable. However, if the ratio of pseudo-counts to measured counts is greater for one base, it will appear less unfavorable than the other three bases. As it happens, when calculating relative specificity for the bases, the underrepresented uracil is preferred.

3. Training for specificity

The sufficiency of the training set may also be considered in terms of interaction types such as hydrogen bonding. While the connection between the prevalence of certain interaction types in the training set does not need to match that in an expected test set, the test set should be a sampling of what is possible in nature as well as the frequencies of the interaction types in nature.

Several works have sought to find the physical interactions that confer specificity. Concentrating on interactions that are more common at a protein interface with nucleic acids, it has been speculated that certain patterns of interactions types may help select one base over another. In addition to positive charges that provide increasing affinity for the nucleic acid phosphate backbone, hydrogen bonds and

stacking interactions between the base rings and aromatic or positive residue side-chains are important for recognizing specific bases as well (Auweter et al., 2006). Hydrogen bonds are most likely to differ between the residues, since the edges of the four bases expose different chemical groups at a number of positions. This was the impetus for previous work in developing a protein-RNA specific hydrogen binding term in conjunction with Rosetta (Y. Chen et al., 2004). We would like a statistical approach to correctly represent important hydrogen bond interactions. In the next chapter (Chapter 3.F below), I compare how the all-atom distance dependent scoring function and other scoring functions perform with the recovery of key interactions.

First, I verified if hydrogen bonds likely to be important to binding are represented in the training set. Using the program `hbplus` (McDonald & Thornton, 1994), I counted the interface hydrogen bonds in the training set; there are in total 303 hydrogen bonds in the training set. However, the number atom types in the all-atom model probably assign different types to some atoms that are chemically identical, so there are relatively few examples of each hydrogen bond for individual protein and RNA atom pairs. The most common bonds in the training set are LYS O4 to U NZ and ARG NH1/NH2 to C O2 with 8 and 10 bonds, respectively, if the arginine symmetry is accounted for. In terms of hydrogen bonds represented as a fraction of total contacts in the training set, the training set is poorer in hydrogen bond contacts than the test set I use in the next chapter. This would be expected to have an effect on calculated affinity, but should not necessarily impede the utility for specificity calculations.

Specificity calculations would be affected if hydrogen bonds to a specific base were better predicted by the scoring function. Above we noticed that plots of potential

hydrogen bonding pairs with distance showed that sub 3 Å scores were more favorable with uracil than with the other bases. This effect could be caused by differences in representation or by unique characteristics of hydrogen bonds with uracil. In terms of raw numbers, the training set does contain more hydrogen bond with G and U (106 and 81) than with A and C (59 and 57). So in terms of representation and base size, we may expect that hydrogen bonds made by G are favored over those with A and similarly for U over C. However, there could be other subtle ways in which a bias for uracil could arise in the context of the Bayesian classifier.

The demands placed upon the empirical scoring function when performing specificity prediction differ substantially from those for performing docking or large-scale rethreading exercises. In order to correctly rank bases based on the score of their interactions with the surrounding protein, the scoring function must be trained on a sufficient amount of representative data. The relatively small number of structures solved to date and the resulting redundancy between possible test sets and training data are a significant obstacle to the success of the pairwise distance dependent approach at this time.

4. Fair scoring matrices

A central problem in constructing a statistical scoring function for protein-RNA interactions concerns the relationship between the training set and the test set. Two approaches can be used to evaluate the effect of overtraining for a specific test: a jack-knife approach to randomly removing subsets of the training data and a directed approach where specific structures are removed from the training set based on similarities with the scoring function. These statistical tests demonstrate the dramatic

performance difference in the same scoring function tested in the protein-DNA case and that observed in the protein-RNA case. The protein-RNA tests are significantly more sensitive to alterations to the training data.

Docking interaction results with protein-DNA complexes continued to perform well with standard jack-knife tests. In docking tests, removing up to 5% of the training structures had little effect on the performance of the all-atom scoring function (Robertson & Varani, 2007). Removing up to 20% of training data did affect performance, but maintained good discrimination for the native structure from most near native decoy structures. Reduced atom potentials where atom types are generalized such that they are reused in similar chemical contexts in other positions and residues were relatively unperturbed by the removal of up to 20% of the scoring function, although the overall performance was significantly worse. The removal of individual structures affected z-score of the native structure but not its ranking in decoy discrimination tests. These results suggest that the training set was sufficiently large for the protein-DNA interactions, but that performance still varied significantly with training set size.

In the full structure docking and rethreading tests, the protein-RNA score was not subjected to the same level of statistical rigor, but signs of overtraining can be seen when these tests are repeated with 'fair' scoring matrices. The structures available for structure-based prediction of binding energy and specificity must be drawn from the same set. In order to perform rigorous tests of the scoring function, I limited the presence of training data from the structure being scored or from close structural neighbors of the test structure. I thus computed 'fair' scoring matrices for all available

RNA structures by compiling the sequence of the structures used in the training set into a BLAST database. Using the sequences of all solved protein-RNA structures in the PDB, the counts for similar structures (determined by BLAST expectation value less than 10^{-6} in the database of training structures) were removed from the count data using difference matrices and a 'fair' scoring matrix was generated for that complex.

Sequence specificity tests with the complete training set were significantly better than those performed with fair scoring matrices. The difference is seen in base recovery tests where each possible base is substituted into a protein contacting RNA position and the predicted base is that with the lowest score against the protein highlights. Figure 2.6 shows that we observe dramatically better base recovery when the entire training set is used than when fair scoring matrices are employed.

The contrast between the training set-sensitive performance of the protein-RNA scoring function with the relatively robust performance of a nearly identical term with the protein-DNA interactions, imply that a certain quantity of training data is required to make the scoring function perform similarly to physical potentials. The performance of the protein-RNA scoring function with fair scoring matrices in RNA base specificity tests at the protein-RNA interface will be examined with larger training sets in the following chapter.

F． Statistical RNA Intra-molecular Term

RNA intra-molecular energies play a substantially more significant role in RNA base specificity than for DNA base-pair specificity. The assumption of independence between neighbors is valid for most short interactions between protein and single stranded RNA. However, many of the structures I tested have significant contacts

between adjacent RNA bases, especially at the edge of a recognition region. I implemented an all-atom distance-dependent statistical pairwise scoring function for RNA intra-molecular energy formally identical to those described for protein-DNA and protein-RNA interactions described above.

The training of the scoring function is similar to that described above but additional training data are available in this case. The scoring function parameters were similar to those used with protein-DNA. Atom types were again defined as all defined PDB residue name and atom name pairs. The scores were constructed using the naïve Bayes classifier previously described before using the empirical reference state. Distances were calculate for atom inter-residue atom pairs and binned in 1 Å bins with a low cutoff of 3 Å and high cutoffs between 6 and 12 Å. Additional scoring matrices were constructed with selective filters applied to the distance data allowing the removal of specified contact data points.

The matrices were initially trained on the relatively small set of RNA structures solved with x-ray crystallography to a resolution better than 2.5 Å. However, the structure of both the large and small subunits of the ribosome had been recently solved by x-ray crystallography to a resolution better than 3 Å. Since the ribosomal sequence is separated from other RNA structures by a very large evolutionary distance, we could assume that the training data was unrelated to any potential test set.

The RNA intra-molecular term was used to score MD decoys of 15 RNA structures with substantial tertiary contacts. The test set in Table 2.2 contains mostly RNA structures that were solved for their functional interactions with other molecules. These structures include tRNA components and self-splicing introns. Decoys were

generated using short MD simulations of a few picoseconds in length using the MOE molecular editing and simulation tools. The simulations were run such that structures from intermediate time-steps could be taken as a set of near native decoys. I scored the decoys using the RNA intra-molecular scoring function to obtain plots of score with RMSD from the native structure. Figure 2.7A shows the score to RMSD relationship for the self-splicing intron 1kxk. The results of these plots can be summarized with a Z-score by assuming a Gaussian distribution with score. For most tests, Figure 2.7B summarizes results with the ribosome-trained scores with distance bins in the 3 to 10 Å range. The average Z-score is better than average by 3 standard deviations. These results were better than those seen in similar tests with the protein-RNA interface.

Results using the scoring matrices trained on smaller RNA structures and those from the selective filter tests suggested that we could be observing an artifact. Discrimination of native from non-native structures was robust in a manner not observed with the inter-molecular pairwise distance scoring functions. Structure discrimination was not disrupted in the non-ribosomal training set when up to half of the training data were removed. With the ribosomal training set, there was no overlap between the training set and the structures tested (Table 2.4). Overtraining for a specific structure motif could not explain the robust discrimination of native from near native structures examined. In order to discover which interactions were contributing the most to the structure discrimination, scores pertaining to highly favorable interactions were eliminated during scoring. Using secondary structure annotation to remove either Watson-Crick base pairs or adjacent bases in the helical regions did not disrupt discrimination of the native structure (Figure 2.8). The stacking interactions in

helical structures were more important in this test. The absence of a strong effect from the removed interaction types contradicts prior knowledge and suggests a problem with the scoring function or the decoys used in the near-native tests.

MD simulations with RNA are difficult to parameterize and may not yield a high quality decoy set of near native structures. Decoy structures created by short MD runs using the MOE program had relatively small (0 to 8 Å RMSD) deviations from native structures. However if the decoy generating MD simulation did not handle intra-RNA scores correctly, the resulting RNA structure decoys may have been an easy test containing unusual structural elements. Recent work using Rosetta to build RNA tertiary structure from fragments has been reported (Das & Baker, 2007; Das, Karanicolas, & Baker, 2010). Sampling from structures generated during the application of the structure-based work might provide for a better test of the performance of the statistical RNA scoring function.

In the end, the purely statistical all-atom intra-molecular term was not included in later specificity calculations. Given questions about the results, I did not invest the significant time required to implement the full score function. A complete scoring function would include a linear combination of protein-RNA scores, protein intra-molecular scores and RNA intra-molecular scores within a RANIER application. At the time of this work, we also lacked a good test set for properly weighting these terms. The implementation of the complete scoring function with all three components within the RAINIER framework would have been very time consuming and ultimately uncertain. However, the re-implementation of the scoring function within Rosetta allows scoring

with an arbitrary number of weighted Bayesian terms. It would now be useful to try the full statistical protein-RNA scoring function including both intra-molecular terms.

G. Summary

Scoring functions employing a naïve Bayes approach to evaluating candidate structures based on learned pairwise distances have been successful in protein structure prediction. These statistical scoring functions have been applied to the protein-folding problem and, in the Varani laboratory, to docking and rethreading applications with protein-DNA and protein-RNA complexes. In this chapter, I described an implementation of the all-atom distance-dependent scoring function for use in predicting specificity of RNA binding proteins for specific RNA sequences.

In this chapter, I have described modifications to the scoring function, selection of a training set, and the implementation of the scoring function for specificity tests. I constructed a new training set to improve the chance of success in predicting the energy of specific mutations to RNA and protein sequence. I extended a program developed in the Varani laboratory for substituting specific base and residue positions at a protein-RNA interface and for calculating changes in score that predict changes in binding energy. The smaller training set and greater structural complexity of the RNA which RNA-binding proteins evolved to bind complicated the application of the empirical scoring approach to this problem.

I am confident that the scoring function could be improved through further changes to training set construction and through optimization of model parameters. However, in order to achieve a performance similar to that seen with protein structure prediction or with DNA binding specificity tests, we fundamentally need more training

data or a different statistical approach that better captures critical interactions between proteins and RNA.

I have also applied the pairwise distance potential to evaluate the structure of unbound RNA structures. RNA forms complicated tertiary structures and can serve a functional role similar to that of proteins in cells. Unlike DNA, RNA structure often adopts non-helical structures and can be understood through their intra-molecular energies. Because of these features, the structure of the RNA may contribute to sequence constraints and specificity when proteins bind to RNA. A scoring function for RNA intra-molecular energies could be used in conjunction with the protein-RNA inter-molecular term to better understand sequence specificity.

This chapter described the theory and implementation of an empirical scoring function for use in the structure based determination of the specificity of RNA binding proteins. The problem of predicting these interactions is of great interest because of advances in our understanding of the role of these proteins in the post-transcriptional regulation of the temporal and spatial expression of functional proteins in cells. The importance and scope of the problem of RBP specificity warrants the development of a computational approach. In the next chapter (Chapter 3), I benchmark the performance of pairwise distance approaches and a Rosetta scoring function in specificity prediction with proteins for which the specificity is known from experimental techniques.

Figures and Tables

Table 2.1: Most prevalent RNA binding proteins from Pfam annotation of known sequences.
Cells employ a relatively small number of domains for the sequence specific recognition of RNA. Pfam uses seed sequences from known structures to train a hidden Markov model for a family and to identify other sequences that fit the model (Finn et al., 2008). The table shows the number of sequences classified in the same Pfam family as known RNA binding proteins (RBPs) in all organisms and in the human genome. This classification provides an estimate for the total number of RBPs whose preferred RNA target sequence needs to be identified. The most abundant domains with RNA specific binding are highlighted in bold. Domains that are employed in recognizing both RNA and DNA sequences are italicized. The last column shows the number of unique proteins for which the structure of an RNA binding domain has been solved wither with (bound) or without (unbound) RNA.

Domain	**name**	**Pfam ID**	**number of identified sequences**	**number of sequences in human genome**	**number unique bound/total structures**
RRM_1	RNA recognition motif 1	PF00076	27590	1173	21/106
RRM_2	RNA recognition motif 2	PF04059	246	0	
RRM_3	RNA recognition motif	PF08777	129	7	
KH_1	K-homology domain	PF00013	9244	359	6/22
KH_2	K-homology domain	PF07650	4644	2	
PUF	Pumilio-family RNA binding repeat	PF00806	3565	104	4/6
PAZ	Piwi Argonaut and Zwille	PF02170	842	18	2/4
zf-CCCH	Zinc finger C-x8-C-x5-C-x3-H type	PF00642	3549	133	
zf-RanBP	Zn-finger in Ran binding protein	PF00641	1533	89	

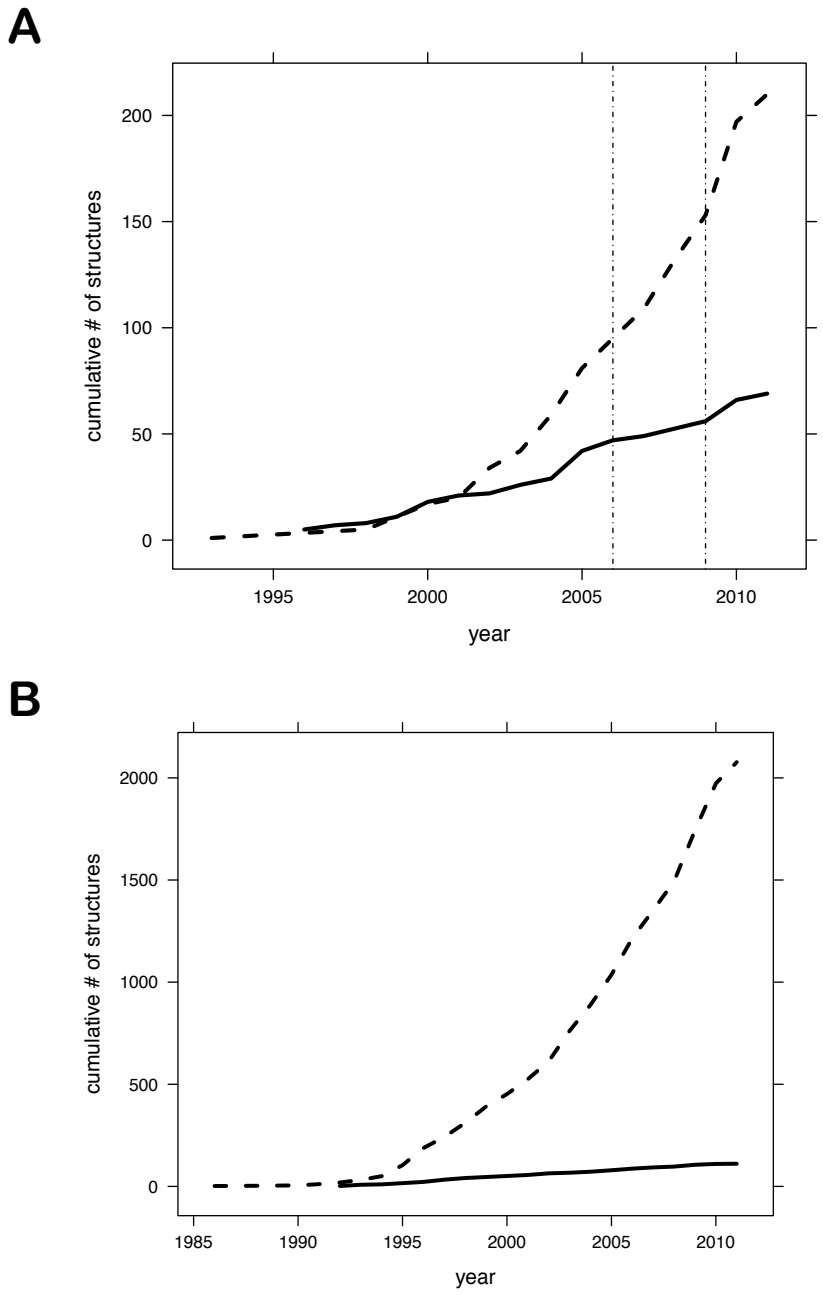


Figure 2.1: Solved structures of protein-nucleic acid complexes deposited in the PDB by year.
A: Number of structures of protein-RNA complexes in the PDB with year solved by x-ray crystallography (dashed) or NMR (solid). Vertical lines indicate years when statistical scoring functions were created. **B**: number of protein-DNA structures with year solved by x-ray crystallography (dashed) or NMR (solid).

Table 2.2: Structure of origin for chains included in the training set for the all-atom scoring function. The table identifies the chain by the PDB ID of the source structure and the letter chain ID from the structure file. The training set chains were selected from the complete set of RNA binding proteins containing both protein and RNA from the PDB accessed in July 2009 (Berman et al., 2000) and culled to contain chains with only 25% identity using the Dunbrack PISCES server (G. Wang & Dunbrack, 2005)3) . The training set structures consisted of the identified unique chain and all RNA contained in the PDB file. The sequences of the protein chains were found and included in a local BLAST database so that distances from proteins that are sequence similar to the test protein can be removed from the training set.

PDB ID	chains	protein name
1b2m	A	ribonuclease T1
1c0a	A	aspartyl trna synthetase
1dfu	P	paraneoplastic encephalomyelitis antigen HUD
1fxl	A	trp rna-binding attenuation protein (TRAP)
1gtf	A	matrix protein VP40
1h2c	A	signal recognition particle protein
1hq1	A	tyrosyl-tRNA synthetase
1j1u	A	Restrictocin
1jbs	A	tRNA pseudouridine synthase B
1k8w	A	signal recognition particle 19 kDa protein
1lng	A	pumilio 1
1m8x	A	minor core protein lambda 3
1n35	A	glutamyl-trna synthetase
1n78	A	nuclear factor nfκB p105 subunit
1ooa	A	protein (glutaminyl-tRNA synthetase)
1qtq	A	isoleucyl-trna synthetase
1qu2	A	core protein p19
1r9f	A	50s ribosomal protein L2p
1s72	ABCDEGHIJKLMNPQRSTUVWXYZ	50s ribosomal protein L7ae
1sds	A	trna nucleotidyltransferase
1tfw	A	cysteinyl-trna synthetase
1u0b	B	P2 protein
1uvj	A	hut operon positive regulatory protein
1wpu	A	60-kDa ss-a/ro ribonucleoprotein
1yvp	A	RNA binding domain of rho transcription termination factor
2a8v	A	neuro-oncological ventral antigen 1
2anr	A	transcription elongation protein nusa
2asb	A	tRNA adenosine deaminase
2b3j	A	protein AF1318
2bgg	A	23s rRNA (uracil-5-)-methyltransferase ruma
2bh2	A	MS2 coat protein
2bu1	A	protein VTS1
2f8k	A	double-stranded RNA-specific adenosine deaminase
2gxb	A	probable trna pseudouridine synthase B
2hvy	B	50s ribosomal protein L1
2hw8	A	ribosomal large subunit pseudouridine synthase A
2i82	A	ATP-dependent RNA helicase ddx48

Table 2.2 continued.

2j0s	CT	exosome complex exonuclease 2
2jea	I	serine protease subunit NS3
2jlv	A	ribonuclease III
2nug	A	selenocysteine-specific elongation factor
2pjp	A	probable exosome complex exonuclease 1
2po1	AB	poly(A) polymerase
2q66	A	coat protein
2qux	A	lupus lA protein
2r8s	H	synthetic antibodies
2vnu	D	RRP44
2voo	A	16s rRNA
2vqe	BCDEFGHIJKLMNOPQRST	non-structural protein 1
2zko	A	RNA dependent RNA polymerase
3bso	A	tRNA (uracil-5-)-methyltransferase
3bt7	A	muscleblind-like protein 1
3d2s	A	U1 small nuclear ribonucleoprotein A
3egz	A	ATP-dependent RNA helicase dhx58
3eqt	A	CAP-specific mRNA (nucleoside-2'-o-)-methyltransferase
3er9	AB	ATP-dependent RNA helicase DDX19b
3fht	A	tRNA delta(2)-isopentenylpyrophosphate transferase
3foz	A	probable tRNA pseudouridine synthase B
3hax	C	pseudouridine synthase cbf5
3hjw	B	ribonucleoprotein pseudouridine synthase

Table 2.3: List of protein chains included in the data set with their associated Pfam family.
The table suggests that the selection based on sequence identity yielded a training set with diverse tertiary structures and evolutionary origins. The training contains many of the subunits of the ribosome that are unique in terms of both sequence and fold.

PDB ID	chain	pfam accession	pfam ID
1qu2	A	PF08264.7	Anticodon_1
2j0s	T	PF09405.4	Btz
2r8s	H	PF07654.9	C1-set
1u0b	B	PF09190.5	DALR_2
2b3j	A	PF00383.1	dCMP_cyt_deam_1
2nug	A	PF00035.19	Dsrm
2zko	A	PF00600.13	Flu_NS1
2hvy	B	PF04410.8	Gar1
2jlv	A	PF00271.25	Helicase_C
3fht	A	PF00271.25	Helicase_C
1wpu	A	PF09021.5	HutP
3foz	A	PF01715.11	IPPT
2anr	A	PF00013.23	KH_1
2vqe	C	PF07650.11	KH_2
1s72	T	PF00467.23	KOW
2voo	A	PF05383.11	La
2bu1	A	PF01819.11	Levi_coat
2j0s	C	PF02792.8	Mago_nashi
3hax	C	PF04135.6	Nop10p
3hjw	B	PF04135.6	Nop10p
1tfw	A	PF01909.17	NTP_transf_2
2q66	A	PF04926.9	PAP_RNA-bind
3er9	A	PF01358.12	PARP_regulatory
2qux	A	PF09063.4	Phage_coat
2bgg	A	PF02171.11	Piwi
3er9	B	PF12630.1	Pox_polyA_pol_N
2i82	A	PF00849.16	PseudoU_synth_2
1m8x	A	PF00806.13	PUF
1uvj	A	PF00680.14	RdRP_1
3bso	A	PF00680.14	RdRP_1
1n35	A	PF07925.5	RdRP_5
1ooa	A	PF00554.16	RHD
2a8v	A	PF07498.6	Rho_N
1b2m	A	PF00545.14	Ribonuclease
1jbs	A	PF00545.14	Ribonuclease
2hw8	A	PF00687.15	Ribosomal_L1

Table 2.3 continued.

PDB ID	chain	pfam accession	pfam ID
1s72	I	PF00298.13	Ribosomal_L11
1s72	J	PF00572.12	Ribosomal_L13
1s72	K	PF00238.13	Ribosomal_L14
1s72	M	PF00827.11	Ribosomal_L15e
1s72	H	PF00252.12	Ribosomal_L16
1s72	L	PF00828.13	Ribosomal_L18e
1s72	N	PF00861.16	Ribosomal_L18p
1s72	P	PF01280.14	Ribosomal_L19e
1s72	A	PF00181.17	Ribosomal_L2
1s72	Q	PF01157.12	Ribosomal_L21e
1s72	R	PF00237.13	Ribosomal_L22
1s72	S	PF00276.14	Ribosomal_L23
1s72	U	PF01246.14	Ribosomal_L24e
1dfu	P	PF01386.13	Ribosomal_L25p
1s72	V	PF00831.17	Ribosomal_L29
1s72	B	PF00297.16	Ribosomal_L3
1s72	W	PF00327.14	Ribosomal_L30
1s72	X	PF01198.13	Ribosomal_L31e
1s72	Y	PF01655.12	Ribosomal_L32e
1s72	Z	PF01780.13	Ribosomal_L37ae
1s72	C	PF00573.16	Ribosomal_L4
1s72	D	PF00281.13	Ribosomal_L5
1s72	E	PF00347.17	Ribosomal_L6
1s72	E	PF00347.17	Ribosomal_L6
1sds	A	PF01248.20	Ribosomal_L7Ae
2vqe	J	PF00338.16	Ribosomal_S10
2vqe	K	PF00411.13	Ribosomal_S11
2vqe	L	PF00164.19	Ribosomal_S12
2vqe	M	PF00416.16	Ribosomal_S13
2vqe	N	PF00253.15	Ribosomal_S14
2vqe	O	PF00312.16	Ribosomal_S15
2vqe	P	PF00886.13	Ribosomal_S16
2vqe	Q	PF00366.14	Ribosomal_S17
2vqe	R	PF01084.14	Ribosomal_S18
2vqe	S	PF00203.15	Ribosomal_S19
2vqe	B	PF00318.14	Ribosomal_S2
2vqe	T	PF01649.12	Ribosomal_S20p
2vqe	D	PF00163.13	Ribosomal_S4
2vqe	E	PF00333.14	Ribosomal_S5
2vqe	F	PF01250.11	Ribosomal_S6

Table 2.3 continued

PDB ID	chain	pfam accession	pfam ID
2vqe	G	PF00177.15	Ribosomal_S7
2vqe	H	PF00410.13	Ribosomal_S8
2vqe	I	PF00380.13	Ribosomal_S9
3eqt	A	PF11648.2	RIG-I_C-RD
2po1	A	PF03725.9	RNase_PH_C
2po1	B	PF03725.9	RNase_PH_C
2vnu	D	PF00773.13	RNB
1fxl	A	PF00076.16	RRM_1
3egz	A	PF00076.16	RRM_1
2jea	I	PF00575.17	S1
2f8k	A	PF07647.11	SAM_2
2pjp	A	PF09107.5	SelB-wing_3
1hq1	A	PF02978.13	SRP_SPB
1lng	A	PF01922.11	SRP19
1r9f	A	PF03220.7	Tombus_P19
2bh2	A	PF01938.14	TRAM
3bt7	A	PF05958.5	tRNA_U5-meth_tr
1j1u	A	PF00579.19	tRNA-synt_1b
1n78	A	PF00749.15	tRNA-synt_1c
1qtq	A	PF03950.12	tRNA-synt_1c_C
1c0a	A	PF00152.14	tRNA-synt_2
1yvp	A	PF05731.5	TROVE
1gtf	A	PF02081.9	TrpBP
1k8w	A	PF01509.1	TruB_N
1h2c	A	PF07447.6	VP40
2gxb	A	PF02295.11	z-alpha

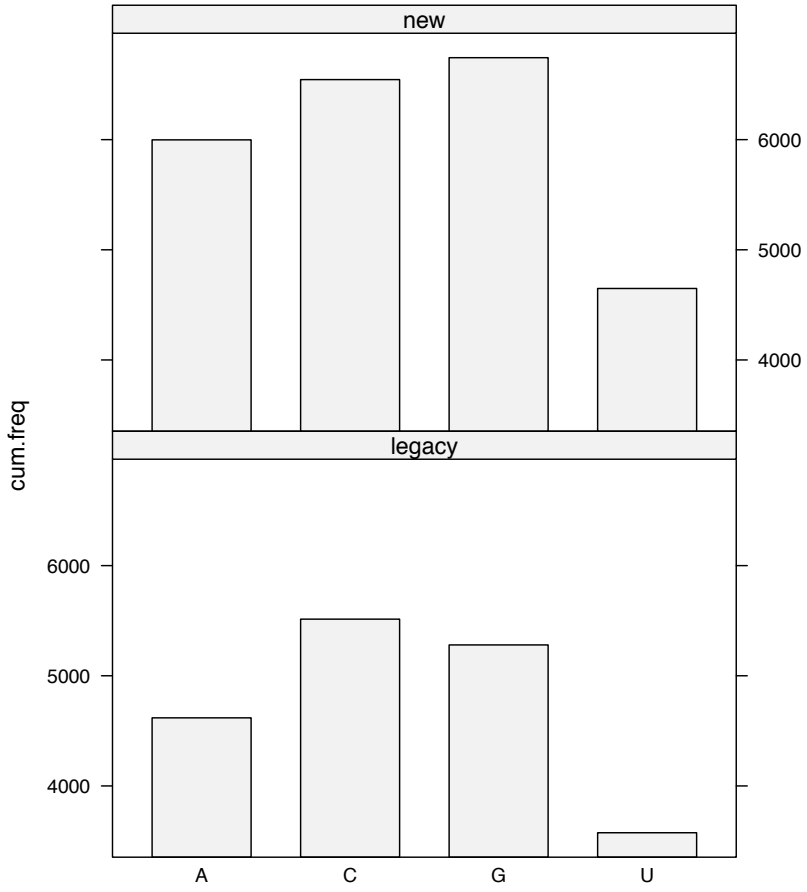


Figure 2.2: Number of side-chain nucleobase contacts in training set by base type.
The count of contacts between a base atom indicating the relative position of the glycosidic bond (C1′) and the protein residue atom marking the beginning of side-chain dihedral (χ) angles (C) occurring within 10 Å of each other is a good metric indicating how well represented a base or side-chain is in the training set. In contrast to a pure base count, a high contact count suggests that many interactions occurred between those bases and the protein in a recognition region of the protein. The bar chart contrasts contacts represented in training the scoring function for the RNA docking decoy discrimination paper (Zheng et al., 2007) (**legacy**) and those represented in my new training set (**new**). There are many more contacts in the new training set and uracil is less severely underrepresented, although still less well represented than the other nucleotides.

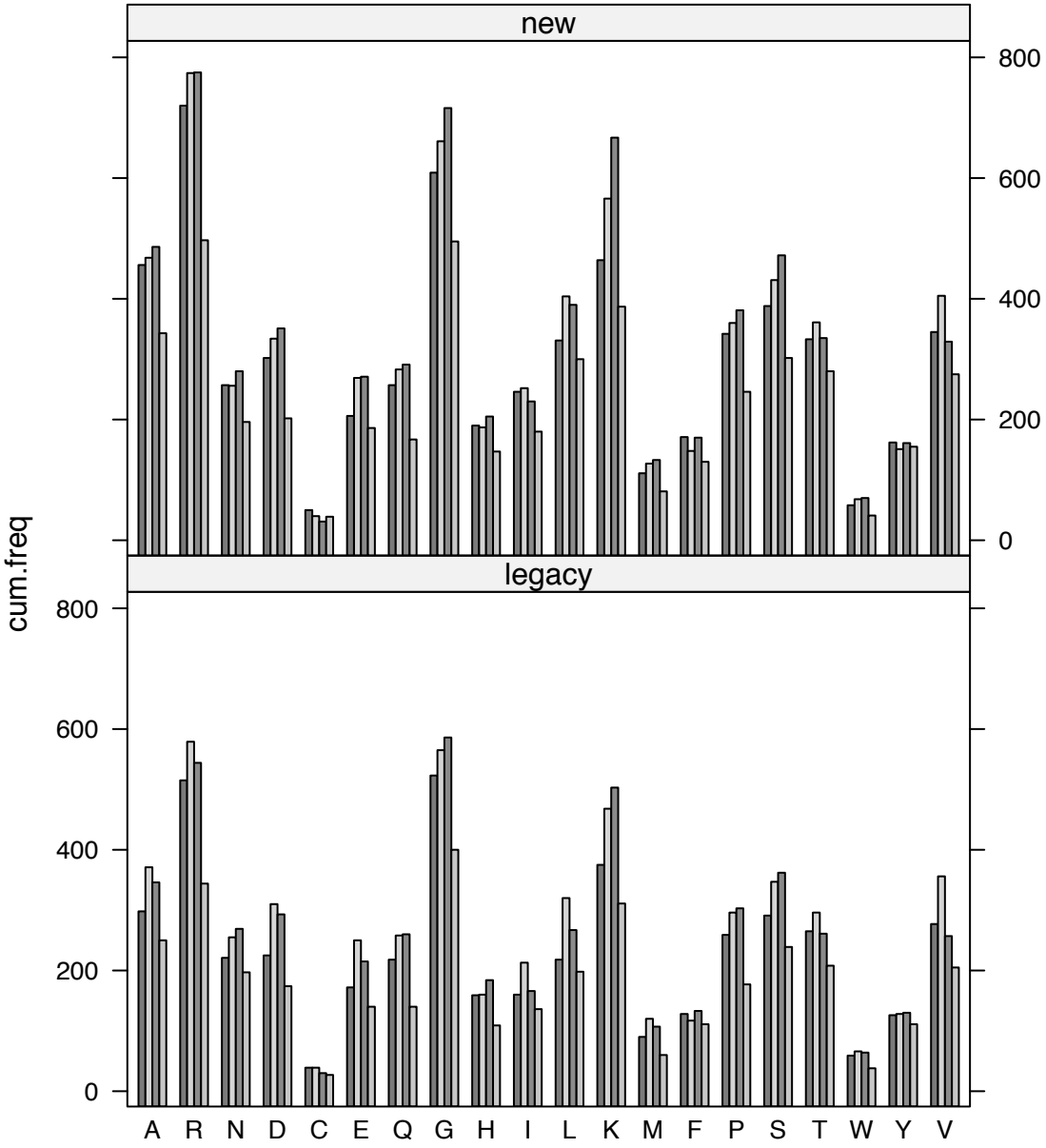


Figure 2.3: Number of side-chain nucleobase contacts in training set by amino acid residue and base. The bar plots show the number of contacting base-residue pairs with a C1′ – Cα distance less < 12 Å by residue and base in used in the previously published results (**legacy**) and my new training set. As would be expected for protein nucleic acid interfaces, arginine and lysine are most highly represented in the training set.

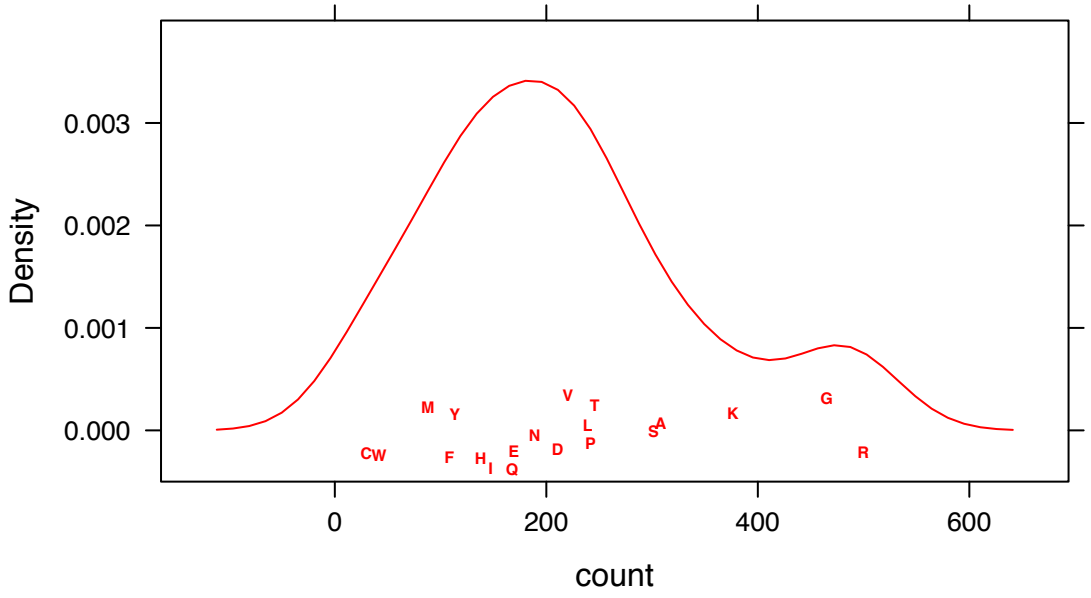


Figure 2.4: Density plot of the number of base contacts by residue type in the all-atom training set. The under-represented residues TRP and CYS and the over-represented residues ARG, GLY and LYS are those are often mis-predicted in the recovery tests. These bases are similarly under-represented in the test set, so statistics of low numbers may also explain why they appear to be incorrectly predicted.

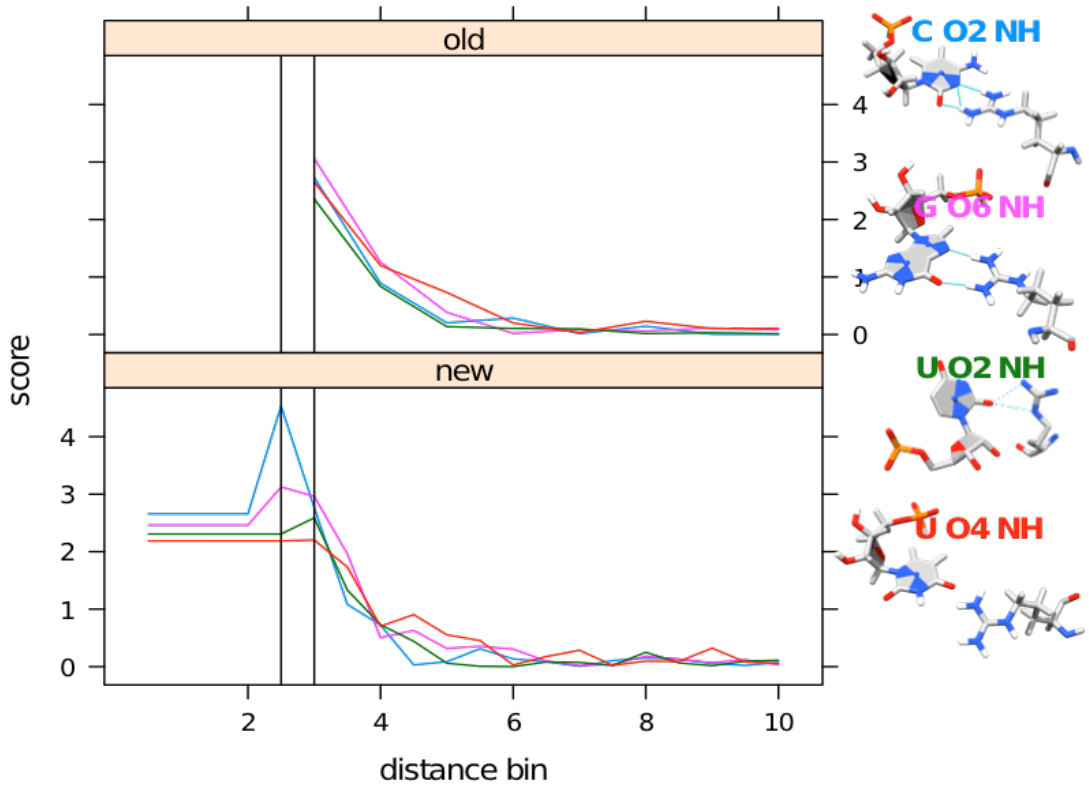


Figure 2.5: Score profile for contacts between arginine NH atoms and all the base oxygen acceptors. These atom pairs frequently participate in similar hydrogen bonding interactions. The plots show the log-odds score contributions with distance in the scoring function used by Zheng, et al. (2007) (**old**) and that used in the present work (**new**). I note that the lower representation of uracil relative to other bases in the new training set does not significantly affect the overall score with distance profile of the hydrogen bond interactions. However, the lower representation of the uracil may explain why uracil is less unfavorable (less positive) at 2.5 Å.

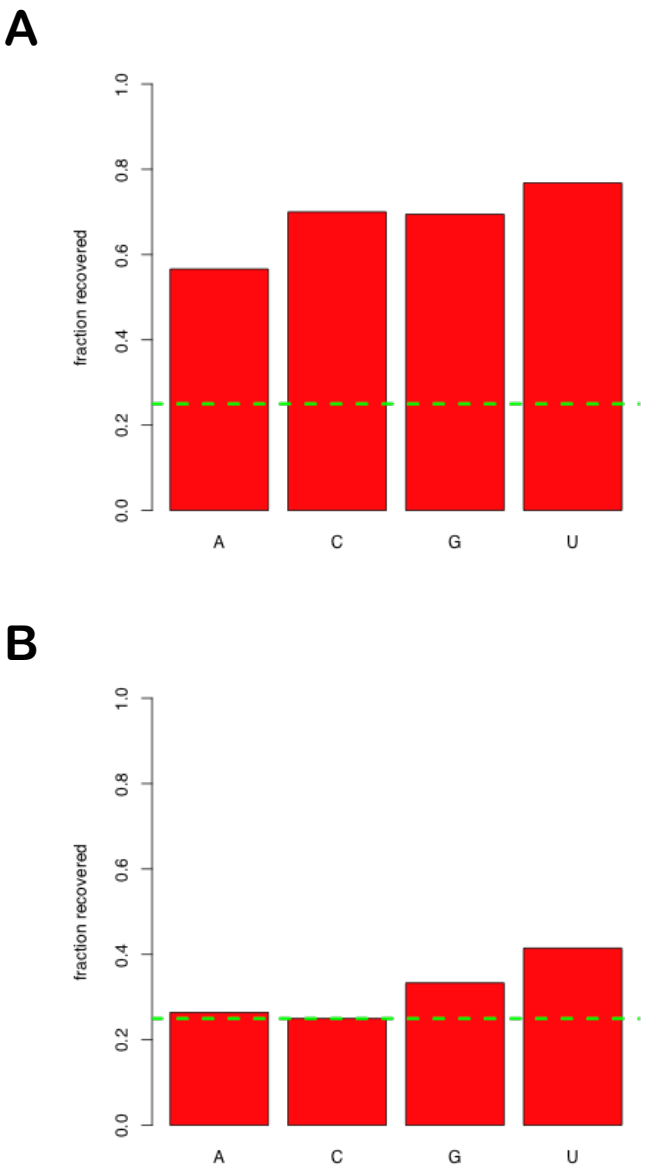


Figure 2.6: Difference between sequence recovery using scores from complete and fair matrix.
In tests using 30 representative proteins where the proteins bind primarily to single stranded regions of RNA. RNA bases at positions making contacts with the protein were substituted to each possible canonical base, repacked contacting protein side-chains, and scored against the protein. The predicted base for the position was that obtaining the lowest score. A base was recovered if the predicted base matched the base in the experimental structure. Much better results were obtained when repacking and scoring was performed with the entire training set (**A**) than with a fair scoring matrix (**B**).

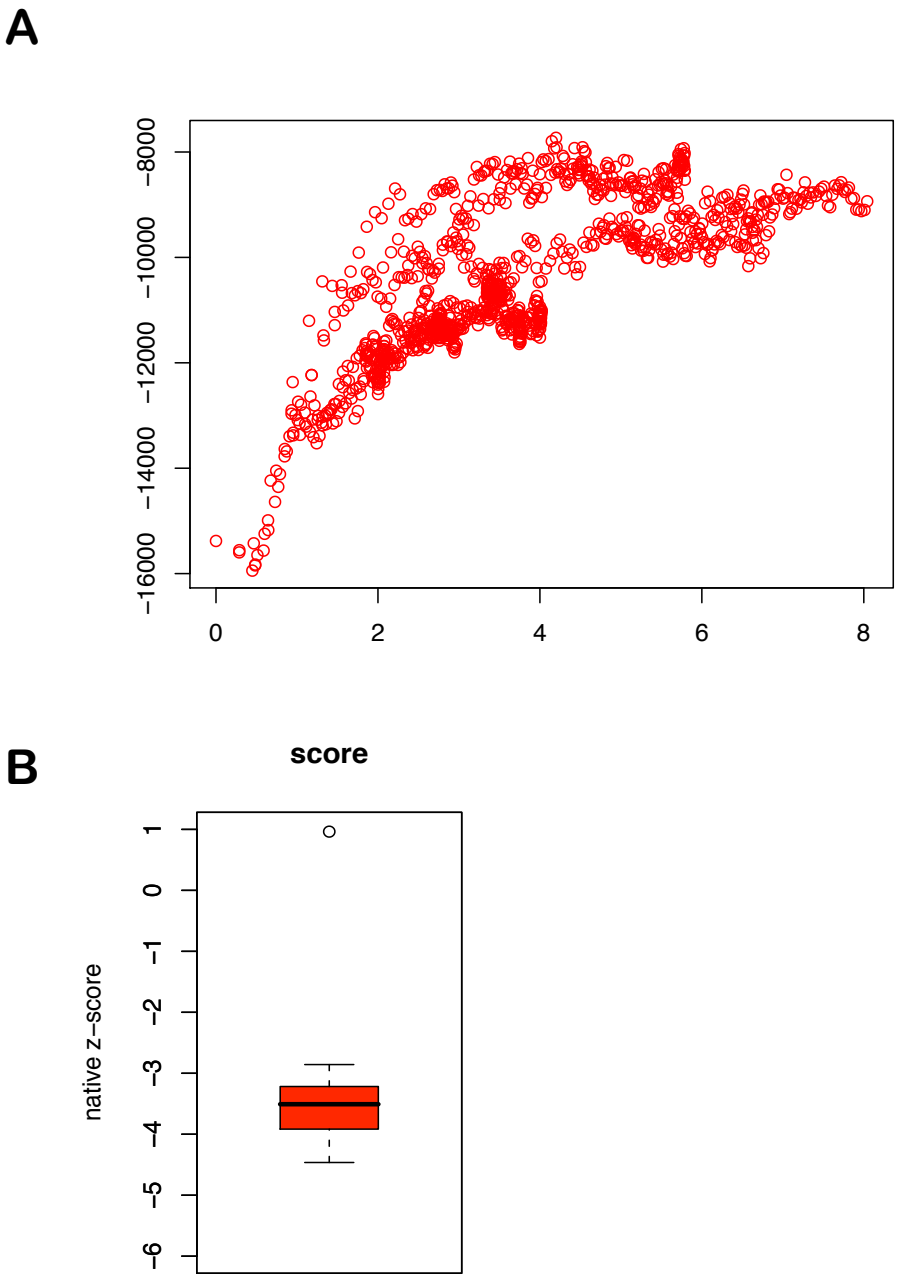


Figure 2.7: Discrimination of a correct RNA tertiary structure from near native decoys.
I applied the all-atom distance-dependent scoring function trained on ribosomal RNA data in decoy discrimination to score near native MD decoys. Decoys were generated using short 5-10 ps runs of an MD simulation and all have a reported RMSD of 0-8 Å. A sample decoy discrimination test is shown with the yeast group II self-splicing intron in PDB structure 1kxk (**A**). The significance of the native structure prediction in all 16 structures tested is summarized in a box plot (**B**).

Table 2.4: List of structurally diverse RNAs included in the test of RNA self-scores.
Many of the structures used in the RNA structure set were solved because of their interesting functional roles in cells. While many of them have tracts of Watson-Crick base pairs in regular A-from double helices, they exhibit many interesting tertiary structures as well.

PDB ID	structure title	molecular weight (g)
157D	RNA duplex containing G(anti).A(anti) base-pairs	7702.78
1CSL	RRE high affinity site	8989.52
1I9V	tRNA-neomycin complex	25445.56
1KD5	r(GGUCACAGCCC)2 metal free form	6962.34
1KFO	RNA helix recognized by a Zn-finger protein	6131.60
1KXK	domain 5 and 6 of Yeast group II self-splicing intron	22669.21
1MHK	hook-turn RNA motif	8370.00
1NLC	HIV-1 DIS(Mal) duplex Zn-soaked	15197.32
1NYI	hammerhead ribozyme	13443.82
1XJR	RNA element W from SARS Virus Genome	15483.81
280D	RNA dodecamer	15281.28
361D	domain E of thermis flaceis 5S rRNA	12969.90
3TRA	phenyalanine transfer RNA crystals	24205.81
422D	GAUCACUUCGGU	7578.62
437D	RNA psuedoknot	9292.84

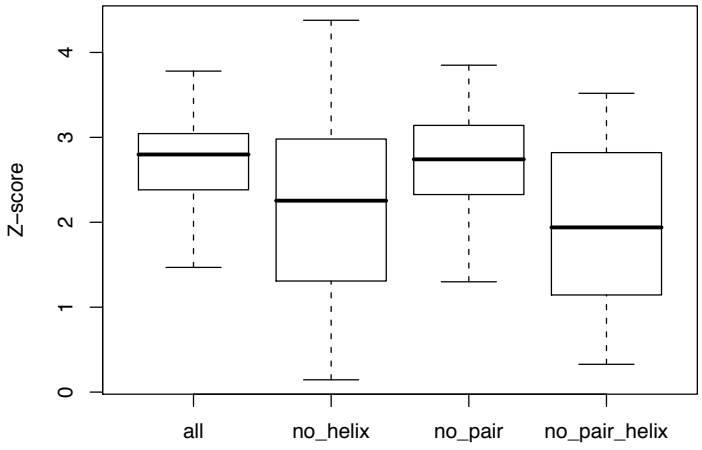


Figure 2.8: Test of the empirical scoring function to identify correct RNA tertiary structure.
Each box and whisker plot summarizes z-scores for tests with the RNA structures listed in Table 2.4. For each structure, near native decoys were created using short molecular dynamics simulations as shown in Figure 2.7. The individual box and whisker plots address whether base pairing interactions or base stacking contribute the most to the identification of the correct structure. A more positive z-score indicates better performance in selecting the native structure. All contacts were included in scoring 'all'. The 'no_pair' scores explicitly avoided scores between Watson-Crick base pairs while 'no_helix' explicitly skipped neighboring pairs in a helical region. The 'no_pair_helix' omits scores from base pairs and adjacent helical residues. The structure information was extracted from secondary structure annotations.

Chapter 3. Scoring Functions in RNA Binding Protein Specificity Prediction

A. Introduction

An understanding of the specificity of RNA binding proteins (RBPs) would best be demonstrated by retargeting an existing RBP to bind to an alternate target sequence. The redesign of a RBP places additional demands on the scoring function and structure search beyond what would be required for simply finding the target sequence of the native protein. The scoring function must guide the search of structure space following a base or residue substitution. Additionally, the scoring function must be sensitive enough to correctly rank base preferences so that relative improvements in candidate designs may be evaluated. In this chapter, I benchmark how statistical scoring functions and a Rosetta potential perform on tasks where the experimental data allows us to test challenging aspects of design.

In order to check whether the scoring functions and design approaches are ready for designing a specificity switch, I used existing experimental data to evaluate key indicators of the performance of the approach to specificity. The structure-based approach to specificity requires that important contacts between bound protein and RNA are correctly predicted. However, correct interface structure prediction is necessary but not sufficient for specificity prediction. Specificity prediction requires that the relative binding energies of the possible bases in a bound RNA sequence be

predicted with sufficient precision that the preferred base can be determined by comparing those energies.

We need more detailed specificity information for design applications than we might need for simple prediction of the RNA target of a RBP. For each possible RNA sequence that may be presented to a RNA binding domain, we would like to know the relative preferences for each base in the form of a position weight matrix. Assuming independence between base positions, the position weight matrix (PWM) is the relative frequency with which a sequence would be bound by a protein with each canonical base in a given position (Stormo, 1998). Knowing precisely how the PWM representation of base specificity changes with changes to the protein sequence would provide the information needed for designing a specificity switch in a step-wise manner.

A structure-based approach to design requires an effective scoring function. As discussed in the previous chapter (Chapter 2.A), proteins make use of sequence variants of a handful of domains to specifically recognize nearly all possible single stranded sequences. However, the possible sequence and structure space within the region of RBP in contact with each base is still large. An ideal scoring function would predict the relative changes in base binding specificity resulting from individual amino acid changes. Thus, specificity calculations are a bigger challenge for scoring functions than our previous docking tests.

The next chapter (Chapter 4.C) will apply the benchmarked scoring functions to a design application. In this chapter, I compare the performance of two pure statistical scoring functions and an implementation of the Rosetta scoring function for predicting the target specificity of RNA binding proteins. I benchmark the scoring functions for

their ability to reproduce interface structure and important residue-base interactions. I evaluate the performance of the scoring functions to constrain sequence base on energy in known structures of protein-RNA complexes. Additionally, I compare specificity predictions with experimental data for protein-RNA complexes for which a protein-RNA complex has been solved and for which position weight matrices representing binding preferences have been determined experimentally.

B. Background

Statistical and non-MD based physical approached have shown promise in predicting protein-DNA (Morozov et al., 2005) and protein-protein specificity (King & Bradley, 2010), but have so far not been applied extensively to RBPs. In this chapter, I present the first application of local structure optimization using pure-statistical scoring functions and a mixed physical and statistical scoring function to explore the RNA binding specificity of RBP. The present computational development has been possible in part because new experimental approaches have been used to characterize the RNA sequence specificity of RBPs that yield information on large number of sequences, as opposed to traditional biochemical and biophysical methods that probe a single protein-RNA pair at the time by measuring thermodynamic features of the interaction. These experimental approaches provide much needed insight into the recognition of the RNA by RBPs and allow for comparison with computational predictions, thereby providing a tool to validate computational models of these interfaces.

The same basic computational approach may be used to predict the required amino acid composition of a RBP needed to bind a given RNA sequence, but

experimental approaches are more limited. *In vitro* selection techniques provide a limited amount of information about alterations to protein sequence on the binding surface that maintain or improve RNA binding affinity. For example, one selection experiment found a number of single amino acid mutants of the binding domain of U1A protein with similar affinity for its native target RNA sequence (Y. Chen, Mandic, & Varani, 2008). In general, methods for discovering high-affinity protein-sequence variants and changes in sequence that alter RNA sequence specificity are labor intensive and are able to discover only the most tightly binding proteins rather than explore the energetics of an interface comprehensively.

Successful computational re-targeting of RNA binding proteins has not been reported in literature. However, a survey of work with transcription factors suggests approaches to design that should be applicable to RBP design as well. A recent book chapter outlines a basic computational approach to designing a zinc finger to recognize a specific DNA sequence *in silico* (Alibés, Serrano, et al., 2010). The approach consists of using a structural scaffold that is most similar to the desired recognition interaction, mutating the DNA to the desired sequence, altering a small number of protein side-chains to improve affinity and filtering by specificity. This is generally the approach that has been successfully used in the retargeting of homing endonucleases as well (Ashworth et al., 2006, 2010). A formal definition of affinity and specificity in terms of Boltzmann distributions of residue and base scores is the basis of this structure-based approach to design (Ashworth & Baker, 2009). This approach to design should generally transfer to work with RNA, but validation of this approach has been limited

by the availability of experimental information about specificity as well as by the lack of experimental design examples for validation.

In this chapter, I investigate the use of structure-based statistical approaches to recover experimental binding-specificity information for protein-RNA interactions. I use both a naïve Bayes classifier and the Rosetta scoring function along with the Rosetta computational platform to both predict interface structure by repacking protein side-chains and to predict preferred nucleic acid and amino acid composition at the interface. I find that both the simple statistical approach and Rosetta capture sequence specificity with RNA binding proteins. The Rosetta approach performs better in all tests and provides a promising starting point for a structure-based exploration of binding specificity in the most common RNA-binding protein motifs. The tools described herein are a starting point for designing a protein to target a specific RNA sequence. I validate concept of designing a specificity switch by demonstrating that the scoring function captures selection-based designs of a domain of the protein Pumilio1.

C. Development of Tools for Specificity Prediction

Structure-based specificity prediction requires a scoring function able to guide the search of local structure and sequence space. I search sequence space by substituting bases and residues and then search local structure space to optimize the structure. Ideally the same scoring function is used to evaluate steps in the structure search and to assign correct relative energies to base and residue mutations. The interface repacking following base or residue substitution requires both the inter-molecular scoring term for the evaluation of base specificity and the intra-molecular

protein scoring term since protein side-chains interact with each other, and when one is modified, the protein self-energy also changes.

Our structure-based approach to discovering target specificity of RBPs in complex with single stranded RNA involves substituting each base position and repacking surrounding residues. We assume independence between base positions and use repeated Monte Carlo minimization of base-adjacent protein side-chains to infer the order and magnitude of changes in binding energy (a measure of specificity) for the four canonical bases.

1. Scoring functions

I benchmarked three different scoring functions for specificity applications. I employed two purely statistical potentials trained on protein-RNA structures or protein-small molecule data and an implementation of Rosetta optimized for protein-DNA design problems. While the Rosetta scoring function is a mixed statistical and physics-based scoring function that handles protein intra-molecular scores, packing protein side-chains required the inclusion of a statistical term for protein intra-molecular energies (Samudrala & Moult, 1998).

a. intermolecular energy terms

I assessed a number of models for assigning atom types and pair probabilities to statistical-based potentials for use in packing and specificity prediction. The main model employed in this study is the all-atom model previously developed in our group (Robertson & Varani, 2007; Zheng et al., 2007). The model is trained on x-ray structures of protein-RNA complexes and specified in the form of a scoring matrix. I also performed the same tests with a scoring function recently described by Bernard and

Samudrala (2009). This second model circumvents problems with overtraining and sparse data that may limit the performance of the protein-RNA statistical score. The small-molecule trained *generalized scoring function* was generated by training on an unrelated data set consisting of protein interactions with small molecules covering atomic configurations similar to those seen in nucleic acid bases. Additionally, I used the Rosetta scoring function that combines several knowledge-based terms with a simple electrostatic potential. These knowledge-based approaches are more computationally tractable than purely physical approaches exemplified by potentials used in molecular dynamics simulations.

all-atom distance-dependent scoring function: I have previously developed an all-atom distance-dependent statistical potential for the evaluation of protein-RNA complexes. The potential is a naïve Bayes classifier based on distances between heavy atoms and was previously described (Robertson & Varani, 2007; Zheng et al., 2007). Parameterization of this scoring was described in the preceding chapter (Chapter 2.C.6).

Generalized distance dependent potential: A variation of the distance-dependent statistical potential that circumvents the problem of overtraining for protein-DNA and, potentially, protein-RNA structures has recently been implemented (Bernard & Samudrala, 2009). I refer to this function as the *generalized scoring function*. This scoring function is trained on protein-small molecule complexes from the Cambridge Structure Database (Allen, 2002). Though it was tested on protein-DNA interactions, it should be applicable to protein-RNA interactions as well. The interactions observed with RNA may in fact be more similar to the interactions molecule case since there are

significantly more contacts with the nucleobases and ribose in RNA than in DNA (Bahadur et al., 2008). The smaller bin size used in training the protein-small molecule scoring function may compensate for some of the difficulties from using fewer atom types.

In the generalized scoring function, the reference state from the primary reference was part of the scoring matrices provided by Dr. Brady Bernard. The study by Bernard and Samudrala (2009) found that discrimination between decoys was improved with the introduction of a radial distribution function and a mean reference state. In this approach, contacts are quantified in terms of contact density, by dividing the number of contacts by the volume of the bin's spherical shell. Instead of using the distance-distribution from an atom type-agnostic analysis, the reference state used was the mean of the prior-probabilities. This implementation sacrifices atomic specificity, compared to the all-atom potential, but gains considerably in terms of size and resolution of the training set.

Rosetta scoring function: Finally, I used the Rosetta scoring function with parameters similar to those previously described for the DNA-binding protein homing endonuclease design (Ashworth et al., 2006, 2010). The parameter weights used are shown in Table 3.1. While better performance would likely be achieved by modifying some of the weights for protein-RNA interfaces, the Rosetta weights are typically chosen to optimize performance on structure recovery tests. Since one of the principal difficulties the present work has been in constructing the test sets, optimizing the weights could lead to over fitting. RNA and DNA are chemically similar with the exception of the ribose 2'-OH and a methyl group on thiamine. Thus, I expected the

unmodified score to work well enough to justify avoiding the possibility of generating an over-trained data set.

The same term weights (Table 3.1) are used for scoring contacts between proteins and nucleic acids as is used for scoring contacts between protein residues and nucleic acid residues. The residue definition files are used to define which atoms participate in the interactions for which non-zero term weights are defined. The scores assigned by a score component such as hydrogen bond interactions between protein residues and nucleic acid bases reflect developments to the Rosetta terms aimed at protein-RNA interactions (Y. Chen et al., 2004). The energy function is dominated by a 12-6 Lenard-Jones potential and an orientation-dependent hydrogen bonding term (Kortemme, Morozov, & Baker, 2003).

In a test aimed at using the scoring function for a design application, I also tested a variation of the Lenard-Jones (LJ) repulsive term. Because we are not allowing the protein or RNA backbones to relax, the repacked structures have less conformational freedom to avoid atomic clashes. In some tests, I employed a linearly rather than exponentially increasing repulsive term. With the normal LJ, the energy of clashing atoms quickly dominates the interaction. Since the search of structure space is limited to repacking, the structure may not be able to adjust to minimize all clashes. The lower LJ term makes sense because we are interested in estimating energy change due to base or amino acid substitutions.

b. intra-molecular terms

The repacking of protein side-chains is necessary following residue mutation. Amino acid side-chains have very different chemical properties, sizes and degrees of

freedom. Hence, following substitution it is necessary to establish the preferred side-chain dihedral angles. Additionally, following the substitution of a base or amino acid, a more realistic energy should be obtained by allowing all contacting residues to repack.

We seek to calculate the energy of a particular amino acid at each position in an interface, with respect to its complete molecular environment. To discover the amino acid dihedral angles for a given side chain that work best at a protein-RNA interface, we need both the inter-molecular terms described above (a), and an intra-molecular score describing interactions between amino acids

Rosetta scoring function: The Rosetta scoring function is mostly based on general statistical and physical terms whose weights are more dependent on the molecular properties of interacting residues than on the class of molecules that are interacting. The same term weights (Table 3.1) are used for scoring contacts between proteins and nucleic acids as is used for scoring contacts between protein residues and nucleic acid residues. The intra-molecular interactions between protein residues important for properly packing protein side-chain are the most well developed part of Rosetta. The scoring function has been extensively tested in applications of constrained modeling, *de novo* design, and loop modeling (Rohl, Strauss, Misura, & Baker, 2004). Correct protein intra-molecular scoring is an important component for accommodating base mutations and for predicting the relative energies of nucleobase binding. When calculations are executed with Rosetta, they contained both intra-molecular terms.

While additional work has been done to extend Rosetta for predicting RNA tertiary structure, I do not integrate those developments in the current work. The current work concentrates on proteins that recognize single stranded RNA in regions

where there are no significant canonical Watson-Crick or Hoogstein base pairs (Leontis & Westhof, 2001). Understanding RNA base pairing and subtle differences in nucleobase-dependent stacking energies is required for properly modeling RNA tertiary structure.

For the purpose of this work, the terms included for protein structure and for binding will also capture electrostatic and hydrogen-binding interactions between bases. These terms are not sufficient for RNA modeling applications (including for allowing complete freedom of base rotation around the glycosidic bond). However, the included energy terms should be sufficient to discriminate between bases where one base type interacts with adjacent RNA atoms in an energetically favorable manner. The single stranded RNA recognition problem I am considering is more similar to the protein-DNA recognition application of Rosetta (Yanover & Bradley, 2011) that to the RNA tertiary structure prediction.

The work with Rosetta in the *de novo* prediction of RNA structure will be useful in future protein-RNA specificity and design applications involving RNA with more complex tertiary structure. A coarse RNA intra-molecular term was developed for use in building RNA structures from fragment libraries (Das & Baker, 2007). The recent correct prediction of the structure of small highly ordered RNAs that have challenged RNA structure prediction applications (Das et al., 2010) will be important for solving the general problem of the sequence specificity of RBP binding. The inclusion of these terms would require significant parameterization that is beyond the scope of this work. The Rosetta scoring function used in this work captures only the basic intra-molecular

interactions for nucleic acid energy that have been demonstrated in work with DNA binding domains.

empirical scoring functions: The empirical scoring functions rely on learning interactions from training data. The learned model is highly dependent on the types assigned to atoms, which are often linked to residue type. Due to the data available in the PDB, potentials of mean force must be derived independently for the stand-alone molecular structure of proteins or of RNA from those aimed at correctly predicting the bound structure. For repacking applications and for correctly capturing the energy of a bound structure, it is necessary to be able to predict the self-energies of the molecules that adopt different energy in the bound and unbound state. An empirical term for scoring protein self-energy is thus included from previous work.

For the empirical scoring applications, I used an energy function trained for intra-molecular protein-protein interactions that was formally similar to that used for protein-RNA interfaces. This intramolecular protein term was implemented as described (Samudrala & Moult, 1998). The implementation is identical to that of the all-atom, distance-dependent term used for inter-molecular calculations, but the resolution of the potential is significantly higher due to the much larger training set that was used to train this intramolecular term. The purely empirical term used for repacking and residue scoring is a weighted sum of the protein-RNA intermolecular term and the protein intramolecular score.

I considered including RNA intramolecular scores by adding an additional term. I previously investigated a Bayesian term for RNA tertiary structure (Chapter 2.F). However, I could not independently demonstrate the quality of those scores. The term

was omitted from specificity calculations. I expected that for RBPs recognizing single stranded regions, the protein-RNA and protein intramolecular terms would be most important.

D. Searching Structure and Sequence Space

A considerable challenge in the structure-based approach to protein design coincides with searching structure space in such a way that meaningful comparisons can be made between candidate or decoy structures. Sampling sequence space adds to the complexity of the search, because the size of the combined space that could be searched is enormous. By asking more narrow questions about the relative favorability of single mutations in sequence space, I sought to make the problem tractable and extract information useful for a simpler retargeting task. A conservative approach to retargeting asks whether base preference at a RNA position bound by and RBP can be switched by mutating a small number of contacting residues. The retargeting task is less demanding because the protein structure is not significantly altered.

Scoring functions containing statistical terms allow for a quantitative comparison of possible structures but do not attempt to model the path between those structures. Statistical scoring functions thus sample structure space elements from those observed in known structures. The search of structure space is nearly as important as the scoring function. Searching too many (or improbable) conformations may challenge even good scoring functions. The correct conformation may be missed if the structure space is over-searched. I seek to alter the binding specificity of a protein, by making the minimum number of sequence changes. Since single amino acid changes

are unlikely to significantly alter the fold of the protein, side-chain repacking is likely to account for most of the structural changes between the original and altered structure.

The scoring function can be used for specificity prediction if we can correctly model the energy of the structure with each base or residue substitution of interest. We seek to model specificity, which is the relative energies of the canonical bases. Even if all contacts are not predicted, important contacts are likely to be more frequent. If the important contacts are modeled correctly, specificity prediction may still be perfect even if the interface is imperfectly modeled. However, correct modeling of the interface is useful in assessing the scoring function.

1. Interface repacking

I repack all interface side-chains using well-established side-chain modeling techniques and the scoring functions described above. I sample side-chains from the Dunbrack backbone-dependent rotamer library (Dunbrack & Karplus, 1993), then identify the most favorable conformations using a Monte Carlo heat-bath algorithm (Newman & Barkema, 1999). This simplified search is similar to one recently employed for protein-protein interfaces (M. Lu, Dousis, & Ma, 2008b). The score of residue pairs are only dependent on the relative conformations of residue pairs. Since I sample from a finite set of relative orientations, the scores can be pre-calculated and cached.

2. Specificity search

I have further introduced amino acid and nucleotide packing and substitution methods required for the design of RNA-binding proteins. Starting with a high-resolution x-ray structure, the program calculates specificity by mutating a position,

repacking the neighboring residues and scoring that position using the distance-dependent potentials described above. I establish amino acid or nucleotide specificity at each position by independent substitution of each monomer position, one at a time.

A few experimental techniques have been developed to discover the position-weight matrix (PWM) for proteins binding to RNA independently of structure. These experiments include the SELEX experiment (Tuerk & Gold, 1990), NMR-SIA (Beuth, Garcia-Mayoral, Taylor, & Ramos, 2007) and the recently developed 'RNAcompete' approach using DNA microarrays (Ray et al., 2009). RNAcompete discovers the relative preference for each base at a binding position by ranking bound sequences in order of binding affinity. Aligning the sequences yields the relative probabilities for each base at each position. These results may be visualized as a PWM (X. Chen, Hughes, & Morris, 2007) and directly compared with the results from our calculations. I used a set of overlapping structures for which experimental PWM data are available and for which an experimental PDB structure was available as the basis for calculating the PWM using our computational approaches.

The probability of finding a specific base at a position is related both to the energy of the bound states and to the number of states available for binding. Each time we perform a point mutation *in silico*, I optimize the contacting side-chain positions using a Monte Carlo algorithm. However, in the score landscape there exist a number of local minima. If the scoring function is a correct representation of the energy, the local minima are physically relevant states. By repeating the same base substitution many times, I map the set of local minima scores that represent the energy levels accessible to the substituted base and the probability that the energy state is accessed.

In order to obtain the relative frequencies of bases at an RNA position, we perform a virtual binding competition. The virtual competition approach allows estimation of relative nucleobase binding probabilities based on accessible energy states. The diagram in Figure 3.1 illustrates the virtual competition approach. One score is randomly sampled with replacement from the set of calculated scores for each base substitution. The base predicted by each competition is that for which the lowest score was drawn. This virtual competition is repeated 1,000 times. The frequency of each base binding at the position is then taken to be equal to the fraction of competitions won by that base.

E. Comparing Specificity Predictions with Experimental Results

With the advent of high throughput methods for discovering binding motifs for transcription factors, several groups have worked on the problem of representing the binding preferences in the form of a PWM (Badis et al., 2009; Stormo & Zhao, 2010; Y. Zhao & Stormo, 2011) and in methods for quantifying the similarity of independently discovered motifs (S. Gupta, Stamatoyannopoulos, Bailey, & Noble, 2007; Tanaka, Bailey, Grant, Noble, & Keich, 2011). I draw upon these other works to solve the slightly different problem of evaluating the correctness of motifs from structure-based predictions against a set of motifs from a heterogeneous set of experiments.

The quantification of the performance of PWMs calculated using RBPs from the PDB against experimental results from experiments such as SELEX and RNAcompete is performed in two steps. First, since the experimental position frequency matrices do not have structural annotation about contacts with the RBP, I must align the sequence in the solved structure with the experimental binding motif. Secondly, I compute the

average per-position difference between the calculated and experimental PWMs representing the same protein.

1. Alignment of position weight motifs

Alignment of structure sequence to an experimental PWM is performed using ideas introduced for finding matches for TF motifs (van Nimwegen, 2007). I assume that the bound RNA sequence in the structure is a high affinity sequence. To obtain the alignment that will be used for comparing calculated and experimental motifs, I find the best alignment of the structure sequence to each experimental motif to be that with the highest binding probability. In the common model where base positions are assumed to be independent, the binding probability is simply the product of the probability of finding a specified base in each position.

$$P_s = \prod_i p_i(b_i) \tag{3.1}$$

The unique alignment of the structure sequence to the experimental motif is then stored and used to compare calculated motifs based on the query structure using each of the scoring functions.

2. Position weight matrix differences

The calculation of the difference between calculated and experimental position weight matrices is performed using the Kullback-Leibler divergence (KLD) and Euclidean distance (ED) (S. Gupta et al., 2007). The KLD is a measure of the extent to which the assumed underlying probabilities of predicting a base at a position agrees with the experimentally observed probabilities. The KLD is the number of additional bits needed to express a message given the incorrect underlying assumption. Since both

the experimental and calculated PWMs are approximations of the underlying position preferences, we use a symmetric form of the KLD that averages the KLD alternatively assuming that the experimental and calculated position probabilities are correct.

$$\mathrm{KLD}(X,Y) = \frac{1}{2}\sum_{a\in A}\left(X_a \log\frac{X_a}{Y_a} + Y_a \log\frac{Y_a}{X_a} \right) \tag{3.2}$$

X and *Y* are the sets of probabilities for all possible residues composing an alphabet *A*; *a* is a specific letter or residue in the residue alphabet. The symmetry is introduced because the experimental and calculated values are only representations of the underlying motif. A problem with the KLD is encountered when any letter receives a zero probability and the logarithm cannot obviously be taken. To avoid this problem, probabilities were adjusted so that the probabilities are $X_a = 0.01 + 0.96 f_a$, where f_a is the frequency of residue *a*. This adjustment avoids calculation errors arising from substitutions with zero probability without greatly altering base probabilities.

The ED has less justification from an information theory perspective, but yields more easily comparable distances (Tanaka et al., 2011). The ED assumes that the set of base fractions at each position represents a point in a space of dimensionality equivalent to the number of possible residues. The ED is the shortest distance between the points in this space.

$$\mathrm{ED}(X,Y) = \sqrt{\sum_{a\in A}(X_a - Y_a)^2} \tag{3.3}$$

X_a and Y_a are again the experimental and predicted probabilities of binding residue *a* from the alphabet, *A*, of possible residues. The advantages of the ED metric include more linear scaling and none of the problems of the KLD in dealing with base probabilities near 0 or 1.

While each metric posits underlying assumptions that do not capture fully the complexity of our problem of comparing structure based calculations to a variety of experimental measures, they allow us to quantitatively rank the performance of the scoring function on each related pair of experimental and calculated motifs. The difference between two motifs is the sum of the KLD or ED score for the aligned experimental and calculated positions.

For most RBP's, multiple experiments inform our knowledge of the binding preference of a RBP domain and multiple structures solved in complex provide a good basis for structure-based calculations of binding preference. Ideally, given a large set of experimental data, I would derive a consensus binding-motif from the experimental data by clustering the experimental PWMs and averaging them using an averaging metric weighted by confidence in the experimental data. However, the data set is sparse and quantifying the quality of the experimental data cannot be summarized in a simple metric. Since I only need a quantity that captures the performance of different computational approaches, I choose to average the different scores across all pairs of experimental results and structure-based predictions. To correct for differences in the size of predicted motifs, I use the average per position difference.

F. Structure Benchmarks

The structure based computational approach to specificity prediction is predicated on predicting likely contacts between proteins and RNA bases. Conversely, incorrect prediction of these interactions may be diagnostic of limitations in the scoring function at the binding site. I use structure recovery to benchmark the performance of

the scoring function in reproducing native contacts believed to be important for conferring specificity to the protein.

An ideal design program would be able to reproduce interactions observed within an interface with accuracy, and to also accurately reproduce the energetic consequences of changes in sequence within that optimized structure. While imperfect scoring functions are a key problem, the computational cost of comprehensively exploring structure space is also prohibitive; thus, sampling must be evaluated as well.

I first assessed how well the combination of search algorithm and scoring function performed by evaluating whether an interface could be correctly reproduced following the repacking of amino acid side chains, because satisfactory performance in this test is essential for a design program. Furthermore, structures where the amino acid sequence is altered must be allowed to relax to a low energy structure, so that neighboring residues can accommodate a change in size and chemical composition at that position. Correct repacking of neighboring amino acid side-chains is necessary for correct estimates of the energetic cost of sequence substitutions. Thus, an analysis of the recovery of contacts important for specificity at the protein-RNA interface provides a benchmark for use of the potential function in specificity applications.

1. Dihedral angle recovery

I assessed repacking quality by comparing predicted side-chain dihedral angles to the original x-ray crystal structures, after repeatedly repacking all interface residues using a Monte Carlo algorithm. I use recovery of side-chain dihedral angles at the re-packed interface as a metric of the prediction of the native structure at the interface. I

repacked the interfaces of 30 representative RBPs binding regions binding to primarily single stranded regions of RNA (Table 3.2).

Interface protein residue side-chains were repacked twenty times, and the lowest (best) scoring complex was selected for further analysis. Residues were selected as interface residues if side-chain heavy atoms (C, O, or N) were less than 6 Å from a nucleobase heavy atom. The tests used 9^n+20 conformations per residue (where n is the number of χ-angles for each amino acid); the nine values per rotamer were selected evenly from the Dunbrack backbone-dependent library (Dunbrack & Karplus, 1993) and the 20 additional conformations were randomly generated. Recovery statistics are reported for the best scoring calculated structure for each binding complex. A χ_i value was considered successfully recovered if the computed value is within 40º of the structure value and all χ_i where $i < n$ are recovered. The choice of the ±40º cutoff values was made to select conformers within the correct rotamer bin.

As a general test of side-chain packing, Figure 3.2 shows the fraction of χ-angles for all residues by fraction of structures recovered correctly using the Rosetta, generalized and all-atom scoring functions. The results demonstrate that there is some variation in the quality of side-chain packing across all structure. Yet, the lower mean recovery of χ_i indicates that Rosetta performs better for most structures compared to the statistical potentials.

Since side-chains differ in chemical properties, the quality of side-chain packing could differ substantially by residue. Figure 3.3 shows the fraction of side-chains recovered to each more distal dihedral angle by residue. The fractions of side-chains recovered at the first dihedral angle are 0.86, 0.82 and 0.78 for Rosetta, the generalized

and all-atom scoring functions, respectively. However, for all residues, the performance of the purely statistical scoring functions is substantially worse than Rosetta at more distal dihedral angles. This suggests that the physical terms such as the Lennard-Jones terms and the explicit statistical representation of hydrogen bonds in Rosetta are helping to correctly place the side-chains.

When using scoring functions such as the purely empirical functions or Rosetta that employ statistical scoring terms, the exploration of structure-space is necessarily separated from the function which scores (assigns energy) to a conformation. In a physics-based approach the potential (scoring function) is used to minimize structure down an energy gradient. In the statistical approaches energy terms may exist in a discretized space for which a derivative is undefined or the scoring terms may not completely define scores over the entire physical structure space, thus the program component that provides candidate structures is decoupled from the evaluation of scores for large steps. The decoupling of structure explorations means that any structure not generated as a candidate structure will not be considered. During the repacking step a large number of side-conformations are generated using the general side-chain statistics described by Dunbrack and Karplus (1993).

In order to validate side-chain dihedral angle recovery errors are due to a problem with the scoring component and not the structure generation component, I performed a control calculation. I repeated analysis shown in Figure 3.3 with the dihedral angles from the experimental structures added in. Figure 3.4 summarizes the fraction of side-chain conformations recovered at each dihedral (χ) angle. With each scoring function and at each angle, the control calculations (*-ctrl*) recover a few percent

more of the correct conformations. This demonstrates that a small amount of the error can be attributed to the comprehensiveness with which structure space was explored. However, the error rate from this is acceptable since increasing the exploration of the side-chain angles dramatically increases computational time.

The side-chain recovery analyses (Figure 3.3 and Figure 3.4) show that all three scoring functions recovery the general position of residue side-chains. Furthermore, the misplacement of side-chains is due to a mismatch between the native conformation and the lowest calculated energy. Since the amino acid side-chains employ a variety of different functional groups, the analysis could have shown that for a given side-chain the conformation is mis-predicted at a higher rate with a given training set. Missing specific conformations would indicate a problem with the parameterization or training of the scoring function. The analysis of side-chain recovery by residue (Figure 3.3) did not suggest problems with the scoring functions *vis-à-vis* a particular residue type.

2. Recovery of interactions specifying base recognition

The recovery of characteristic classes of intermolecular interactions allows us to investigate failures and limitations on the potential. Important interactions characteristic of RNA binding interfaces include hydrogen bonds, cation-π and π-π interactions (Auweter et al., 2006). I identify these important interactions in a set of test structures and test for their recovery following repacking. Since hydrogen bonding and interactions with the aromatic ring are implicated in base recognition, the recovery of these interactions is diagnostic of the utility of the scoring function for specificity prediction. Using the test set of 30 RBPs binding single-stranded RNA (Table 3.2), I analyzed the recovery of key interactions using the three scoring functions in Table 3.3.

interaction recovery statistics: I define an interaction to be recovered if it exists both in the experimental structure and in the best (lowest energy) predicted structure calculated by me. Since in all tests I only allow side-chains to move and I am interested only in direct interactions with the nucleic acid, I consider only contacts between amino acid side-chains and the nucleic acid. I use a combination of often-used analysis programs and our own implementation of heuristics to recognize interactions important for base recognition interactions. I tested interaction recovery by comparing those recognized in the experimental structure to those in the structure with lowest energy repacked interface.

quantifying interaction recovery: I drew upon a well-tested program to identify hydrogen bonds. I employed the program hbplus to identify hydrogen bonds in PDB structures (McDonald & Thornton, 1994). The interactions of interest were culled from the PDB output and the hydrogen bond interactions in the predicted structures were compared to the x-ray crystal structures. The definition of a hydrogen bond was strict so both experimental and predicted hydrogen bonds are likely to be under-estimated. The reported hydrogen bond recovery results are probably under-reported for all three scoring functions.

Stacking interactions are characteristic of protein-RNA interfaces and, like hydrogen bonds, contribute to sequence specific recognition. While advanced computational methods have been developed to describe interactions with the aromatic nucleobases (Wintjens, Liévin, Rooman, & Buisine, 2000), the interactions can be reliably recognized using fairly simple heuristic tests. Cation-π interactions were identified using a heuristic where the vast majority of cations forming stacking

interactions on aromatic rings can be found in a conical region with the apex at the center of the ring, a base diameter similar to that of the ring and a height of 4.5 Å normal to the plane of the base (Wintjens et al., 2000). Cation-π pairs matching these criteria were identified using an in house program. π-π interactions were identified using a similar heuristic criterion but with the additional constraint that the angle between the ring-plane normal vectors could not exceed 30º (Morozova et al., 2006; Šponer, Leszczynski, & Hobza, 2001). The heuristics used for identifying interactions with proteins, due to base aromatic properties were used to identify possible cation-π and π-π interactions before and after repacking.

recovery results: The results of the tests of the recovery of interactions involved in base recognition are summarized in Table 3.3. With hydrogen bonds, which probably contribute the most to specific nucleobase recognition, the Rosetta scoring function outperforms the empirical scoring functions recovering nearly half of the hydrogen bonds. Rosetta recovers 49% while the generalized and all-atom scoring functions recover 36% and 43%, respectively. Most importantly Rosetta recovers more of the side-chain to nucleobase contacts than the other scoring functions. These results are expected as Rosetta contains explicit hydrogen bonding terms that are a large component of its interface structure recovery (Y. Chen et al., 2004). That the all-atom scoring function that is explicitly trained on protein-RNA interactions outperforms the generalized scoring function in hydrogen bond recovery is not surprising.

The cation-π stacking results demonstrate an interaction type for which all three scoring functions need improvement. The generalized scoring function recovers 40% of cation-π interactions compared with 32% for Rosetta and the all atom scoring function.

The generalized scoring function properly predicts most ARG stacking interactions, but none of the LYS interactions. Rosetta recovers fewer ARG interactions than the empirical approaches, but recovers some of the LYS. Overall the results are poor with cation-π interactions. The Rosetta scoring function does not contain a term that explicitly captures these interactions. While cation-π interactions are important components of specific binding by RRMs and PUF domains (Auweter et al., 2006; Y. Chen & Varani, 2011), the training set does not contain many examples of these interactions. Improving the description of cation-π interactions may improve predictions of RBP affinity and specificity.

The π-π stacking interactions are mostly (81% to 86%) recovered. The high recovery is not surprising since only side-chains are repacked and steric constraints limit the conformations allowed for these bulky side-chains. The test would be more informative when more degrees of freedom such as backbone conformations are allowed.

lessons from interaction statistics: The recovery statistics for interactions implicated in the recognition of specific nucleobases illustrate areas where all three scoring functions need improvement. The specific approaches to improving the performance of the scoring function are scoring function dependent. The empirical scoring functions are primarily improved through training while the Rosetta scoring function may require different terms or parameter weights.

lessons for empirical scoring functions: The empirical scoring functions are formulated in such a way that hydrogen bonding and cation-π interactions should be equally captured. The poor performance of the all-atom scoring function suggests that a

more high quality training data may be needed. The better performance of the generalized scoring function with cation-π interactions and worse performance with hydrogen bonds likely reflects the types of small molecules in the Cambridge Structure Database (CSD). The generalized scoring function demonstrates that, with sufficient data, more cation-π interactions may be recovered. An alternative approach with Bayesian statistics is to alternatively formulate the interface interactions in terms of groups and orientation (M. Lu et al., 2008a). However, whether alternate approaches outperform the inter-atomic distance dependent formulation when sufficient training data is present would require further investigation.

lessons for Rosetta scoring function: The Rosetta scoring function with term weights selected for DNA applications performs better than empirical approaches with hydrogen binding, but no better or worse with base stacking. The hydrogen bond results may reflect problems with searching structure space, but the stacking interactions with the aromatic bases may represent an interaction type for which the scoring function needs improvement.

The 49% recovery of hydrogen bonds is surprisingly low since some of the principle terms ('hbond_' terms from Table 3.1) in the scoring function score for hydrogen bonding. The limited exploration of structure space using only protein side-chains may be partially responsible for the low recovery. Structures from the PDB may have orientations reflecting an energy minimum while the distances between atoms and bond-angles may deviate from the ideals values for the statistical terms in the Rosetta scoring function. With more degrees of freedom including backbone angle optimization, Rosetta may be able to recover more hydrogen bonds. However,

additional freedom in structure space can dramatically increase the computational complexity of the problem and decrease the probability of finding the best solution. Hydrogen bond statistics should be monitored as other avenues to improving the algorithms for searching structure and to improving the scoring function for RNA applications.

With respect to interactions between protein side-chains and the aromatic nucleobases (cation-π and π-π interactions), no term in the Rosetta scoring function directly captures these interactions. Lower resolution (not atomic level) approaches to RNA structure have explicitly included terms that include base stacking (Das & Baker, 2007; Sykes & Levitt, 2005). Neither a statistical scoring term nor a physics-based term that calculates the electrostatics of the π molecular orbitals is currently available for the Rosetta scoring function. Including a term for these interactions may be more important for protein-RNA interactions than it has been for the previous applications of Rosetta to protein-DNA and protein-protein interactions (Yanover & Bradley, 2011; King & Bradley, 2010).

summary: In a structure-based approach to predicting the binding preference of RBPs for specific nucleobases, the correct prediction of interface structure is of central importance. The findings from the interaction recovery statistics shows that the recovery of key interactions thought to be important for specificity predictions are far from ideal. The missed interactions help explain some of the difficulties in applying the empirical and Rosetta scoring functions to specificity calculations (section G below).

Many hydrogen bond interactions, which contribute the most to specificity, are likely missed in empirical calculations because of insufficient training data. However, in

both the empirical and the Rosetta scoring functions, the limited exploration of structure probably fails to provide candidate conformations that have the characteristic of ideal hydrogen bonds.

The stacking interactions have been shown to contribute to nucleobase selectivity (Morozova et al., 2006). However, these interactions are likely to contribute most to protein-RNA binding affinity and to defining the binding site when modeled structures are considered. The distance-dependent empirical scoring functions should be capable of discovering these interactions. While base stacking interactions are key to specificity and affinity, they are likely to be under-represented in our training set (Table 2.3), which was selected to represent diverse structure rather than specifically bound structures. The interaction between cations and aromatic residues with the aromatic bases is a feature that would be worth further developing for the application of Rosetta to RBPs. However, the structure recovery statistics were intended to assess the performance of the scoring function with interactions that may contribute to specificity prediction. Thus, I benchmark the performance with specificity predictions more directly.

G. Specificity Calculations

My work aims to develop computational, structure-based approach to predict the preferred target sequences of RNA binding proteins. The rapidly expanding knowledge about protein structure potentially provides a wealth of knowledge about the structural determinants of specificity in RBPs. The development of a general computational method for predicting binding specificity would be generalizable to include predicting the specificity of structures for experimentally uncharacterized RNA

binding domains and to designing proteins to bind novel RNA sequences. A computational approach that correctly captures relative binding preferences at each position is likely to be useful in a design application where an existing protein is modified to recognize a different base.

In order to design a tool that specializes in specificity prediction and in designing protein with altered specificity, we must create challenging tests for the computational approach to directly test specificity. I reviewed experimental techniques that measure RBP binding specificity in Chapter 1.B.2. Additionally, I discussed the sparseness of the set of known RBP complexes with RNA in the context of selecting novel structures for training the empirical scoring function (Chapter 2.C.3). The structure-independent and structure based approaches to understanding sequence specific recognition of RNA by RBPs illustrate the challenges of relying on experimental approaches. However, the experimental data that are available do provide a basis for constructing tests of my computational approach. In a first test, I used a set of bound structures as a specificity library. In a second test, I performed a more stringent test with protein complexes for which we have both a structure and an independently assayed binding motif.

The set of solved protein-RNA structures can be viewed as a library of specifically bound nucleotides. Much insight into specific recognition has been gained by studying the structures as I described in Chapter 1.B.1. Evaluating the ability of computation tools to recover interactions that confer specificity as in section F (above) provides mechanistic insight into the performance of computational tools. However, the set of nucleotides in a structure form a set of proven instances of specific, high affinity binding. In section 1 (below), I use the set of RBPs binding single stranded RNA (Table

3.2) as a test set for specificity calculations. In many of the cases the specific interaction will be maintained through highly specific interactions. The predicted energies with nucleobase or residue mutations are indicative of the effectiveness of the scoring functions in specificity calculations.

Data about RBP binding motifs in the form of a position weight matrix (PWM) have been obtained for a small number of RNA binding proteins relative to the set of characterized RBPs. Modern techniques such as NMR scaffold independent analysis (NMR-SIA) and the technique employing a cDNA microarray (RNAcompete) provide high quality data about RBP specificity (Beuth et al., 2007; Ray et al., 2009). I reviewed the experimental approaches to obtaining binding motif data in Chapter 1.B.2.b. While these high quality techniques do not currently provide a scalable mechanism to experimentally test RBP specificities, the experiments provide good tests for our specificity predictions.

The experiments yielding detailed binding motifs were recently compiled into the RNA binding protein database (RBPDB) (Cook et al., 2010). Data from the RBPDB may be cross-referenced with the PDB to find examples for which we have both structures and PWM data. This subset of data allows for a more detailed examination of the performance of computational approaches to predict binding preferences from structure data.

The determination of binding specificity is of central importance to understanding post-transcriptional gene regulation, but progress on this challenge has been limited by the difficulty of validating specificity predictions. Validation against experimental measurements of binding specificity is an important assessment of the

performance of the scoring functions and of the algorithmic approaches to design. I used the limited amount of known structures to assess the performance of specificity prediction approach using the two empirical scoring functions and the Rosetta scoring function.

1. Recovery of RNA and protein sequence at RNA-protein interfaces

In order to investigate whether our computational energy scores can recapture the sequence information content of an interface, I substituted interface residues one at a time and calculated the score of the modified base or residue. By substituting all possible residues into each interfacial position, it is possible to calculate the most probable residue for a given position, as predicted by the design algorithm, and this prediction can be immediately compared with experimental results. As an initial test of these specificity prediction capabilities, I examined how well the sequence of the protein and RNA at the interface is recovered by the different scoring functions.

a. RNA sequence recovery

The recovery of the correct nucleobase at a protein interface is a good test of the selectivity of the potential with regards to nucleic acid bases. Thus, I performed a simple base recovery test with the two statistical potentials and with Rosetta, where I simply substituted each base one at a time at every possible position in the interface and calculated the predicted energy. This test is strict and undoubtedly under-estimates recovery rates, because it does not account for positions whose identity is not constrained by specificity requirements, nor does it restrict the calculations to bases directly (as opposed to indirectly) recognized by amino-acid residues.

I performed the base recovery tests with all three scoring functions selecting identical base positions and applying the same side-chain interaction cutoff and repacking conditions. The heat maps in Figure 3.5A show the probability of predicting each base given the specified based in the original structure. In order to prevent over-training, I executed the base recovery tests with scoring matrices where distance counts from any sequence within the training set with a high sequence similarity to the test structure were removed.

The rates of recovery of the actual base are compared for the three scoring function (Figure 3.5B). As a simple quantitative metric, I used the Hamming distance which representing the probability that an incorrect base is chosen in a given structure (see Appendix 2.B). The average normalized Hamming distance with Rosetta, the all-atom and generalized scoring functions were 0.45, 0.64 and 0.63 respectively, where 0.75 would indicate chance and 0.0 would be a perfect score. Thus, Rosetta significantly out performs the other two scoring functions. More notably, the Rosetta scoring function is more consistent with its predictions across the bases, while the all-atom scoring function is extremely sensitive to the composition of the training set.

A fairer assessment of the scoring function performance in the base recovery test would make allowance for some incorrect ordering of the base preference. Bases may be mispredicted for structural reasons or because of small errors in energy prediction. Some binding sites may select for a nucleobase using steric constraints. The purines (A and G) and the pyrimidines (C and U) are differing greatly in molecular size. A transversion is a substitution of a purine for a pyrimidine or *vice versa*. A switch between nucleobases with rings of like aromatic ring size is a transition. A site selecting

bases on size may be tolerant of a transition but not tolerant of a transversion. Additionally, calculated base energy scores are often similar, so slight imperfections in the interatomic scores may lead to a mis-ordering of bases.

Figure 3.6 visualizes the base recovery analysis from Figure 3.5 in terms of recovery range using a cumulative distribution function (CDF) plot. If transitions were allowed at many positions we would expect the correct base to receive a calculated rank of 2 or better. The rank 2 or better recovery rates are 0.71, 0.62 and 0.59 (Figure 3.6). The small increase in cumulative recover rank 2 or better values for Rosetta from the rank 1 of 0.57 suggests that the mis-ordering are not significantly affected by non-selectivity of the position. With only for possible bases with similar chemical properties the CDF plot will only reveal a strong effect of non-selectivity at a position recognized by an RBP.

A limitation of this approach to testing this outcome of the simulation is that the recovery of the correct base may be significantly underestimated. In this strict approach, I have not included a strong requirement for significant base contacts that will result in forcing a base selection where the base identity is unconstrained. If the scoring functions were completely accurate, I would expect the scoring function to discriminate potentially ambiguous sites by assigning to each possible base substitution a similar score. The information content metric ($\sum_i p_i \log p_i$) should then correlate with certainty about a base position with higher information content implying greater certainty. However, I did not find that positions with higher calculated information content were more likely to yield a correct base prediction (Figure 3.7).

I executed this control because Rosetta was shown in the case of peptide binding proteins to perform well in predicting the information content of the position in a bound peptide (King & Bradley, 2010). Given a correct scoring function, I expected that sites with high information content, reflecting meaningful energy differences, would be more likely to be correct. That is to say, sites with a strong physical preference for one base would not be subject to mis-predictions due to minor errors in the scoring function. I observed instead the opposite result, suggesting that there are still significant limitations in the scoring functions, or that structure-space is sampled insufficiently.

b. amino acid sequence recovery:

The inverse problem of protein side chain recovery is more indicative of the sequence interface constraints and a direct test of the situation relevant to RBP re-design. Indeed the correct identification of which protein side-chain or combination of amino acids recognize a nucleotide sequence is a pre-requirement for any protein design application. Thus, I performed the amino acid recovery test with the two statistical scoring functions and with Rosetta in order to evaluate whether any of these approaches would be of value in design applications.

I performed the base recovery tests with all three scoring functions selecting identical residue positions and applying the same side-chain interaction cutoff and repacking conditions. Repacking and scoring was repeated 20 times following replacement of the candidate residue. The heat maps in Figure 3.8A show the energetically preferred amino acid side-chains recovered using each scoring function for positions corresponding to each residue in a structure. The map also reveals which

contextually incorrect residue is suggested at each position. Both pure statistical functions predict ARG and TRP as a common replacement for most other residues. The selection of ARG is not surprising as its positive charge makes it a residue very frequently found at RNA binding sites, while the TRP prediction is understandably unreliable given its low representation in both the training set and the test set (Chapter 2.E.2).

The correct recovery of each residue is more easily compared in the bar plot of fraction correctly recovered by structure residue (Figure 3.8B). Rosetta outperforms the statistical potential in recognition for almost all of the residues. The performance can again be summarized using the Hamming distances (see Appendix 2.B), where 0.95 represents chance and zero represents perfect recovery. The Rosetta, all-atom and generalized scoring functions had Hamming distances of 0.67, 0.79 and 0.77, respectively.

In this test, I expect some replacement of residues by residues with similar chemical properties. The cumulative distribution plot (Figure 3.9) compares the performance of the scoring functions by the rank assigned to the correct residue. I described the CDF plot in section a (above) with respect to base substitution. However, the CDF analysis is more informative in the residue recovery test since many of twenty standard amino acids share functional groups and chemical properties. Evolution has demonstrated that proteins often tolerate residue substitutions to similar residues. Rosetta has a ~70% chance of assigning a score in the top 3 to the correct residue. The top 3 recovery data in Rosetta scores suggests that the recovery predictions may reflect a physical tolerance for a chemically similar residue at a position.

I checked for chemical and structural similarity between the residues in the top three ranked positions at positions where the correct residue is found within the top three. When the residues along the axes are sorted by chemical property of the side-chains grouping hydrophobic, polar, charged and aromatic residues, the resulting heat maps of predicted with actual residues (Figure 3.10A) show most incorrect substitutions fall close to the diagonal. This analysis suggests that with all three scoring functions, the best scoring residue is chemically similar to the native structure residue.

The ordering of the residues on the axes of the residue recovery heat maps (Figure 3.10A) can be compared to the natural base substitution explored through evolution. Most permissible base substitutions observed in homologous proteins (orthologs and paralogs) are substitutions that have no effect on protein function. Thus substitution matrices, such as the PAM250 matrix, from genome biology largely capture the frequency with which side-chains sharing functional groups may be interchanged (Wilbur, 1985). Figure 3.10B shows a heat map of the PAM250 substitution matrix with residues in the same order as (Figure 3.10A). We see that the evolutionarily expected substitution rates are similar to the predicted permissible substitutions using the scoring functions. The similarity supports my contention that the scoring functions are correct enough to identify reasonable substitutions.

A related explanation for mispredicted interface residues is that the position does not participate in specific binding. A position that is unconstrained may not participate in energetically favorable or unfavorable contacts with the RNA base.
As discussed in section a (above), the information content (IC) metric should correlate with certainty about a residue position with higher information content implying

greater certainty. Figure 3.11 shows how the top *x* positions perform with recovery of the correct residue. We see that the Rosetta scoring function does slightly better with residues for which the calculated substitutions imply a high information content. The IC test is a strict test in that it uses the residue scores both to assign information content and to predict the scores. The lack of a good correlation implies that scoring function is not completely capturing specific interactions.

c. recovery test summary

Together, the base and amino acid recovery tests query the extent to which the scoring function captures sequence restraint imposed on a position by neighboring sequence and structure. This assumes that proteins bind to RNA like puzzles and each position has been optimized though evolution to fit into that position. This assumption is most certainly too strong, but was required to make a useful comparison in situations where structure is our only source of experimental information.

The tests were successful in so far as each scoring function more often than would be expected by random identified the structure base or amino acid as the preferred residue for that position. Additionally, since I expected the constraint between structure and sequence to be high, the Rosetta scoring function performed better in these tests indicate that its performance is indeed better. The residue recovery tests showed that especially with Rosetta the correct residue is likely to be among the top 3 and that that highest scoring residues are likely to be functionally related.

However, to better understand the performance of the scoring function in specificity prediction, we need to correlate the structure information with more direct

sources of specificity data from literature. The structure does not directly contain information about how strong the binding preference is for one base over another or rank the preferred bases. The CDF plots gave some clues as to how the predicted and structure residues agreed beyond the top prediction. However, the structure information is insufficient to fully evaluate the relationship between structure and specific recognition. Information from experimentally determined binding motifs would allow a direct evaluation of the specificity predictions.

2. Direct Comparison of Predicted and Experimental RNA Binding Motifs

The tests conducted above do not fully capture the ability of scoring function to model the interaction energy at a protein-RNA interface. With an accurate scoring function, the computational approach allows us to evaluate the order and magnitude of the preference for each base. This specificity information may be exploited to quantify binding sequence specificity and to provide clues as how the specificity could be switched at binding position.

We wish to evaluate the extent to which the scoring functions capture specificity of the binding interaction (Ashworth & Baker, 2009), because this specificity reflects how important a base site is to the the recruitment of RBP to that overall sequence. A computational approach could allow us to gain a more detailed view of the recognized RNA sequence than can be provided by simple a consensus sequence recognized by a given protein. Thus, the ability to rank candidate binding-sequences by specificity and affinity is a key step in being able to discover the mRNA binding targets of RBPs that form the basis of post-transcriptional regulatory networks (Mansfield & Keene, 2009).

The prediction of RNA binding motifs is a much more rigorous test of the scoring functions.

a. Prediction of sequence specificity landscapes

A complete characterization of an RBP specificity landscape would require predicting the relative binding energy of all possible sequence substitutions at an interface, and comparing these predictions with experimental measurements. However, studies with DNA have found that base recognition is mostly captured by site-specific base preferences. In the case of DNA binding sequences, higher order constraints on base sequence do not generally extend beyond adjacent base pairs (Y. Zhao & Stormo, 2011). Thus, each base position can be considered to be independent to a good approximation.

In the case of RNA, the tertiary folded structure of many functional molecules increase the probability that adjacent and non-adjacent base positions may create a higher order sequence constraint or affect the binding affinity of a RNA sequence (Xiao Li, Quon, Lipshitz, & Morris, 2010). However, this problem is beyond the scope of current computational approaches and a much less significant problem for the single-stranded RNAs that are the subject of this investigation. Thus, we will assume positional independence.

A common representation of the sequence binding preference of a protein, as generated using these methods, is the position weight matrix (PWM). The PWM represents the probability of recognizing a specific base at a nucleic acid position (Stormo, 2000). Assuming that positions are recognized without dependence on neighboring bases, the relative affinity for binding a sequence can be immediately

derived from the PWM. A PWM may also be easily visualized and compared using the sequence logo representation (Schneider & Stephens, 1990)

b. Validating against experimental position weight matrices

The RNA Binding Protein Database (RBPDB) provides the first comprehensive survey correlating sequence specificity data with sequence and structure information about RBP's (Cook et al., 2010). However, only a small subset correlating 30 unique proteins with data from SELEX (Tuerk & Gold, 1990) or RNAcompete (Ray et al., 2009) contains sufficient information to construct a binding site PWM. Of this set, only 20 proteins have structural information, and only eight proteins have both an experimental PWM and a solved structure of the relevant protein-RNA complex. Thus, the set of examples available for a complete comparison between experimental results and modeling is not large.

Nonetheless, the structures solved as complex provide a first basis for assessing the ability of the scoring function to reproduce the correct base preference at any position. Since we wanted to establish if any calculations could recover experimental base preference, we used all solved complex structures for the eight proteins as starting points. Table 3.4 lists the structures with information about the Pfam fold family (Finn et al., 2008) and structure quality. We note that six of the eight proteins for which a complex is available are in the Pfam RRM_1 family (PF00076), and therefore the sample of structures is relatively homogenous but is defined from the most abundant class of RNA-binding proteins.

I performed the calculations in the same manner as the preceding RNA sequence recovery tests. I included all models from ensemble of NMR structures and independent

x-ray crystal structures to increase the structure space searched to include near native backbone conformations. At each interface base position, I substituted each possible base and repacked all neighboring amino acid residues with a side-chain heavy atom within 6 Å of a base heavy atom. The sequence and structure space was explored identically with all three scoring functions.

With the Rosetta calculations, I performed additional minimization steps allowing the optimization of rotameric dihedral angles. The minimization step could not be performed with the statistical scoring functions because the scores are discrete and are not differentiable. This minimization step greatly improved base scores reported using the Rosetta scoring function. If the all atom function were differentiable, the score improvement with minimization would not be as significant because of the discretely binned interactions of this function.

I performed each base substitution 20 times in the context of each model. This resulted in base scores that reflected the local minima for the scores of the surrounding side-chains. I converted the sampled scores to a probability of binding at that location using a sampled competition approach as described in methods and Figure 3.1. Base order was consistent with that which was reported based on the alternative minimum score method.

I used the PWM data from the RBPDB to benchmark the scoring functions against experimental position weight matrices. Comparing PWMs yields more information about the recovery of natural binding preferences than comparing sequences, since both the preference order for recognized bases as well as the magnitude of the preference is considered.

We can judge the relative performance of the potentials by visually comparing the sequence logos (Schneider & Stephens, 1990) of the predictions with each scoring function against the set of experimental results. For example, Figure 3.12 shows the performance of the scoring potentials for small nuclear ribonucleoprotein A (U1A) comparing a single PWM from a SELEX experiment with a calculation based on the x-ray crystal structure 1urn. An example of aligned PWM predictions for each remaining protein with each scoring function is shown in Figure 3.14, Figure 3.15, Figure 3.16 and Figure 3.17. For each protein, there were often more than one experimental PWM and more than one structure was used as a basis for PWM calculations. Only the best experiment and calculated PWM pair is shown in Figure 3.14, Figure 3.15, Figure 3.16 and Figure 3.17, but all such pairs are used in quantifying the comparison as shown below.

The experimental data represent a heterogeneous set consisting of PWM estimations from multiple experiment types and potentially different constructs. For each protein, we aligned base positions in the NMR and x-ray structures to the experimental motifs by finding the motif frame in which the sequence from the structure yielded the highest probability score. For each pair of aligned experimental and calculated motifs, we calculated the average per-position difference using the Euclidian distance (ED) metric (section E.2 above).

I report the average difference between all experimental and calculated PWMs for each gene using each of the three scoring functions (Figure 3.13). As can be inferred visually from the representative sequence logos, the Rosetta scoring function outperforms the purely statistical scoring function for all proteins except PBPC1, a

structure for which there is only one experimental binding set of results and one structure. The generalized potential performed better than the all atom scoring function in all instances, except ZRANB2. The better performance of the generalized scoring approach suggests that the increased size and resolution of the training set may be critical to capturing the interface detail needed for specificity calculations.

The results of this test demonstrate that the Rosetta scoring function substantially captures the order and information content of base specificity in the experimental data. We should note that only two of the experimental results were obtained using the RNA compete microarray method. There is likely significant error in the information content reported for the rest of the experimental PWMs from heterogeneous sources. However, the performance of the Rosetta scoring function suggests that it is ready for tests with design tasks. In the next section, I apply all the scoring functions to the task of predicting the specificity of experimentally designed PUF domains, and concentrate on results using the Rosetta scoring function.

H. Design Test with Pumilio1 Protein

Pumilio and FBF homology (PUF) repeats are attractive domains for use as a modular recognition platform for the specific recognition of bases (Cheong & Tanaka Hall, 2006; Xiaoqiang Wang, McLachlan, Zamore, & Tanaka Hall, 2002). PUF is an attractive domain for design because each repeat makes predictable contacts with one and only one RNA base. Recognized bases are generally sandwiched between aromatic and positively charged side-chains on the α-helical regions of adjacent repeats, and amino acid side-chains at conserved positions make contact with the base edge. The

regularity of its interaction has made the PUF domain a favorite candidate for creating proteins to bind to arbitrary RNA sequences.

While a basic code for RNA recognition by PUF domains had been described (Xiaoqiang Wang et al., 2002), it had not been shown that by altering one or two residues the specificity of a domain could be modified to recognize all four canonical bases. Specifically, rationally altering a protein to recognize cytosine had proven difficult (Cheong & Tanaka Hall, 2006). Thus, achievement by selection experiments of a PUF domain repeat from Pumilio1 capable of recognizing any of the four canonical nucleobases using only small sequence changes was unique. The completeness of the specificity switch at a single PUF repeat provides the basis for an interesting test case for the computational approach.

Recently, two groups have demonstrated and provided structural evidence that a single PUF repeat may be used to recognize any of the four canonical RNA bases by performing substitutions of only two residues that contact the RNA base edge (Y. Chen & Varani, 2011; Dong et al., 2011; Filipovska, Razif, Nygård, & Rackham, 2011). The residue positions of Pumilio1 repeat 6 (PUM1 R6) that allow retargeting to each of the canonical bases are illustrated in Figure 3.18. The change in specificity with only two substitutions makes this the experimental analog of the basic computational retargeting problem and provides an experimentally well-defined system in which to validate my computational approach.

1. Computational Approach to PUM1 Specificity Switch

I tested whether the computational method could be used to recapitulate the shift in base specificity caused by the set of two residue mutations investigated by Dong

et al. (Dong et al., 2011). Mutations of residues at positions A1043 and A1047 in PDB structure 1my8 have been shown to alter the binding specificity in PUM1 R6 from uracil to each of the other three bases (Dong et al., 2011). The structure of the native human PUM1 domain (1m8y) identifies contacts between the native residues N and Q and the uracil base edge. The PUM1 mutant structure 2yjy, mutated at A1043 and A1047 to SER and ARG respectively (henceforth referred to as the SR mutant), has experimentally verified preference for binding cytosine (Dong et al., 2011). The native (1m8y) and SR mutant structures can both be used as starting point for the computational prediction of the position weight matrix (PWM) column for the base position recognized by PUM1 R6.

The reported PUM1 R6 mutants involved in a specificity switch all involve single or double mutations to PDB positions A1043 and A1047. Since simulating the entire set of double mutants at two positions is computationally tractable, I tested all double mutants but focus my analysis on the subset of residue swaps for which experimental results are reported. I refer to each tested residue swap by a pair of one-letter-residue codes of the amino acids replaced at positions A1043 and A1047. Starting with each the two copies in the crystal unit of each of these structures, I used our scoring approach to predict the base preference with each of the residue combinations for which base specificity has been experimentally tested (AR, CQ, CR, GR, NQ, SE, SR, TR) as well as all substitution pairs except those involving PRO at these positions. The PWM column representing the relative binding preference for each of the canonical RNA nucleobases can be determined as performed above (section G.2.b).

The double mutants to PUM1 R6 can be simulated with a conservative approach concentrating on side-chain conformations because these mutations do not significantly

alter the backbone structure of the protein or RNA. The conservation of the backbone can be inferred from structure alignment of PUM1 R6 with the native NQ sequence binding U in structure 1m8y and the SR mutant binding C in structure 1yjy. I use each structure as a starting model for prediction of binding preference with the other possible double mutants.

I pack only the residues and bases participating in recognition. I allow only the two specificity switch residues to repack and a few degrees of rotation of the target base about the glycosidic bond in the repacking procedure. Each base substitution and repacking calculation was repeated 50 times with each pair of residues substituted. As in the base specificity calculations, the base preference is determined through the sample competition approach, where one of the 50 base repacking scores is drawn for each base (as described in D.2 above and shown in Figure 3.1). The PWM prediction differs from the previous approach in that base scores from the same substitution using the different starting structures 1m8y and 1yjy are combined for the sampled competition. The predicted base probability is the fraction of simulated competitions won by that base. The use of different starting structures allows the test to be completed on a set of valid backbone structures and can be thought of as incorporating some of the effects of local changes in the protein backbone.

2. Target Switch Results

I performed the PUM1 R6 nucleobase preference test for all double mutants with all three scoring functions. The all-atom and generalized scoring function correctly predicted some base preference switches, but exhibit the same prediction biases we saw in the base recovery test and the logo recovery tests (section G above). The

empirical scoring functions continued characteristic mis-predictions including an over prediction of U for all-atom scoring function. In the remainder of this section, I concentrate on the Rosetta scores since they more closely approximate the experimental results and provide the best opportunity for understanding base prediction in this well-defined structural model.

When the base preference prediction is performed with the Rosetta scoring function on each of the four structures, our approach captures many of the altered specificity of the PUF domain variants (Figure 3.19). Many of the transitions to C are predicted and the calculations based on the 2yjy structures capture the preference for A with the CQ substitution and G with the SE substitution. Calculations based on the 1m8y structures almost never accommodate a purine and have a stronger preference for its native U with all side-chain sequence substitutions. This suggests that the subtle variation in the backbone between calculations starting with 1m8y and those starting with 2yjy is important for selectivity and for accommodating larger bases.

If instead I consider the four structures as valid backbone models, as we did within structures in the motif recovery calculations, I recover all but one base preference. Figure 3.19 shows the results using all 200 (50 trials x 4 structures) of each base substitution. Base switches at the PUM1 R6 position are observed in many of the 361 (19×19 pairs of non-PRO mutations) residue switches tried. For the set of double mutants for which we have experimental knowledge of the binding preference, the computational predictions almost always recover the altered base preference (Figure 3.19).

The native NQ sequence of PUM1 R6 results mis-predicts a preference for A over the native U. The high preference for A and G with the native NQ sequence can be rationalized by looking at the best predicted scoring structure with each of the bases. Both N and Q are predicted to make favorable hydrogen bonds as donors with A or as acceptors with G. I suspected that both the incorrect hydrogen binding pattern with U and the failure to predict a purine (except in the case where A1043 was ALA, GLY or an aromatic residue, data not shown) using the 1m8y structures could be attributed to the Lenard-Jones (LJ) repulsive term. When I repeated the calculations with a diminished LJ repulsive term, I recovered the native hydrogen-binding pattern with U, though the predicted preference remained for A and G.

The results of the analysis of the binding specificity of PUM-1 demonstrate some of strengths and weaknesses of a structure-based approach to specificity prediction. Using established and empirical scoring terms, we can reproduce the magnitude and order of preference for binding bases at a specific position. Furthermore, I can recapitulate experimentally demonstrated specificity shifts in this system. However, the PUM1 R6 case is a very specialized case where two residues are positioned perfectly to contact the base edge and the RNA interface is highly constrained by stacking interactions with residues on the PUF domain α helices. A major part of the problem in applying this approach to other systems is in recognizing cases where the RNA interface is consistent and in identifying the key residues that confer specificity.

Ⅰ. Summary

I benchmarked the performance of three scoring functions for predicting binding specificity. The scoring functions included the all-atom distance-dependent scoring

function, the generalized scoring function trained on small molecule interactions, and a Rosetta scoring function with weights optimized for transcription factor structure prediction. Base and residue recovery tests showed that all three scoring functions were significantly more predictive of base and residue identity constraints than what would be expected from a random selection of bases or residues of equal probability.

A microarray based specificity assay and a recent survey of specificity data in the literature allowed a more detailed analysis of how each scoring function captured the information content of binding positions, but the set of structure for which there is a solved complex structure and for which detailed binding motif data are available remains small. Calculations on this set of structures demonstrated that the statistical scoring functions performed poorly in predicting the position weigh matrix representing the binding motif. The Rosetta scoring function predicted half of specifically recognized RNA positions and captured the information content of many of those positions. We expect that it will be possible to extend these results to homology models of some of the structures for which there is binding profile data.

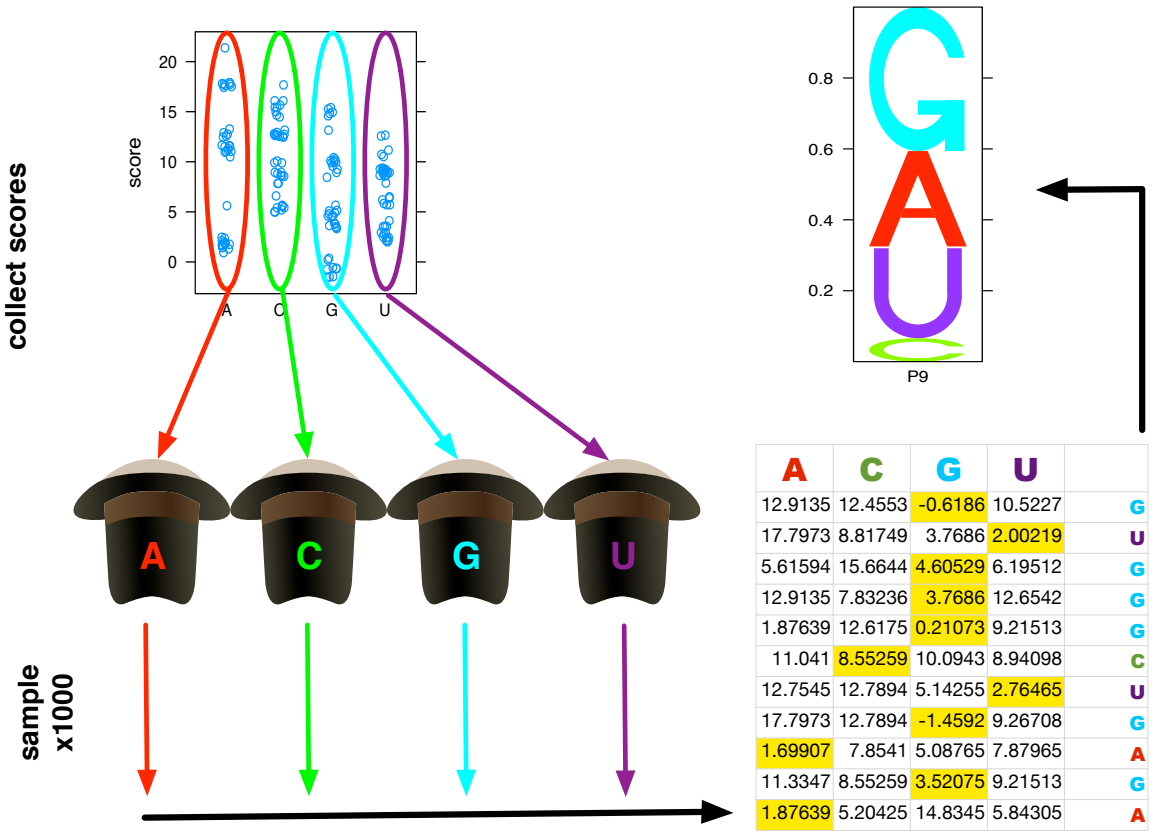
I performed a test of design application using experimental data from a PUF domain. Experimental selection of the PUF domain had revealed two amino-acid positions where selected mutation could retarget the position to recognize any of the canonical bases. I performed the same mutations using the Rosetta framework and scoring function. The calculated base preferences recapitulated the experimental preference in all except the native case. Predictions were made for the other possible double mutants. This test validates our hypothesis that RBPs may be retargeted to recognize non-native RNA sequences through a small number of changes to the protein

sequence. In the next chapter (Chapter 4 below), I will apply the lessons learned here to suggest mutations for a biologically relevant retargeting of an RRM domain.

Figures and Tables

Table 3.1: Component weights used for scoring applications with the Rosetta scoring function.
The following weights were taken from work by others that optimized the weights for correct structure recovery of protein-DNA interfaces

Rosetta term	weight	description
fa_atr	0.800	Lennard-Jones attractive
fa_rep	0.440	Lennard-Jones repulsive
fa_sol	0.650	Lazaridis-Karplus solvation energy
fa_intra_rep	0.004	Lennard-Jones repulsive between atoms in the same residue
fa_plane	0.000	pi-pi interaction between aromatic groups
fa_dun	0.560	internal energy of sidechain rotamers as derived from Dunbrack's statistics
ref	1.000	reference energy for each amino acid
hbond_lr_bb	1.170	backbone-backbone hbonds distant in primary sequence
hbond_sr_bb	1.170	backbone-backbone hbonds close in primary sequence
hbond_bb_sc	1.170	sidechain-backbone hydrogen bond energy
hbond_sc	1.100	sidechain-sidechain hydrogen bond energy
p_aa_pp	0.640	probability of amino acid at phi-psi
dslf_ss_dst	1.000	distance score in current disulfide
dslf_cs_ang	1.000	csangles score in current disulfide
dslf_ss_dih	1.000	dihedral score in current disulfide
dslf_ca_dih	1.000	ca dihedral score in current disulfide
pro_close	1.000	proline ring closure energy
hack_elec	0.500	Simple electrostatic repulsion term
omega	0.500	omega dihedral in the backbone



A	C	G	U	
12.9135	12.4553	-0.6186	10.5227	G
17.7973	8.81749	3.7686	2.00219	U
5.61594	15.6644	4.60529	6.19512	G
12.9135	7.83236	3.7686	12.6542	G
1.87639	12.6175	0.21073	9.21513	G
11.041	8.55259	10.0943	8.94098	C
12.7545	12.7894	5.14255	2.76465	U
17.7973	12.7894	-1.4592	9.26708	G
1.69907	7.8541	5.08765	7.87965	A
11.3347	8.55259	3.52075	9.21513	G
1.87639	5.20425	14.8345	5.84305	A

Figure 3.1: Virtual competition approach to calculating base probabilities at recognized RNA sites. Structure space is searched using a Monte Carlo search of local Dunbrack allowed side-chain conformations. Depending on the temperature of the simulation, the returned structure is usually a local minimum rather than the absolute structure minimum. Additionally, while the scoring functions used in this search likely scale with energy, we cannot easily discover the scaling factor for each of these approaches. Instead, I used a competition-based approach to discover the probability of binding each base from a sampling of local minima in an entropy dependent approach. All local minima scores for each base substitution are recorded. Using these values I simulate 1000 winner-take-all competitions. The probability of a base binding is taken as the fraction of virtual competitions won by that base.

Table 3.2: Test set of thirty representative RBP structures binding to single stranded RNA.
The analysis of RBP binding sites by Bahadur et al. (2008) investigated difference in the physical properties based for sets of protein binding RNA in a similar manner. One of the categories investigated was a diverse set of proteins binding to primarily single stranded regions of RNA. The table lists descriptions and Pfam classifications of the proteins. I used these proteins as a basis for interface repacking and base recovery tests.

PDB ID	structure	Pfam Acc [Res Range]	Pfam ID	res (Å)	PubMed ID
1a9n	spliceosomal U2B"-U2A' protein complex	PF00076 [9-80]	RRM_1	2.38	9716128
1av6	vaccinia methyltransferase VP39	PF01358 [2-294]	PARP_regulatory	2.80	9660928
1cvj	poly(A)-binding protein	PF00076 [13-169]	RRM_1	2.60	10499800
1g2e	HUD and AU-rich element	PF00076 [5-159]	RRM_1	2.30	11175903
1jbs	ribotoxin and restrictocin	PF00545 [24-147]	Ribonuclease	1.97	11685244
1jid	human SRP19	PF01922 [16-116]	SRP19	1.80	11641499
1k8w	e. coli pseudouridine synthase TruB	PF01509 [46-193], PF09157 [265-322]	TruB_N, TruB-C_2	1.85	11779468
1knz	asymmetric NSP3 homodimer	PF01665 [5-164]	Rota_NSP3	2.45	11792322
1kq2	Hfq-RNA complex	PF01423 [5-65]	LSM	2.71	12093755
1lng	SRP19-7S.S SRP RNA complex	PF01922 [1-87]	SRP19	2.30	12050674
1m5o	U1 snRNP A	PF00076 [12-83]	RRM_1	2.20	12376595
1m8v	snRNP SM-like protein	PF01423 [9-74]	LSM	2.60	12409299
1m8w	Pumilio-homology domain from human Pumilio1	PF00806 [21-314]	PUF	2.20	12202039
1n35	Lambda3 elongation complex	PF07925 [1-1263]	RdRP_5	2.50	12464184
1wpu	Hutp antitermination complex	PF09021 [7-145]	HutP	1.48	
1wsu	C-terminal domain of elongation factor SelB	PF09107 [71-122]	SelB-wing_3	2.30	15665870
1zbh	3'-5' exonuclease ERI1	PF02037 [26-60], PF00929 [80-256]	SAP, RNase_T	3.00	
1zh5	Lupus La protein	PF05383 [18-76], PF00076 [114-182]	La, RRM_1	1.85	16387655

Table 3.2 continued.

PDB ID	structure	PFAM Acc [Res Range]	Pfam ID	res (Å)	PubMed ID
2a8v	RHO transcription termination factor	PF07498 [5-47], PF07497 [49-118]	Rho_N, Rho_RNA_bind	2.40	10230401
2anr	NOVA-1 KH1/KH2 domain	PF00013 [7-169]	KH_1	1.94	21742260
2asb	tuberculosis NusA-RNA complex			1.50	16193062
2b3j	tRNA adenosine deaminase	PF00383 [4-105]	dCMP_cyt_deam_1	2.00	16415880
2bx2	catalytic domain of E. coli RNase E	PF00575 [42-126], PF10150 [128-400]	S1, RNase_E_G	2.85	16237448
2db3	DEAD-box protein	PF00270 [80-253], PF00271 [318-395]	DEAD, Helicase_C	2.20	16630817
2f8k	SAM domain of VTS1	PF07647 [13-75]	SAM_2	2.00	16429151
2g4b	U2AF65 variant	PF00076 [7-81, 97-167]	RRM_1	2.50	16818232
2gic	vesicular stomatitis virus nucleocapsid-RNA complex	PF00945 [10-405]	Rhabdo_ncap	2.92	16778022
2i82	pseudouridine synthase RluA	PF00849 [20-168]	PseudoU_synth_2	2.05	17188032
2ix1	rnase ii d209n mutant	PF08206 [44-101], PF00773 [209-537], PF00575 [577-660]	OB_RNB, RNB, S1	2.74	16957732
2j0s	ATP-dependent RNA helicase DDX48	PF00270 [61-227], PF00271 [294-371]	DEAD, Helicase_C	2.21	16923391

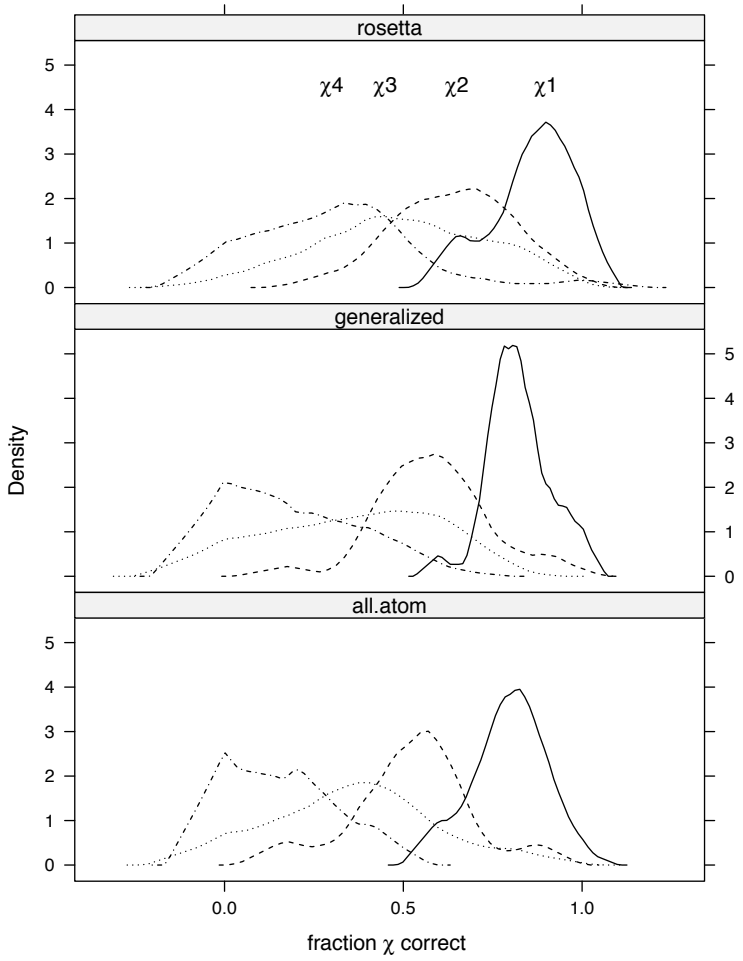


Figure 3.2: Density plots summarize fraction of side-chain dihedral (χ) angles recovered by structure. The fraction χ correct is the number of structures tested for achieving the each $\overline{\chi}_i$. Since the recovery is cumulative, dihedral angle χ_i has a maximum recovery rate of χ_{i-1}. The density plot was created using from the Lattice package in the R Project (Sarkar, 2008, pp. 35–54). The edge effects are due to the smoothing by the Gaussian kernel.

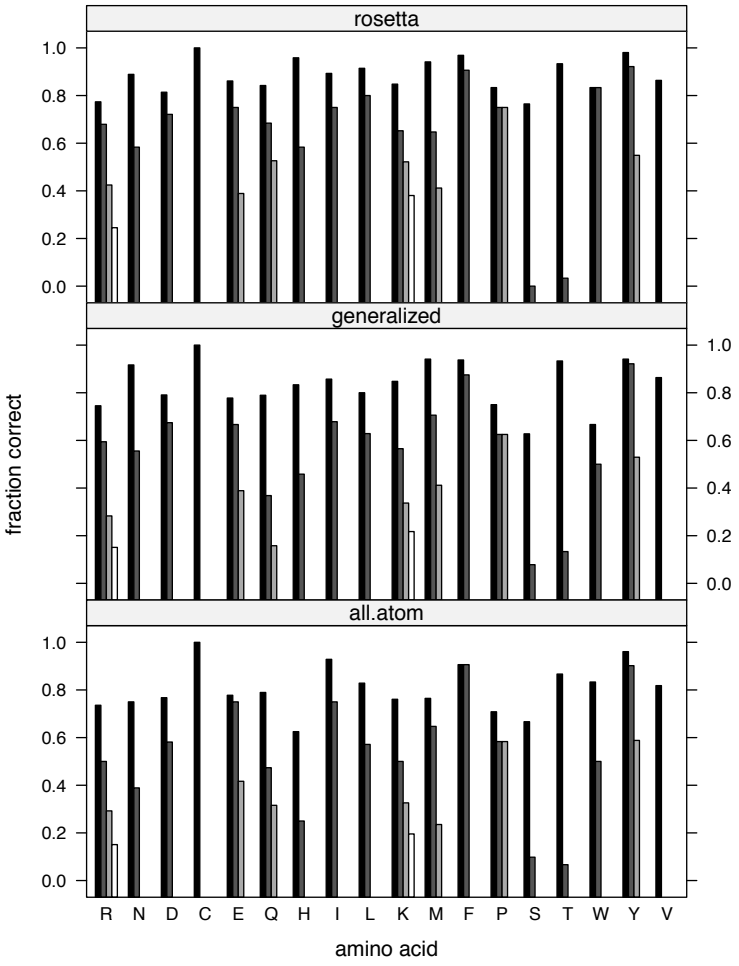


Figure 3.3: Recovery of correct side chain conformations at protein-RNA interfaces by scoring function. Interfacial amino acid side-chains in the set of 30 representative high-resolution crystal structures were simultaneously repacked with the all-atom distance-dependent statistical potential. Each packing-run was independently repeated several times and recovery statistics were taken from the run that provided the lowest calculated interface energy. Bar plot shows cumulative side-chain dihedral angle recovery. Bars are gray-scaled from χ_1 (black) to χ_4 (white). Recovery fraction is cumulative such that $\max(\chi_i) = \chi_{i-1}$.

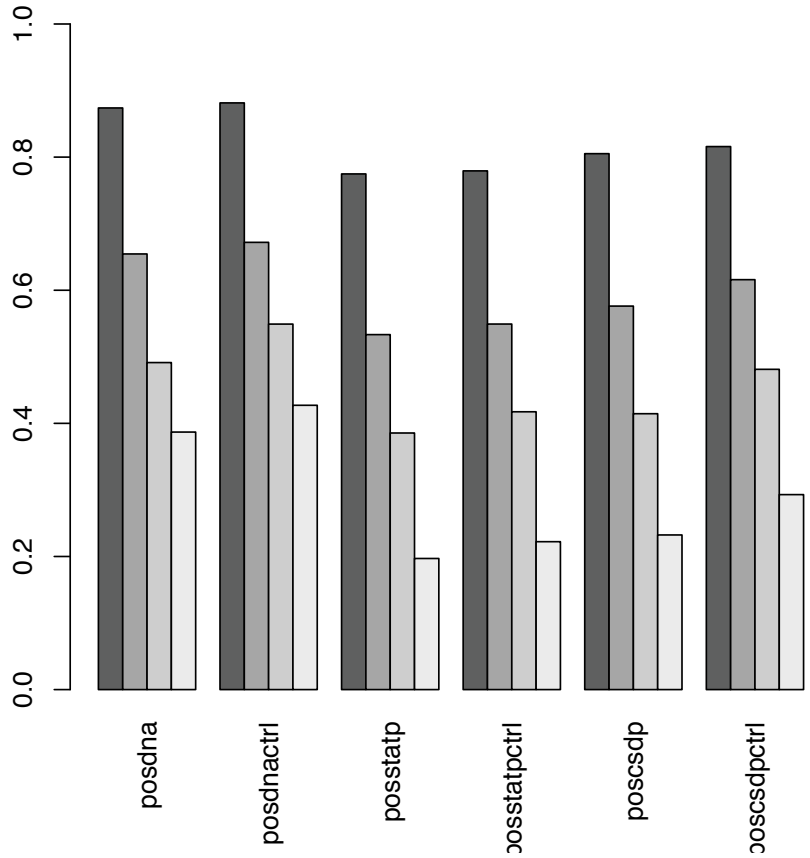


Figure 3.4: Summary of side-chain recovery using Rosetta and the all atom statistical potential. This figure shows the cumulative fraction of each dihedral angle χ_1 (black) to χ_4 (light-grey) for all side-chain types where $\max(\chi_i) = \chi_{i-1}$. Packing tests were performed with Rosetta (**dnap**), the all-atom potential (**statp**) and the generalized potential (**csdp**). Entries labeled with the suffix **–ctrl** indicates that the native side-chain angles were included in the rotamer test set. Bar plot shows the recovery in the simplest case where each interface side-chain (within 6 Å of the RNA) is packed while maintaining all other side-chains fixed.

Table 3.3: Recovery of different types of intermolecular contacts observed in the native structure. Recovery fraction was found following re-packing of amino acid side chains at the protein-RNA interface. The test recovery test was performed on the test set of thirty RBPs binding single stranded RNA (Table 3.2). Only interactions between residue side-chains (sc) and RNA are included in the table, since only these contacts may be lost through repacking. The percentage recoveries summarize the percentage recovery for all contacts of the indicated type (hydrogen bonds, cation-π or π- π) across all sub-categories. The sub category values are reported as number out of the total of that sub-category. The cation-π interactions in the test set are dominated by arginine side chains and are often replaced by the formation of non-native hydrogen bonds.

interaction type sub-category	**all-atom scoring function**	**generalized scoring function**	**Rosetta scoring function**	**total**
hydrogen bonds	43%	36%	49%	
sc-base	44	38	51	106
sc-sugar	29	23	30	57
sc-phosphate	43	37	52	105
cation-π	32%	40%	32%	
LYS-base	0	0	2	3
ARG-base	8	10	6	22
π-π	81%	86%	86%	
PHE/TYR-base	29	29	31	34
TRP	0	1	1	1
HIS	5	6	4	7

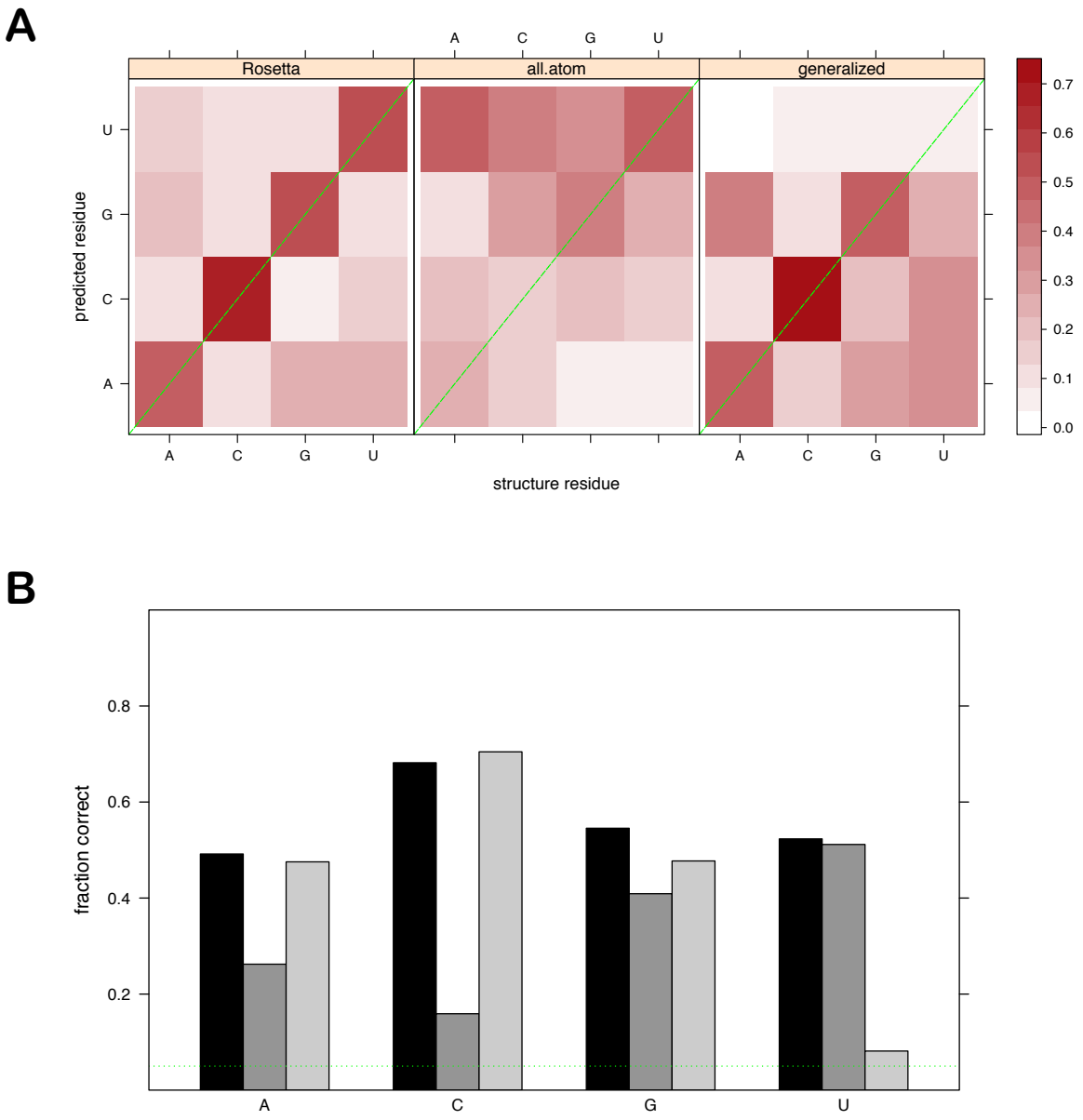


Figure 3.5: Comparison of base specificity recovery for RBP recognizing ssRNA with all scoring functions. **A** The heat maps represent the probability of each possible RNA base being preferred by the scoring function at a position in a structure originally containing the specified base. The green diagonal intersects cells representing the rate of recovering the correct base (self-recovery). **B** The bar plot facilitates the comparison of the performance the three scoring function for base self-recovery (Rosetta: black; all-atom: grey; generalized: light grey).

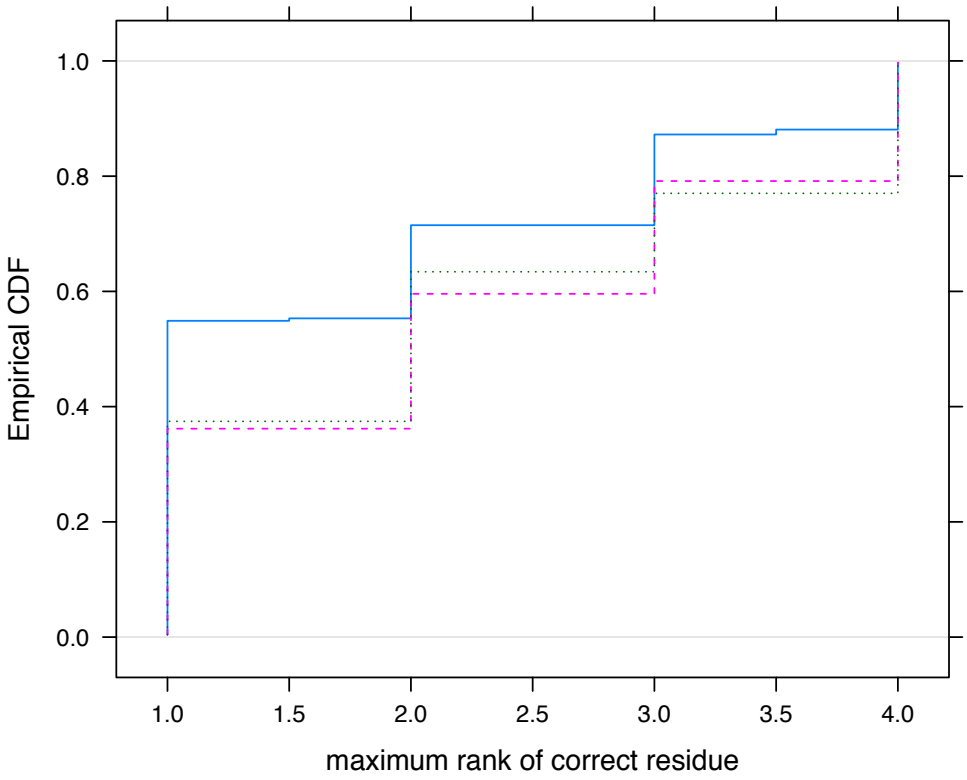


Figure 3.6: Probability of recovering the correct base at each rank or better with each scoring function. The empirical cumulative distribution (CDF) plot gives us a closer look at the preference rank assigned to the correct base at each base-recovery position by the scoring functions. The plot shows the fraction of bases for which the structure base receives a preference rank less than (better) or equal to the rank index on the x-axis. In the infrequent case where *n* possible base substitutions receive the same score, the residues receiving the same score are assigned a rank equal to the average of the next *n* rank indexes (solid cyan: Rosetta; dashed magenta: all atom; dotted green: generalized).

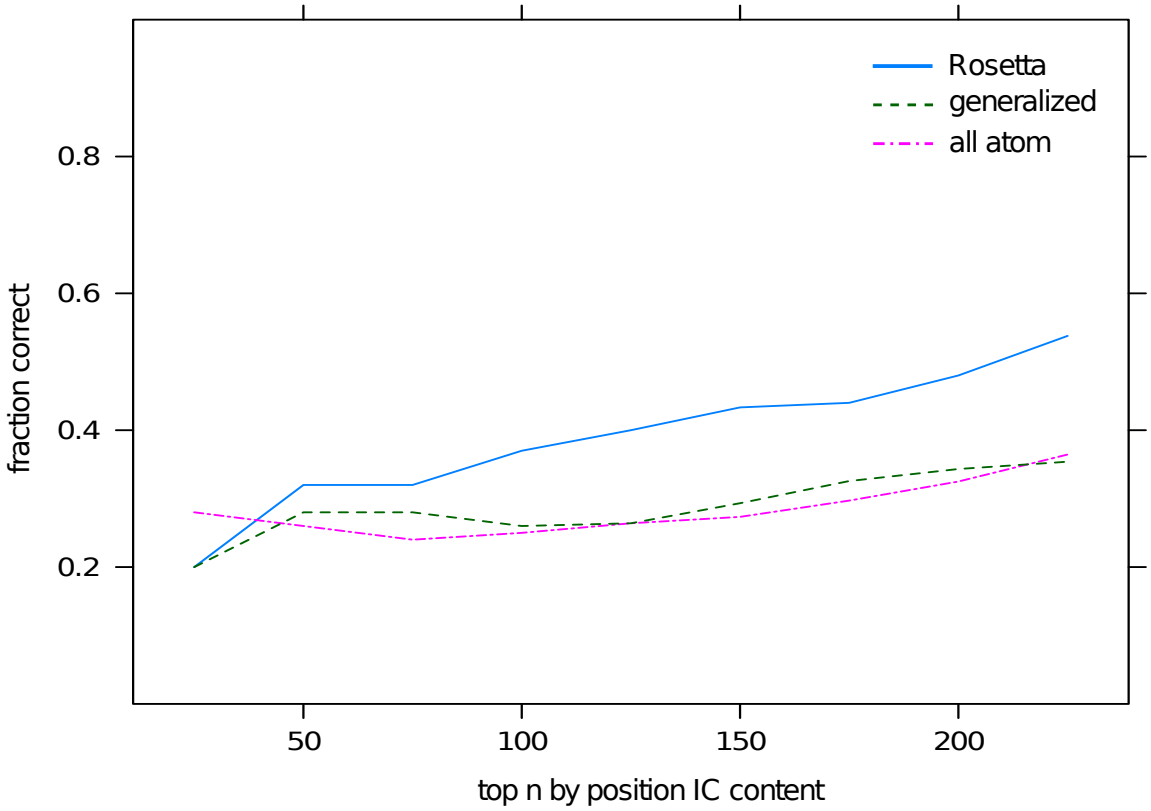


Figure 3.7: Fraction of bases correctly recovered for positions with the strongest predicted preferences. I used the relative scores of the bases at a given RNA position recognized by a RBP as a prediction for energy. The preferred base is that which receives the most negative score. The difference in the predicted energies between candidate bases determines the magnitude of the preference for the base with the lowest score. The spread in the energy prediction at a position is captured by the information content metric $H = \sum_{a \in \{A,C,G,U\}} p_a \log p_a$. Higher H represents a stronger preference for the most likely base at a position. We expect that the protein does not specifically recognize many base positions. If the energy function is correct for all base substitutions, information content should reflect the strength of selectivity at the base position. The subset of n bases with the highest value of H would be more likely to be correct. The plot shows the fraction of bases in the test previously described in Figure 3.5 and Figure 3.6 with a IC content range less than or equal to the index on the x axis that are correctly predicted. The figure shows that higher IC content does not increase the likelihood of making a correct prediction. (Rosetta – solid cyan; generalized – dashed green; all atom – dash-dotted magenta)

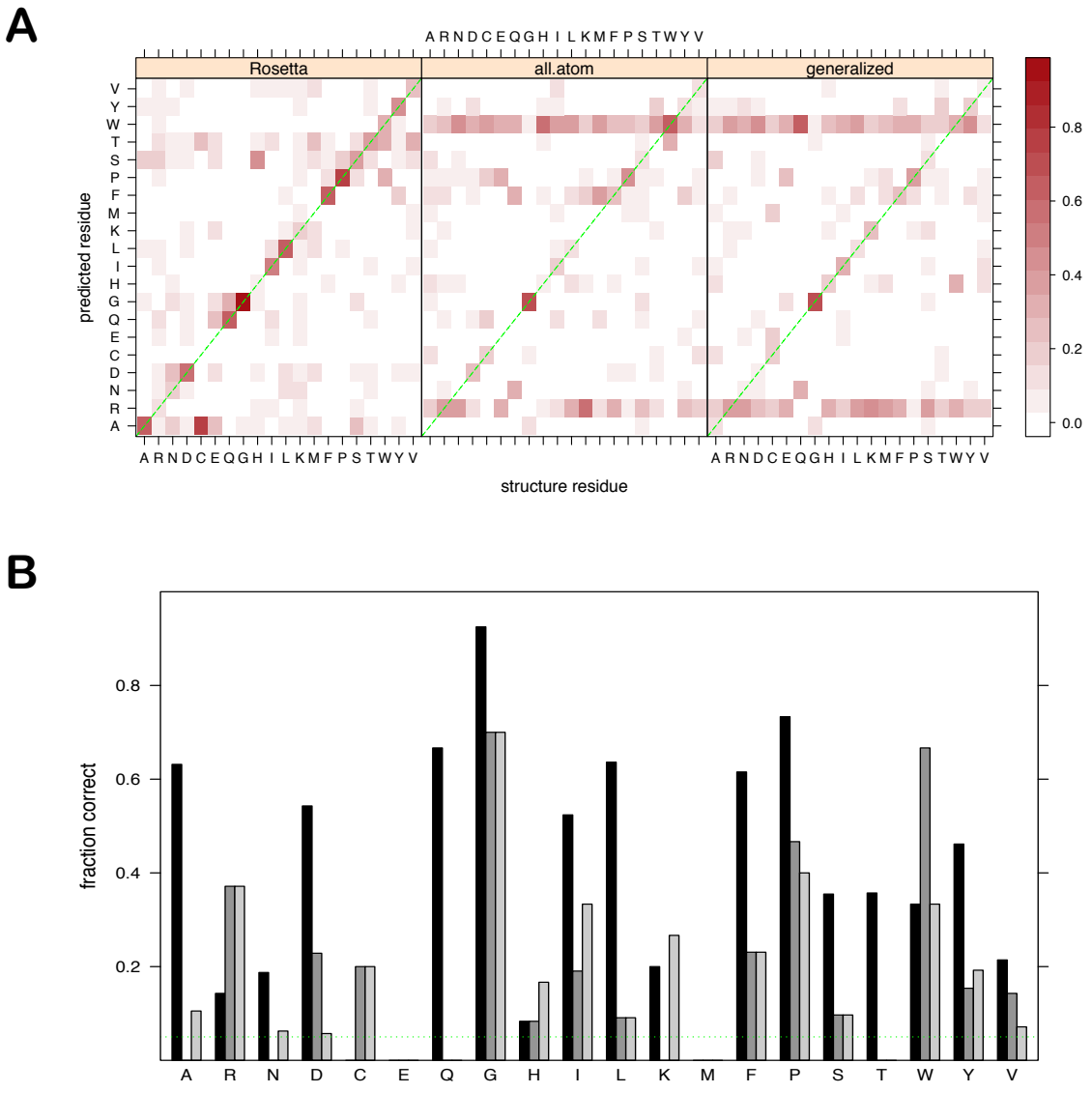


Figure 3.8: Comparison of amino acid recovery at binding interface of RBP recognizing ssRNA. **A** The heat maps represent the probability of each amino acid residue at the RNA interface being preferred (receiving the best score) by the scoring function at a position in a structure originally containing the specified residue. The green diagonal intersects cells representing the rate of recovering the correct residue (self-recovery). **B** The bar plot facilitates the comparison of the performance the three scoring function for residue self-recovery (Rosetta: black; all-atom: grey; generalized: light grey).

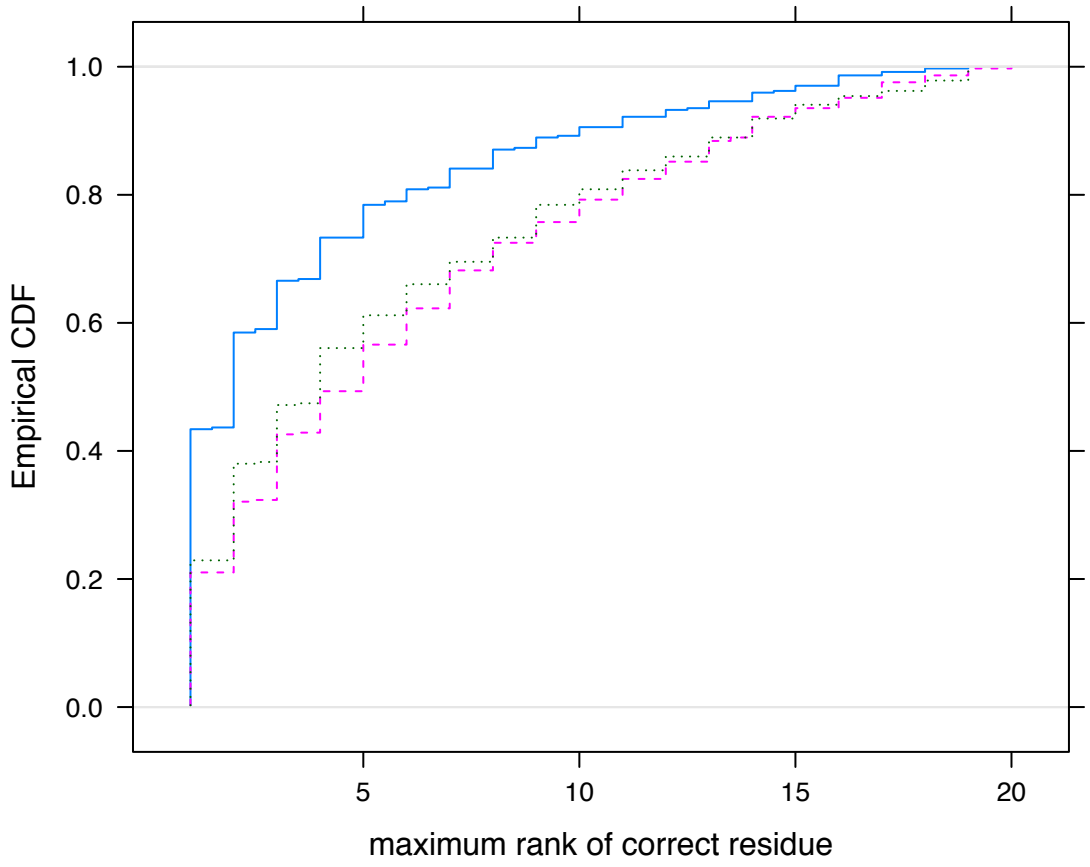


Figure 3.9: Probability of recovering the correct RBP RNA binding residue at each rank or better. The empirical cumulative distribution (CDF) plot gives us a closer look at the preference rank assigned to the correct residue at each interface residue position by the scoring functions. The plot shows the fraction of bases for which the structure base receives a preference rank less than or equal to the rank index on the *x*-axis. In the infrequent case where *n* possible amino acid substitutions receive the same score, the residues receiving the same score are assigned a range equal to the average of the next *n* rank indices (solid cyan: Rosetta; dashed magenta: all atom; dotted green: generalized).

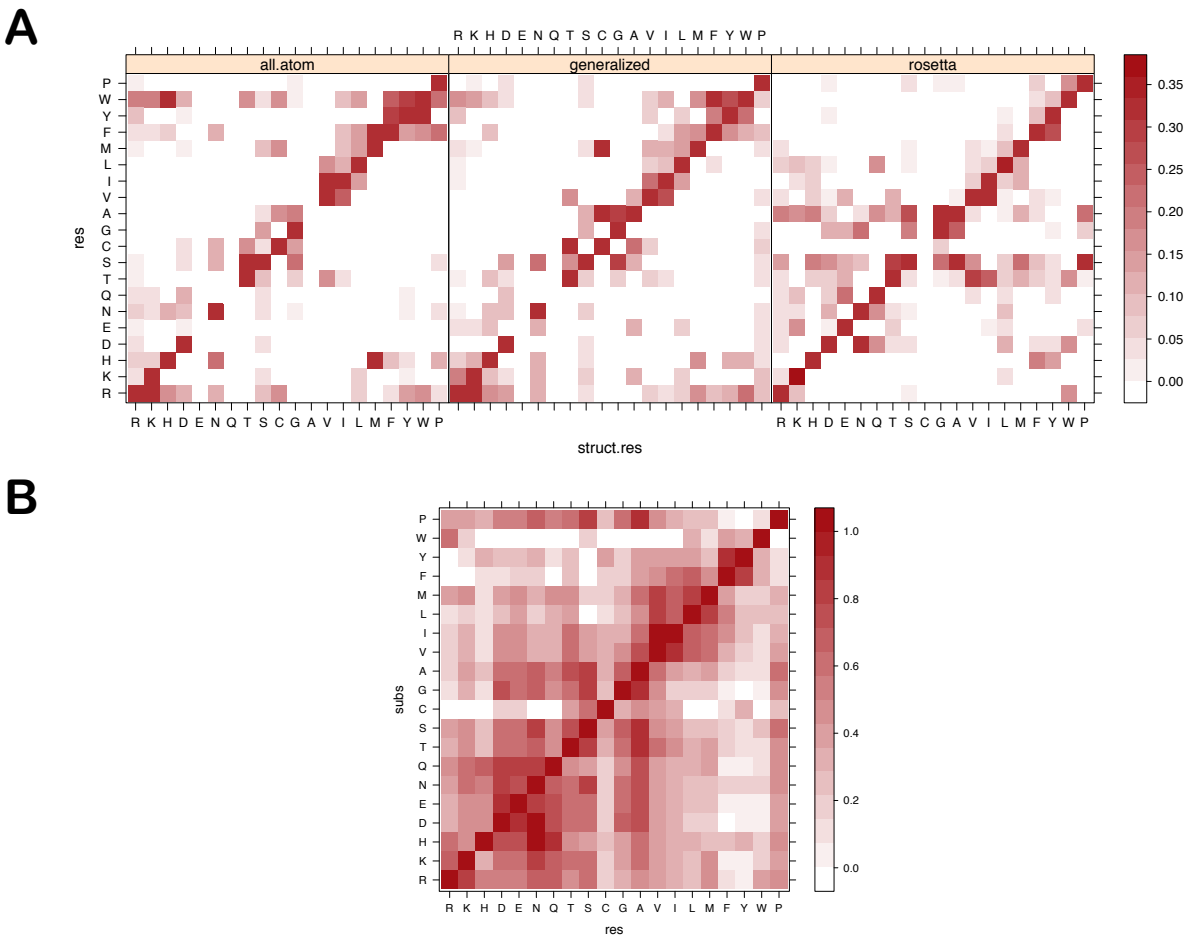


Figure 3.10: Preferred mispredictions of interface residues substitute for correct chemical properties. This figure evaluates whether residue selections based on energy score functions agree with biochemical intuition and sequence substitution matrices. **A** In the significant fraction of cases where the correct residue is ranked in the top 3 (Figure 3.9), I examined what other residues round out the top 3 with each scoring function. The residues are sorted by functional property (positively charges, negatively charged, polar and hydrophobic) and then by size. If residues selected based on chemical properties the off-diagonal fractions should reflect this. **B** A similarly sorted version of the PAM 250 matrix (Wilbur, 1985) is shown for comparison with the relative probabilities of X→Y subsitutions in evolutionarily related homologous proteins. Alignment derived substitution matrices such as the PAM matrices quantify the tolerance for residue point mutations in proteins. The most favorable residue substitutions share chemical or physical properties.

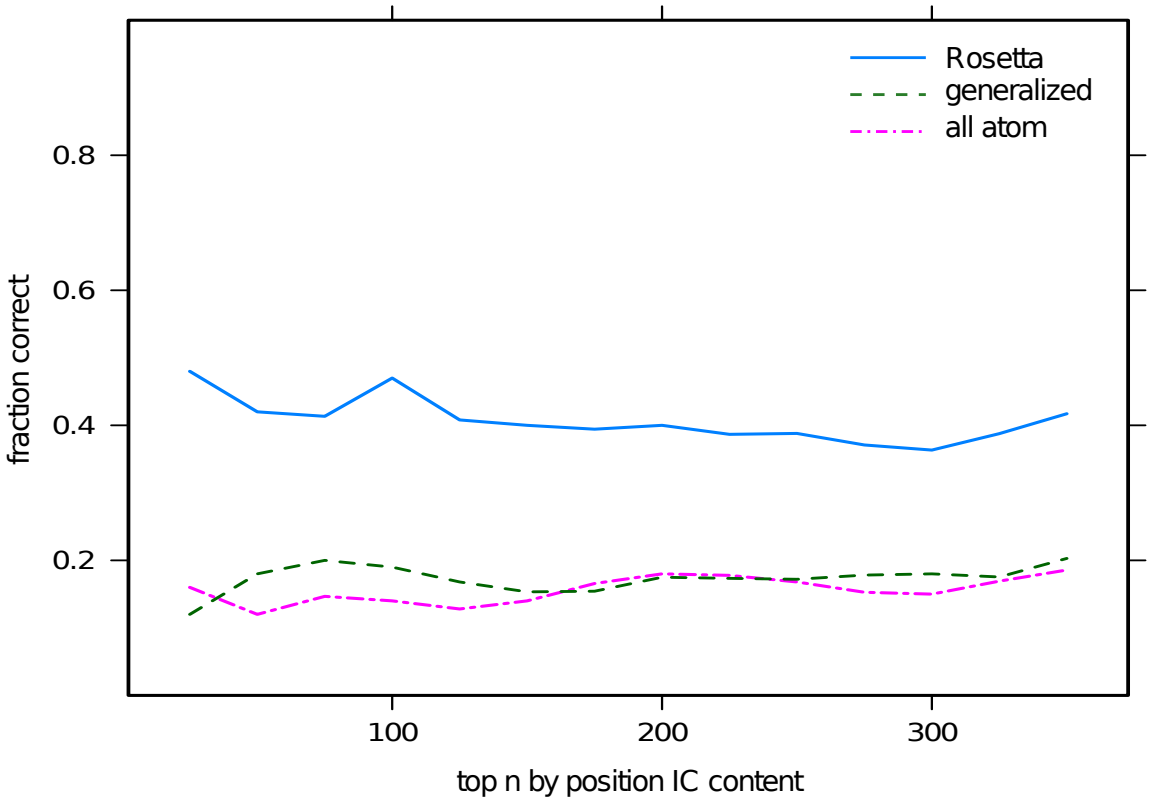


Figure 3.11: Fraction of residues correctly recovered at positions with the strongest predicted preference. The line plot demonstrates the fraction of correct amino acid residue recovery for the residues with an information content rank less than or equal to the index on the *x*-axis. The method and reasoning are described in the caption for Figure 3.8. In this case higher information content slightly increases the probability of a correct prediction only with the use of the Rosetta scoring function. (Rosetta – solid cyan; generalized – dashed green; all atom – dash-dotted magenta)

Table 3.4: Structures of RBPs in complex with RNA with independently determined specificity data. Only a few RBP complexes solved by NMR or x-ray crystallography can be associated with experimental specificity profiles. The binding profiles are from SELEX experiments, RNAcompete (Ray et al., 2009)9), or NMR-SIA (Beuth et al., 2007)7). For NMR structures, we used all submitted models in the calculation. Similarly, for x-ray crystal structures, I used all non-identical copies of the structure taken from the crystallographic unit cell in the calculations. For the fair version of the statistical calculations, we excluded count data from structures in the training set that had a sequence similar to the structure being scored.

gene	PDB ID	chain	exp. type	unique structures*	Pfam accession	Pfam ID	training set exclusions
A2BP1	2err	A	NMR	30	PF00076.16	RRM_1	1fxl, 3egz
ELAV4	1fxl	A	x-ray	1	PF00076.16	RRM_1	1fxl
ELAV4	1g2e	A	x-ray	1	PF00076.16	RRM_1	1fxl
KHSRP	2jvz	A	NMR	20	PF00013.23	KH_1	
MBNL1	3d2s	A	x-ray	4			3d2s
NOVA2	1ec6	A	x-ray	2	PF00013.23	KH_1	2anr
PABPC1	1cvj	A	x-ray	8	PF00076.16	RRM_1	1fxl, 3egz
PTBP1	2ad9	A	NMR	20			
PTBP1	2adb	A	NMR	20	PF00076.16	RRM_1	3egz
PTBP1	2adc	A	NMR	20	PF00076.16	RRM_1	1fxl
RBMY1A1	2fy1	A	NMR	17	PF00076.16	RRM_1	1n78, 1fxl, 3egz
SNRPA	1aud	A	NMR	31	PF00076.16	RRM_1	1fxl, 3egz
SNRPA	1drz	A	x-ray	1	PF00076.16	RRM_1	1fxl, 3egz
SNRPA	1dz5	A	NMR	13	PF00076.16	RRM_1	1fxl, 3egz
SNRPA	1urn	A	x-ray	2	PF00076.16	RRM_1	1fxl, 3egz
ZRANB2	3g9y	A	x-ray	1	PF00641.12	zf-RanBP	

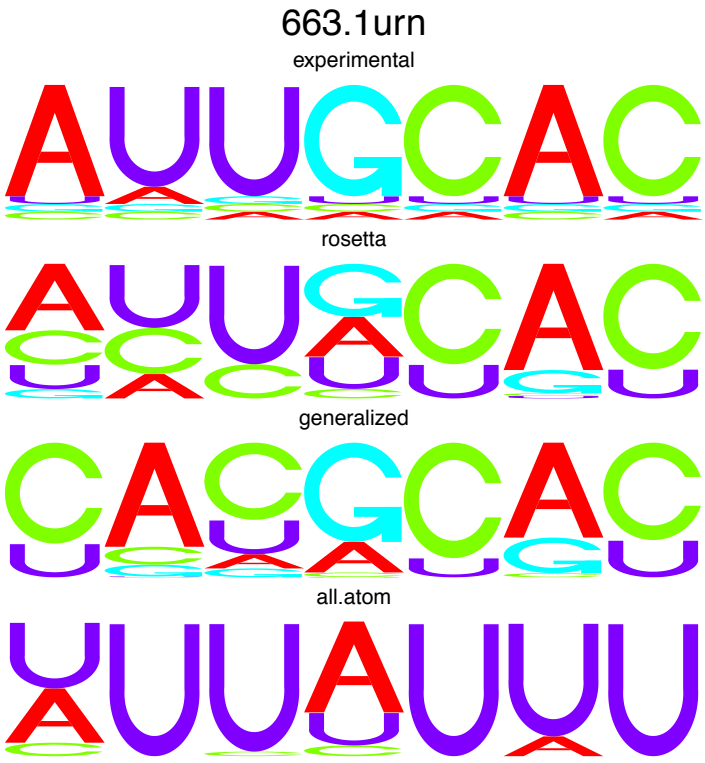


Figure 3.12: Representative predicted RNA binding motifs for the U1A protein.
The prediction of base-specificity by each scoring function can be represented visually using the sequence logo representation. For the binding site of the U1A protein, we show a representative experimental PWM from SELEX for snRNP-A (Tsai, Harper, & Keene, 1991) accessed from the RBPDB (Cook et al., 2010) and predictions using the three scoring function based on PDB structure 1urn. The height of the letters represents the relative probability that the base may be used in that position of the recognized sequence. The probabilities for the calculated logos are obtained from the Bolzmann distribution of the lowest score for each base at that position. The recovery of the experimental binding motif is quantified using a Euclidian distance metric in Figure 3.13.

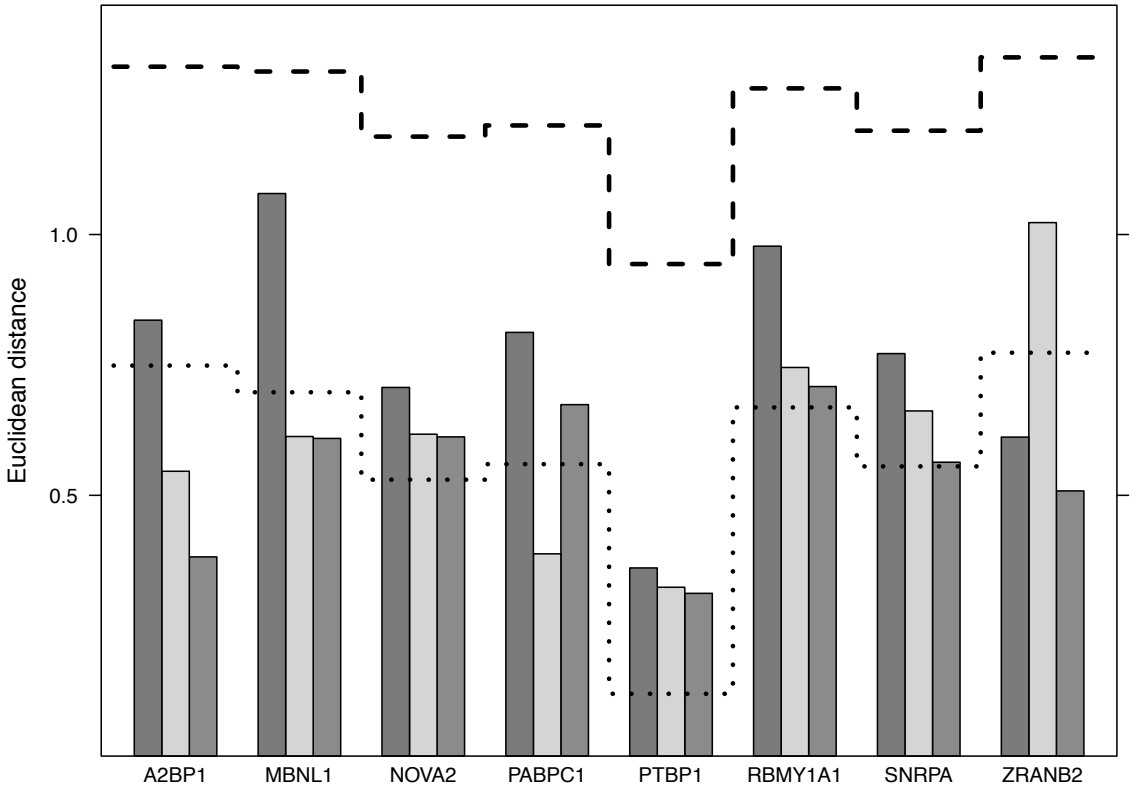


Figure 3.13: Relative performance of scoring functions in RNA binding motif recovery test.
I compare the average Euclidean distances (ED) between the calculated predictions from experimental values for eight proteins. The proteins constitute the complete set for which both experimental position frequency matrices and a solved structure for the RBP-RNA complex structure is available. A lower ED divergence is better since an ED divergence of zero would indicate a structure for which the probability of using each base was perfectly recovered for all positions. Bar group members from left to right, correspond to: all-atom, generalized, Rosetta. The lines provide bounds for the scores of each structure where the top (dashed) line is the distance to a PWM where the probability of the least probable base in the experimental PWM is set to 1 and the bottom (dotted) line is the distance to a PWM where each base has a probability 0.25.

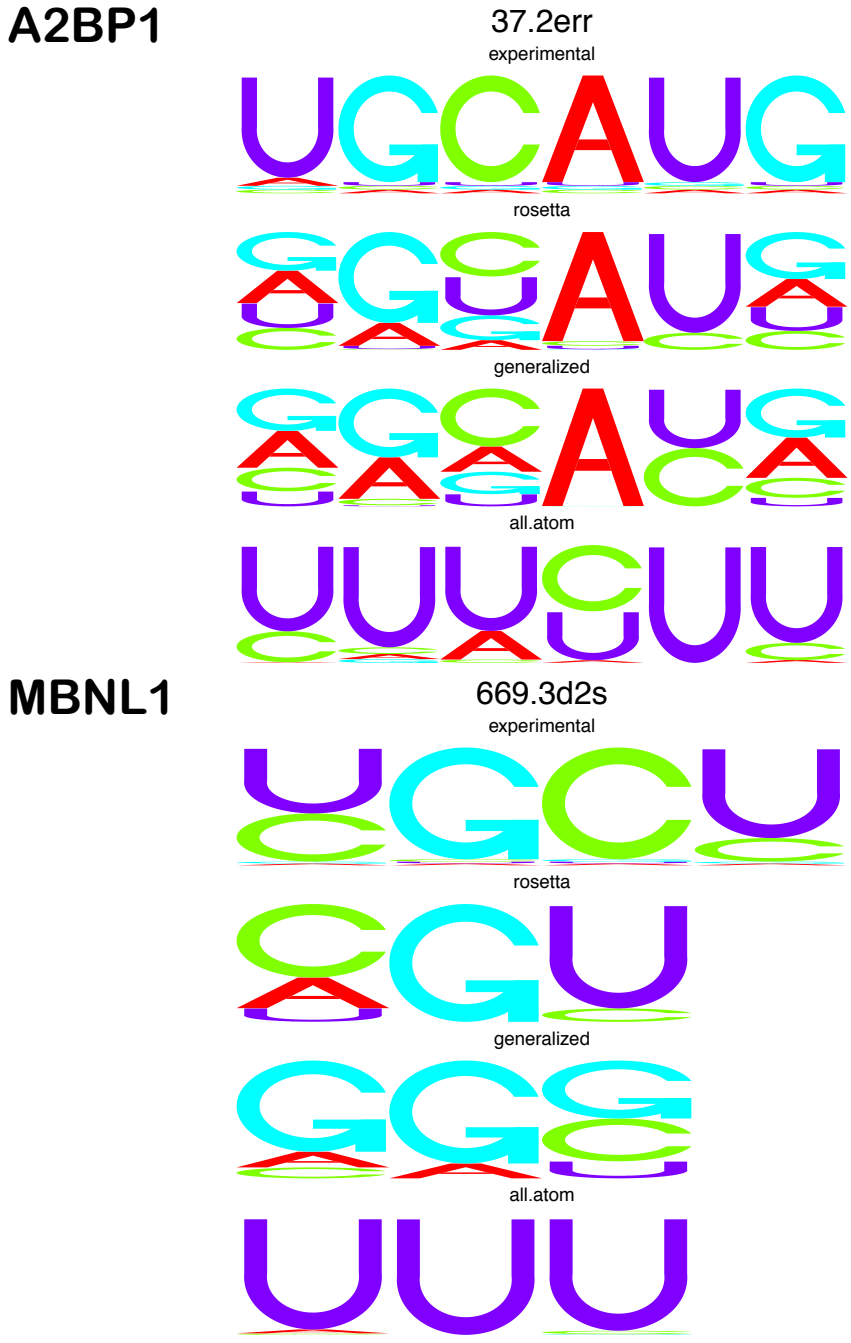


Figure 3.14: Representative predicted binding RNA motifs for the structures of A2BP1 and MBNL1. summarized in Figure 3.13. The experimental data is usually that of higher quality. The calculated data is the one for which Rosetta best recapitulated the binding profile.

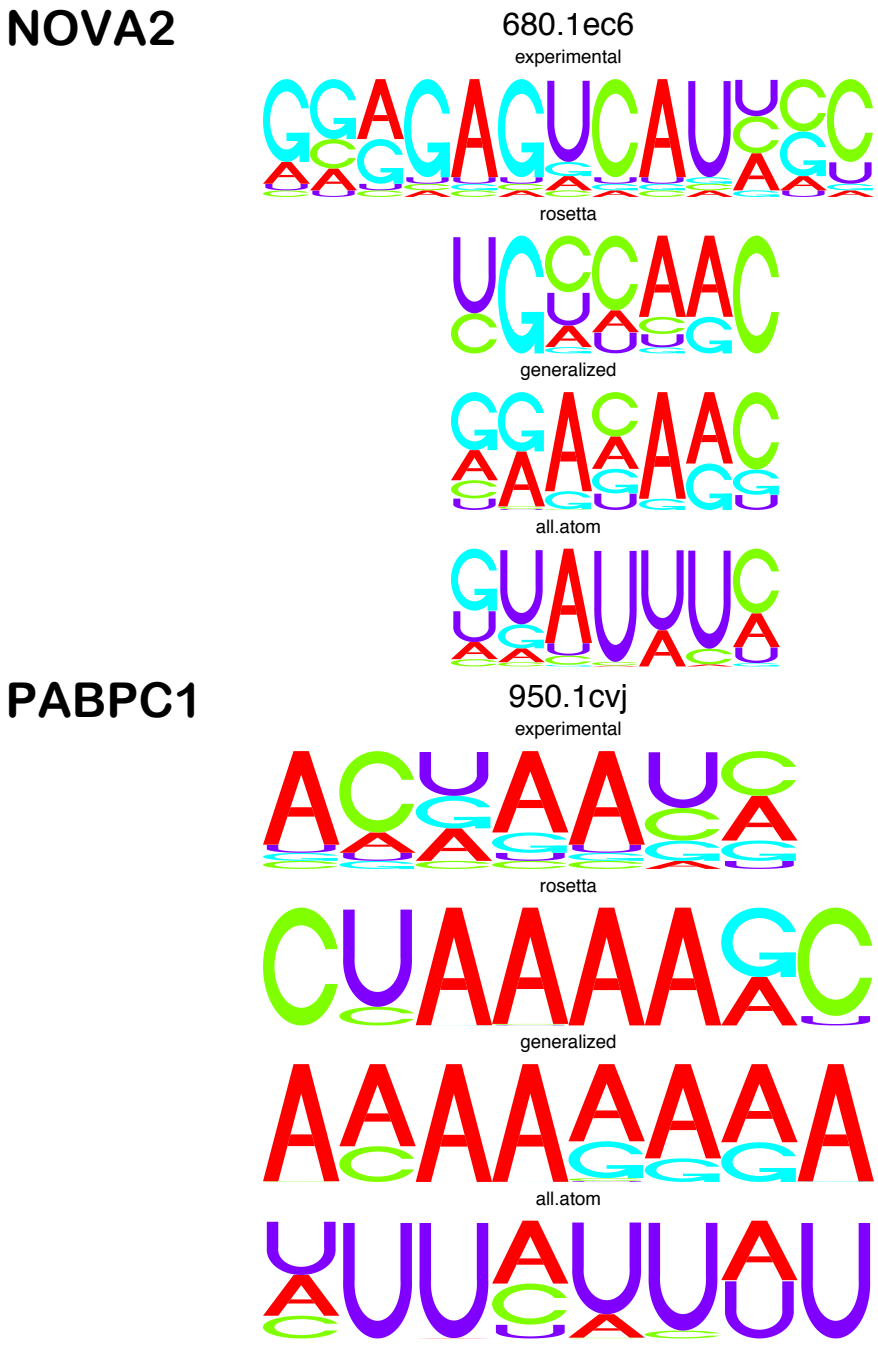


Figure 3.15: Representative predicted binding RNA motifs for the structures of NOVA2 and PABPC1.

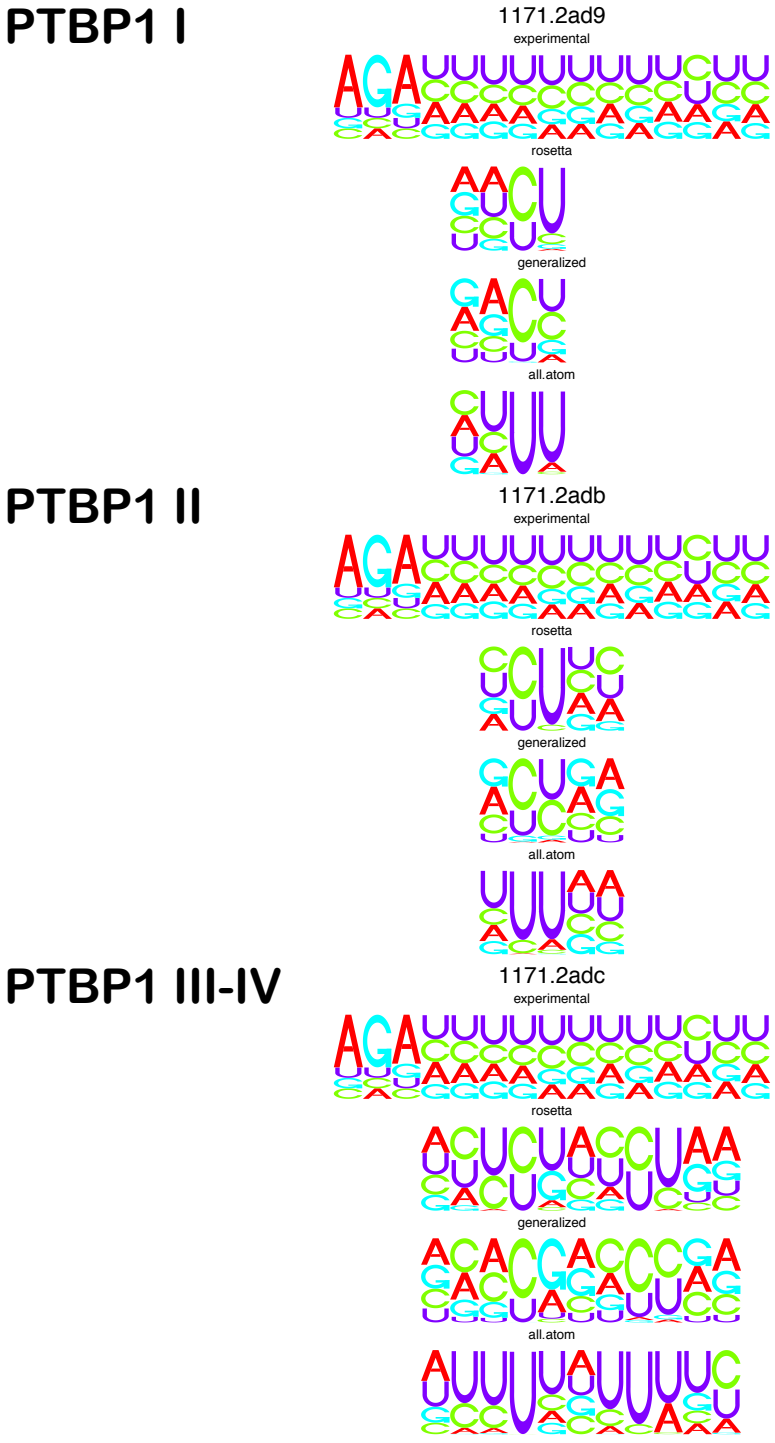


Figure 3.16: Representative predicted binding RNA motifs for the structures of the four subunits of PTB.

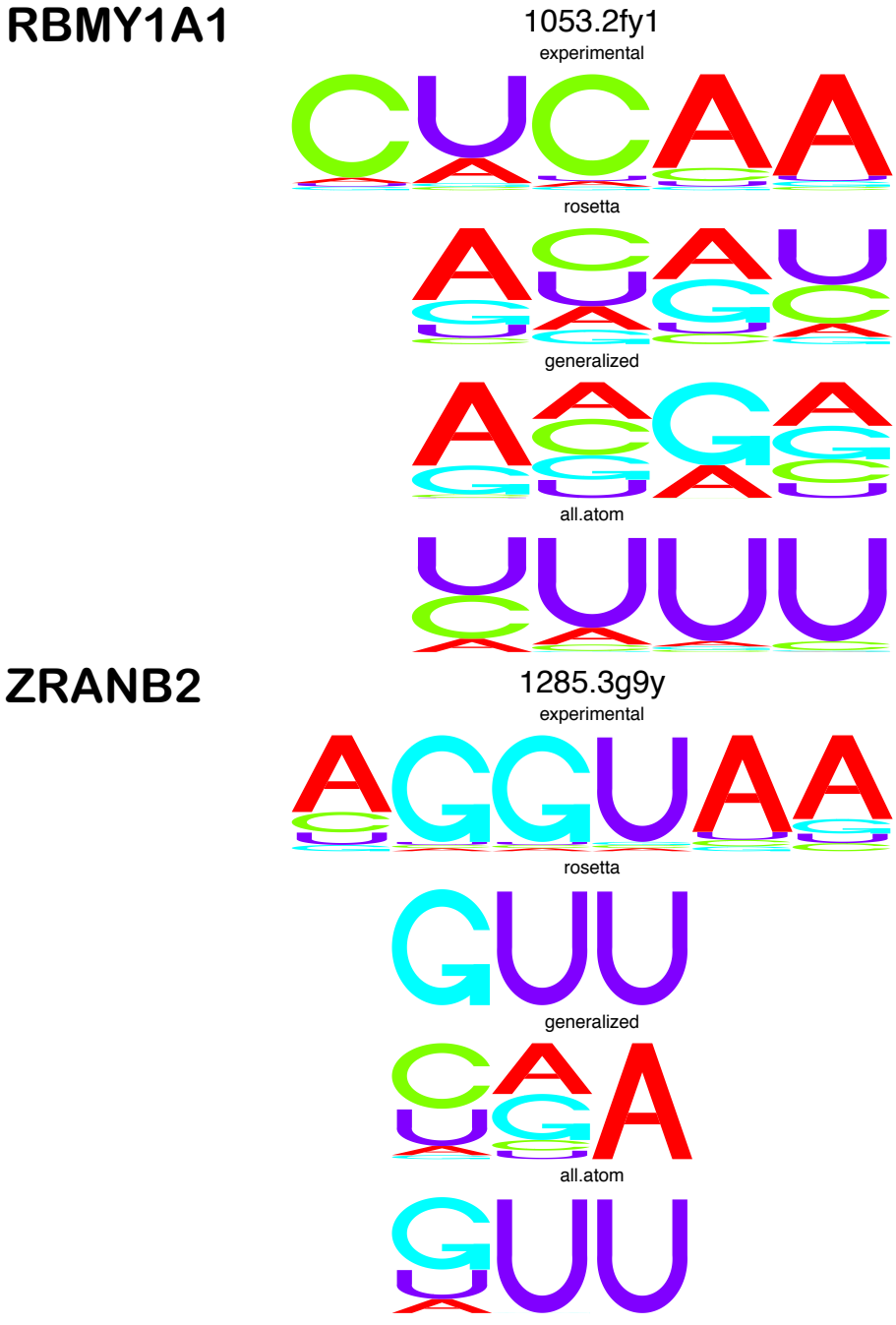


Figure 3.17: Representative predicted binding RNA motifs for the structures of RBMY1A1 and ZRANB2.

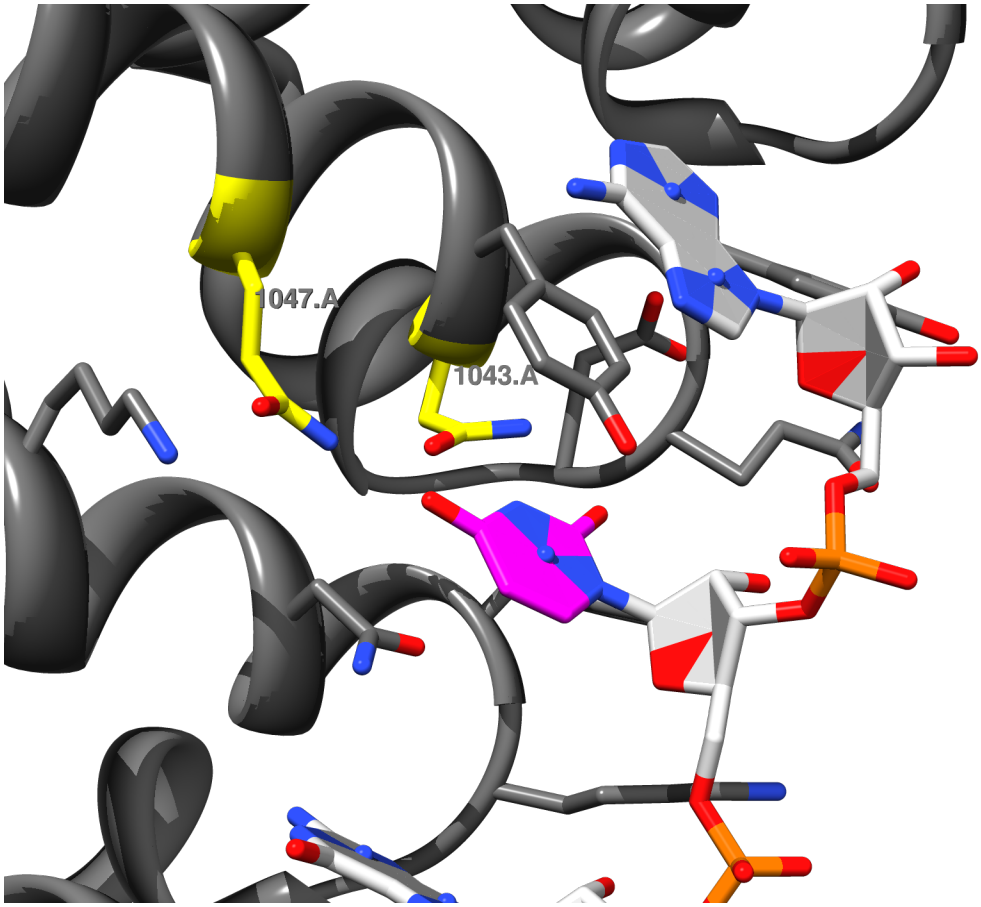


Figure 3.18: Structure illustrating Pumilio1 repeat 6 residues used in retargeting experiment.
The structure 1m8y highlights ASN and GLN at PDB position A1043 and A1047 (yellow) in the native structure of human Pumilio1. The highlighted residues contribute to the recognition of U (purple) in the native protein. Two groups have recently shown that selected double-mutants of the highlighted residues allows repeat 6 to be altered to preferentially recognize any of the four canonical nucleobases (Dong et al., 2011; Filipowicz, Bhattacharyya, & Sonenberg, 2008). I explored the specificity of the double-mutants at A1043 and A1047 computationally.

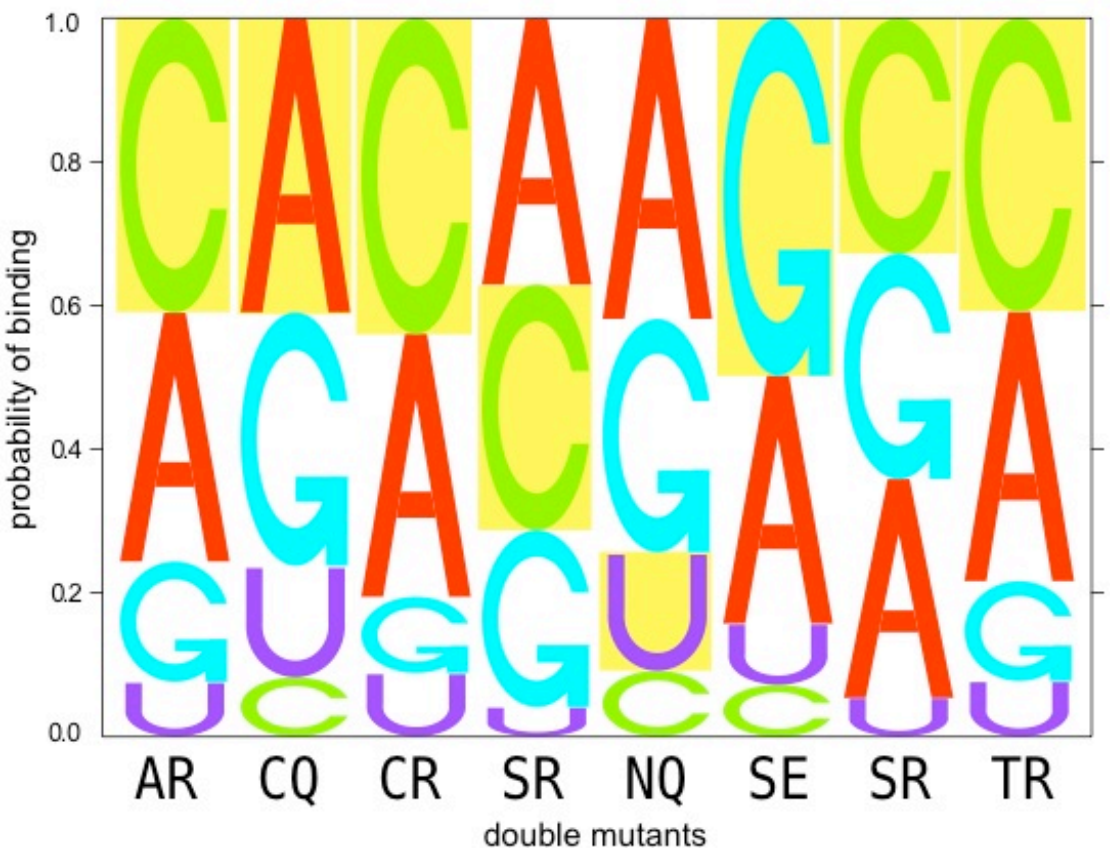


Figure 3.19: The effect of Pumilio1 double mutants on the binding preference at a RNA position.
I performed computational tests of the effects of Pumilio1 repeat 6 double mutants to positions A1043 and A1047 in structures 1m8y or 1yjy as illustrated in Figure 3.18 on nucleobase binding preferences. The calculations shown were performed with the Rosetta scoring function. The mutations to positions A1043 and A1047 are indicated pairs of residue one-letter-codes on the *x*-axis. Following residue mutation, the base preference of Pumilio1 repeat 6 was determined as described in Figure 3.1 with the inclusion of results from different starting structures as described in section H. Each logo column shows the base preference based on a sampled competition between 50 scores for each starting structure. The experimentally preferred nucleobase for the indicated double mutants as reported by Dong, et al. (2011) is highlighted in yellow.

Chapter 4. Application of Specificity Prediction Tools to Protein Design

A. Introduction

A significant difficulty in validating the scoring function for target specificity applications is identifying good test cases that provide a clear metric for success. With the data available for structures and specificity of protein-RNA complexes, I could show that the scoring functions were useful in structure-based specificity predictions (Chapter 3.G). Quantifying the expected performance of the scoring functions for accuracy in predicting a position weight matrix from structure remains difficult. Validation against a new and larger set of specificity data would improve confidence in the utility of my scoring function for use in biological applications. Demonstration of a functional design would provide the strongest evidence that the computational approach to target specificity prediction captures the most important interactions.

Sequence level comparison of experimental and predicted binding preferences – The approach of comparing position weight matrices provides results with good statistical confidence only if a large set of experimental data is available. The RNAcompete microarray method developed by Ray et al. (2009) is in principle scalable to determine position weight matrices (PWMs) for a set of proteins much greater than the 9 reported. However, there are significant limitations to this approach that have made it difficult to obtain PWMs for significantly more proteins. Additionally the short

sequences that are exposed on the microarray cannot describe the entire binding site of a protein, and the time for creating constructs limits the determination of the per domain specificity of the proteins. So the available position weight matrix data will remain small and will continue to originate from various experimental approaches.

In the absence of a large set of experimental PWMs that would allow us to establish confidence in our predictions with respect to an external metric, the comparison between experimental and predicted motifs that currently are available are still useful. Comparison of predicted specificity with additional experimental measurements would support the computational strategy and could lead to insights to improve the scoring function or the approach to exploring the impact of sequence changes. Insight into improperly parameterized terms in the scoring function or into shortcomings in the algorithmic approach to sequence space exploration will require careful analysis of the actual and predicted structures in the context of the specificity prediction. Correct prediction of binding sequences from homology modeled protein-RNA complexes would additionally increase the size of the test set and, if successful, would demonstrate that the predictions are robust. In order to gain insight from the overlapping set of binding experiments and known structures, we may look to additional evidence about binding that is lost in the PWM comparison.

I delved deeper into the microarray data to discover if there is a higher order (non-position independent) relationship between sites. In theory, the microarray-based approach to determining RBP binding motifs, RNAcompete, provides individual binding affinity measurements with the complete library of seven base single-stranded RNA sequences (Ray et al., 2009). This data set allows us to directly ask the key biological

question of whether a specific sequence RBP binds a specific RNA sequence. The question asked of a specific sequence may differ from expectation presented in a PWM if the bases are not recognized independently (as discussed in section B.2.a below). Comparing computational predictions at the level of an individual sequence provides a valuable and strict test of the scoring function.

The potentially interesting sequence-level approach to level approach to the problem was complicated by the quality of the RNAcompete data. I performed the sequence level comparison with a complex for which both microarray data and a a number of structures were available. The additional steps in the required to adapt a cDNA microarray to the RBP assay complicated this potentially interesting analysis.

Approach to RNA binding motif retargeting – A successful computational protein design would provide strong evidence that the scoring function and approach to searching structure and sequence work. Designs may be tested in a similar manner to assays employed to validate binding candidates from evolutionary approaches to experimental design. The selection approach to RBP re-targeting is the state of the art in identifying protein mutations that alter binding specificity (Y. Chen, Mandic, et al., 2008; Danner & Belasco, 2001). However, the selection approach often fails to capture sequences intermediate to that with a specificity switch because selection stringency is hard to adjust. A computational approach to retargeting could significantly speed-up the engineering process.

A successful design would support the argument that the computational approach works. A design prediction is successful if it is more likely to bind than a protein with a similar number of random mutations. Since a protein that does not

natively recognize a sequence is extremely unlikely to be converted to one that does through random changes to protein sequence, most successfully retargeting events provide strong evidence that a redesign method is valid. Biochemists have long sought to perform rational design by intuition, but in the absence of computational assistance with structure minimization and energy prediction, they are rarely successful (Cooper & Waters, 2005). A successful computational design does not provide a better quantitative argument for the performance of an approach than would comparison with experimental binding motifs. However, the success of the design would still provide strong evidence that a computational approach is valid and useful.

Complementarity of approaches – A computational tool for predicting and altering target sequence binding specificity in RBPs is necessary for completing our knowledge of post-transcriptional regulatory networks. Some good techniques have been developed for querying the specificity of proteins or domains that can be expressed in a soluble form. The power to infer binding targets for the hundreds of domains of similar family and fold to those known would allow us to fill important gaps in our knowledge (Messias & Sattler, 2004).

Progress in developing these computational methods will require an iterative approach with specificity prediction and design applications. All existing specificity data are exploited to elucidate scoring function limitations in the structure-based approach to specificity. Design applications allow a directed and finer exploration of the relationship between protein sequence and target specificity. Each approach provides valuable information than can be used to improve the computational approach.

In this chapter, I discuss limitations in matching predicted binding specificities and affinities with data from the state of the art micro-array experiment. I discuss how the specificity prediction tools are applicable to the problem of binding protein design. I apply the design process to design a domain to target a physiologically relevant microarray target that could be used in a drug-like context.

B. Detailed Comparison of Binding Predictions with Microarray Data

The set of proteins of known specificity and the rate of discovery of new specificities will allow direct blind tests of specificity prediction algorithms. New techniques such as RNAcompete provide a potentially scalable approach to discovering the binding specificities of RBPs. The results for these experiments will allow the establishment of confidence boundaries for specificity predictions. While these data are not yet available for a large set of proteins, the higher quality of the data may allow for some additional analysis.

A more detailed view of specificity may be obtained by taking advantage of raw data used in the RNAcompete microarray analysis of RBP binding specificity. With the microarray data, we may be more confident of the quality of the binding motif and we may ask more detailed questions about the binding. The RNAcompete approach by Ray et al. (2009) measures in theory the relative affinities of the complete set of heptamers. The exploration of RNA sequence space is guaranteed to be comprehensive. Having the complete set of binding motifs, we may discover whether the position independent PWM is a good description of the binding sites. Additionally, in order to test whether the RBP preference for RNA targets involves higher order relationships between

adjacent bases, I performed comprehensive rethreading and repacking tests, extending base specificity approaches to the full length of the bound RNA.

1. Experimental Constraints Affecting Comparison with Predicted Values

A generalizable quantitative comparison of predicted and measured RBP target motifs was limited by the amount of data available. The highest quality data comes from the RNAcompete approach that uses DNA microarrays to explore a large segment of RNA sequence space (Ray et al., 2009). The data from the microarray approach is an adaptation of a method that used protein-binding microarrays to directly measure specificity of transcription factors (Berger & Bulyk, 2006). However, limitations on the sequences that can be explored, the constructs that can be made and complications due to the extra transcription and re-annealing steps limit data availability and the analysis that can be performed with the data.

a. Microarray sequence limitations

The length of single stranded RNA that can be exposed is limited by the formation of RNA secondary and tertiary structure. Ray et al. (2009) designed a pool of RNA sequences that would express nearly every seven base sequence of RNA in an unstructured RNA region or as a loop in a stem loop structure. In designing the DNA sequences for their microarray, they used the RNAshapes program (Steffen, Voß, Rehmsmeier, Reeder, & Giegerich, 2006) to computationally predict the secondary structure of the sequences. The limits on single stranded sequences that can be expressed in solution by transcribing DNA from a microarray is limited to sequences that are short in comparison with those recognized by microarrays querying transcription factors.

The limit on single stranded RNA structure that can be searched with the microarray is similar to the limits on native RNA structures exposing single stranded regions. Some RNA tertiary structures may stabilize longer exposed regions of single stranded RNA. But the domains of interest mostly recognize regions shorter than seven bases. However, the Pfam database shows that most of the common RNA binding domains occur as part of architectures containing other RNA binding domains. Exploring sufficient structure and sequence space to query the entire RNA binding surface of the protein is not yet possible.

b. Stable expression

A protein must be expressed in a construct that will yield useful binding intensities with sequences from the microarray. The construct must be stable, soluble and must present the binding domains of interest. While creating a construct that can be used in the microarray binding experiment is simpler than expressing the protein in a manner conducive to structure determination, the expression is still a significant challenge. Expressing all RNA binding domains from even a single genome is a daunting challenge. When the importance of haplotypes of these proteins in disease is factored in, the difficulty is increased. A computational approach would allow information from a few measurements to be generalized to the sequence space of interest.

c. Comparison of DNA and RNA microarray specificity determination

The DNA binding arrays more directly measures binding affinity than its RNA counterparts. The RNAcompete approach by Ray et al. (2009) differs from specificity measurements used for transcription factors because of additional steps needed to adapt the technique. PWMs for transcription factors have been extensively measured

using microchips (Berger & Bulyk, 2009; Philippakis et al., 2008). These microarrays directly query protein affinity for a specific sequence by directly visualizing the binding of labeled proteins to the array. The RNAcompete approach first transcribes all the sequences on a cDNA microarray to RNA. Using a protein of interest, Ray et al. (2009) pull down and label the bound RNA. The RNA bound by RBPs is indirectly visualized by allowing the labeled RNA to anneal to the original cDNA microarray. Thus, the measurement contains two additional steps and is more indirect.

The more indirect measurements used in RNAcompete may differ from its DNA counterpart in some predictable ways. Firstly, we would expect the measurements themselves to be noisier than the DNA counterpart. Ray et al. (2009) performed their measurements in duplicate. The observed intensities in each pair of duplicates strongly correlate with each other, but the raw measurements show substantial differences for individual measurements. Secondly, we expect the sequence specific binding to show a greater difference in intensities between high and low affinity sequences than would be observed in the DNA microarray (Gharaibeh, Newton, Weller, & Gibas, 2010). The RNA method uses the proteins to pull down RNAs following a mixing step. Thus the RNA case more closely resembles competitive binding conditions. This exaggerates the difference between the affinity of bound and unbound sequences.

An RNA microarray would allow a more direct measurement of RNA binding specificity. Recently, RNA microchips were shown to be possible when ribose 2'-OH groups are protected until use (Mikheikin et al., 2008). If the RNA was stable, these measurements could be scaled as they have been scaled recently for transcription factors. The UniPROBE database (`http://uniprobe.org`) now describes 315

transcription-factor binding motifs (Newburger & Bulyk, 2009). This suggests that a much greater amount of data could be obtained if a true RNA binding array were available. However, the RNA molecules that could be exposed would still be significantly limited by structure.

The length limit encountered by the method of Ray et al. (2009) would still apply. The RNA sequence library that may be exposed is limited to small, unstructured RNAs and stem-loop structures exposing short heptamer single stranded regions. While the RNA binding domains which bind single-stranded RNA recognize heptamers or shorter. RNA binding proteins employ multiple binding domains in succession to bind a longer region of RNA. Thus, using the microarray technique the domains must be expressed independently in stable and soluble constructs. The preferred binding motif and affinity for an entire protein cannot therefore be measured in a single step.

2. Detailed Analysis of Predictions with Microarray Data

In theory the data from the RNAcompete approach provides a complete set of binding affinities for every heptamer sequence. The complete coverage of sequence space should allow us not only to compare binding motifs where positional independence is assumed, but should also allow us to compare higher order sequence dependencies in the binding preference. Here I discuss higher order sequence constraints and discuss an approach to test this hypothesis.

a. Debate over higher order sequence constraints

The microarray measurement of RNA binding specificity potentially provides a basis for determining whether bound nucleic acid positions are recognized independently. The DNA binding data for transcription factors in the UniPROBE

database recently sparked a debate over whether DNA positions are recognized independently (Morris, Bulyk, & Hughes, 2011; Y. Zhao & Stormo, 2011). An analysis of the UniPROBE data suggests that higher order relationships between base pairs need to be considered (Badis et al., 2009). However, reanalysis of the derived position weight matrices suggested that most selectivity could be described by a position independent approach (Y. Zhao & Stormo, 2011). Zhao and Stormo (2011) suggest that when positions are not independent, the dependency can be captured by an adjacent pair code. The representation of some RNA targets of RBPs is more likely to sometimes require higher order terms since RNA structure often involves contact between non-sequential bases.

In order to address this concern, I calculated the complete set of binding scores for rethreaded four base sequences for RRMs where we have both microarray data and solved structures. Various visualization techniques have been proposed to make sense of comprehensive measurements of sequence space. I employed the receiver operator characteristic approach and another visualization approach based on number of sequence mismatches.

b. Comparison approaches

Since the RNA microarray approach reports affinity for each heptamer sequence, I should be able to validate predictions without making assumptions about position independence. I applied a receiver operator characteristic (ROC) approach to check whether better binding sequences consistently receive better scores than non-binding sequences. Additionally, I compared the quality of the data with that reported using typical UniPROBE experiments using a reported visualization technique.

Predicted energies for target RNA sequences may be incorrect if independence between positions does not hold. I may look at how the score of an entire sequence predicts binding affinity. I use the ROC curve to check that binding sequences are more likely to receive a better score. By using this approach, I may compare performance when calculations are performed by independently substitution base positions with the same calculations where the entire sequence space of the bound RNA is searched exhaustively. Additionally, this approach promised to allow a quantitative comparison of the performance of the scoring functions.

ROC analysis: The ROC plot is used to capture the performance of a classification function where the classifier reports a scalar value related to that classification (Krzanowski, 2009). The comprehensive data set admits a more rigorous test than PWM models for predicting binding targets in the form of ROC curve (Badis et al., 2009). The microarray data can be divided roughly into two categories of interest: binding and non-binding. We may see whether our continuous distribution of sequence scores can reproduce this categorization.

Our calculated position independent PWMs may be used to assign a probability of binding to each heptamer. The probability of a sequence binding is the product of the position probabilities.

$$P(s \mid \Theta) = \prod_j P(s_j \mid \Theta_j) \tag{4.1}$$

However, in the case where the test sequence is longer than the *k*-mer binding site we must consider the probability of the *i*th sequence s^i of length L^i binding in any frame.

$$g(s^i,\Theta) = 1 - \prod_{t=0}^{L^i-k}\left(1 - P(s^i_{t+1:t+k} \mid \Theta)\right) \qquad (4.2)$$

After setting a true-hit threshold for the microarray data the ROC curve may be elucidated by sampling with a decreasing probability threshold (Figure 4.1). The area under curve (AUC) metric corresponds to the probability that the scoring program will allocate a better score to a binding sequence (Krzanowski, 2009). The comparisons with experimental data using ROC plots will provide a very rigorous test of the performance of the potential.

Alternatively, we may comprehensively rethread and repack the entire interface. From these data we may directly obtain the probabilities of binding each sequence of length equal to the binding region. Since the binding region is likely smaller than the bound sequence we still have to discover the probability of binding the heptamer using equation (4.2). By comparing the ROC plot of the complete rethreading to the position independent approach, we may discover whether higher order sequence constraints are important for RNA binding proteins.

The ROC approach allows us to rank the performance of the scoring functions base on the area under curve (AUC) of the ROC plots. The AUC values are not substantially affected by scaling factors. However, there is only a tiny overlap between the protein domains for which we have structures and whose binding preferences were determined by the microarray approach.

Quality analysis by mismatch plot: To address some of the concerns with quality of the intensities from the microarray method discussed in section Chapter 4.B.1.c, I applied a visualization method that worked effectively in the analysis of data from the

UniPROBE database (Newburger & Bulyk, 2009). The method described by Carlson et al. (2010), uses the consensus binding-motif for a RBP and groups the sequences on the microarray by minimum number of mismatches with the consensus sequence. Within each group, the mismatches are sorted lexically (alphabetically sorted by treating the strings of base one-letter-codes as words). When applied to transcription factor binding microarrays, the score clearly correlates with number of mismatches. Additionally the lexical sorting of the mismatches allows quick visualization of the positions most important for high affinity binding.

3. Results of Comparison Between Predicted Binding Sequences and RNAcompete Data

Previously, I tested how well a structure-based approach to specificity prediction worked at prediction position weight matrices (PWMs) that agreed with measurements from a variety of experimental techniques (Chapter 3.G.2). When compared by visually examining the PWMs represented as logos or quantitatively using motif comparison metrics, I could demonstrate relative performance of specificity prediction methods. However, the binding data from the new microarray experiments in principle yield a complete set of sequence affinity measurements. A more detailed comparison of predictions with these values could give useful insight into prediction errors and could address concerns about higher order sequence dependencies. Figure 4.1 outlines an approach to comparing predicted binding to the complete heptamer to the raw microarray binding intensities.

a. Comparison using the receiver operator characteristic

The overlap between proteins for which there are specificity measurements using the microarray approach and for which known bound structures is small. The

microarray data allow us to quantify the extent to which the energy function picks out sequences that bind. The microarray data for nine RBPs were recently published (Ray et al., 2009). However, structures are known for only 5 of them. Of those, only three are solved in complex with RNA. Given the current PDB, the data for U1A is the only set suitable for this analysis.

The three proteins for which RNAcompete data and structures are available illustrate the quality and limitations of the data. The structure of U1A is well characterized and several structures are available for that complex. The structure of VTS1 (UniProt ID: Q08831) is also available, but the recognition site is between a sterile alpha motif (SAM, Pfam: PF07647) domain and a single unbound G and a following C involved in the structure of the stem-loop tertiary structure (PDB ID 2f8k). The additional motif positions captured by RNAcompete are not due to RBP contacts to single stranded RNA and cannot be predicted from structure without a RNA intra-molecular term. Structures of PTB are available for each of its four RRM domains in complex with RNA. However, the experiment measures the affinity of the full-length protein recognizing a sequence far longer than the seven base sequences on the microarray. Thus, additional calculations are required to map the PTB predictions to the measured binding affinities. The other proteins (such as HuR) for which we have a protein structure are good candidates for future calculations using homology modeling, but it is premature at this stage to use homology models, which could introduce other sources of uncertainty. For my purpose here, only the U1A complex provides a clear example suitable for a more detailed analysis.

As a proof of concept, I calculated the RNA binding profile for the U1A RRM from the structure 1aud using the two empirical scoring functions and the Rosetta scoring function. I predicted the probability of binding the sequence based on predicted PWM. In this special case, the length of the bound RNA was seven nucleotides and matched perfectly with the length of the single stranded regions on the microarray. Figure 4.2 shows the ROC curve as described in methods (section 2.b, above), but assuming only a single binding frame. The area under curve (AUC), for the fair all-atom, the generalized and Rosetta scoring functions were 0.398, 0.394 and 0.568, respectively. Predictions more useful than a random approach have an AUC greater than 0.5. The prediction with the Rosetta scoring function indicates that binding sequence will have a higher score than a non-binding sequence around 57% of the time. The results are obviously far from the ideal (where AUC would approach 1).

While the native binding motif does emerge from the RNAcompete data, the structure of the RNA may explain why the binding sequences are not as strongly preferred as expected. The RNAcompete PWMs are built from the sequences and microarray intensities using the RankMotiff++ algorithm (X. Chen et al., 2007). The PWMs from these data largely match those from previous methods in literature (Ray et al., 2009). However, for a given protein the intensity of the binding of individual sequences may not necessarily reflect the relative energy of recognizing that sequence. If the sequence dependent difference in binding energy is not well resolved, the ROC assay may not perform as expected. I explore the sequence dependent intensities below (section b). For the case of U1A, we may expect that the binding to the correct sequence in the incorrect RNA structural context may be suboptimal.

Aspects of the binding affinity of proteins such as U1A to the preferred RNA target not captured by the RNAcompete assay include the effects of the RNA structure and induced conformational changes within the protein. A multi-step binding model has been proposed to explain the complex dissociation of U1A to its target stem loop 2 (SL2) of U1 snRNA (Anunciado, Dhar, Gruebele, & Baranger, 2011). Conformational changes in the RNA are involved in the transition between a tightly bound and loosely bound RNA state (Allain et al., 1996; Qin et al., 2010). The structural transition upon binding contributes strongly to the affinity for the sequence in the native stem loop in a mechanism often called the 'lure-and-lock' mechanism. The absence of the contextual RNA structure and the interconversion to the 'locked' structure may explain why the correct sequences do not have intensities more than a few standard deviations above the non-binding sequences.

The computational predictions were performed starting from a structure that resembles the 'locked' structure. The predicted preferences thus reflect the most correct conformation. The RNA intra-molecular score-terms are not present in the empirical scoring functions and are not well developed in the Rosetta scoring function. While the use of the structure optimal for the native sequence helped the scoring of binding sequence candidates, the structure interconversion is not likely sufficient to explain the poor agreement in the ROC curve (Figure 4.2).

Performing the same analysis with other proteins for which we have RNAcompete data required additional steps that would complicate the analysis. The U1A protein is the only protein in the RNAcompete and PDB datasets for which I can do a simple ROC comparison of predicted and experimental binding preferences. The

structures for all RRMs of the polypyrimidine tract binding protein (PTB) are available in complex with RNA. The RNAcompete data for PTB is for the binding of the entire protein to the heptamer RNA sequences. The ROC analysis is possible if we assume that only one RRM may bind the heptamer at a time (see section 2.b above). However, the agreement between prediction and sequence intensities would have needed to significantly better that the U1A results for the analysis to work.

The ROC metric still promises a quantitative path for optimizing the potential for specificity. The current data suggests a test for the remainder of the RNAcompete data with the bound RNA modeled or with homology modeling of the protein (as would be required for VTS1). Since in RNA neighboring bases may have higher order specificity relationships than in DNA, an approach to optimizing the agreement between predicted binding preference and experimental measurements may still prove valuable.

b. Relationship between consensus sequence and microarray intensity in RNAcompete data

An additional problem learning from the individual binding intensities may be attributed to the noise in the binding affinity measurements. A visualization technique developed to uncover the most selective binding sites for transcription factors binding to protein binding microarrays can be applied to the RNA data (Carlson et al., 2010). Figure 4.3 shows the application of a flattened version of this visualization approach to the intensities for binding of U1A data to the heptamers queried in the RNAcompete approach.

The consensus binding sequence was used to calculate the minimum number of mismatches. In applications to UniPROBE data, the low mismatch number (1 and 2) sequences show high intensity while the higher mismatch sequences show much lower

intensity (Carlson et al., 2010). For the of case U1A intensities in the RNAcompete data, Figure 4.3 shows a plot of intensity for sequences for each minimum number of mismatches with the U1A consensus sequence UGCAC. The sequences within each panel are sorted lexically along the *x*-axis. In the applications of this approach to protein binding microarrays (PBMs) the average intensity is high but decreases gradually with increasing mismatches (Carlson et al., 2010). A box plot of the intensities by minimum mismatch number in U1A (Figure 4.4A) shows that only the sequences containing an exact match have intensities greater than one standard deviation above the average. The unusual distribution of U1A may further support a role for the contextual tertiary structure described in section b (above) where in its absence only exact matches are bound with low affinity.

In contrast, the plot of intensities scale inversely with the minimum mismatch for the Sam68-like mammalian protein 2 (SLM2) as was expected for U1A. The box plot in Figure 4.4B shows that each additional minimum number of mismatches is accompanied by a statically significant decrease in binding affinity. When intensities are plotted in panels corresponding to their minimum mismatch number and lexically sorted by match sequence along the *x*-axis in the style of Carlson et al. (2010), the intensities for SLM2 in the RNAcompete data in Figure 4.5 provides a more information rich view of the specificity than the PWM.

Specific trends within the panels of Figure 4.5 reveal some higher order dependence in binding. Within the zero mismatch panel (Figure 4.5 panel ‘0’), we observe some dependence on recognition from sequences flanking the consensus recognition sequence. Within the single mismatch panel (Figure 4.5 panel ‘1’), we

observe that binding does depend on mismatches in the first order sequence as captured by the PWM. However, the sequence of flanking RNA sequences does significantly influence binding. The differences in score due to flanking sequences exceed the noise from replicate experiments where the standard deviation of the change in intensity between replicate measurements is 0.1 intensity units. The higher mismatch panels reveal that some sequences with two or more minimum mismatches bind with nearly as high an affinity as sequences with zero mismatches. These high intensity readings in the two and more mismatch panel suggest either that positions within the consensus sequence, which all had high information content in derived PWM, do not contribute to selective binding or that there are higher order relationships governing binding.

The mismatch plots for U1A and SLM2 (Figure 4.3 and Figure 4.5) both suggest that RNA tertiary structure does factor into the affinity of the RBP for a specific sequence. The marked differences between the mismatch plots for U1A and SLM2 demonstrate that the extent of dependence of RBP binding on RNA structure is protein dependent. The binding of SLM2 may be less dependent on RNA tertiary structure than the U1A example. The SLM2 case would be a good test case for the previously described ROC analysis if a structure were available for use with structure based predictions.

Experimental and biological factors make the RNAcompete assay a less optimal source of experimental data for comparison with computational predictions than the PBM approach for DNA from which RNA compete was adapted. The sequence library for RNAcompete was adapted to cover as large a single stranded binding region as possible without introducing the complication of RNA secondary structure. Adding in

secondary structure would increase the search space beyond what is reasonable for the current technology and add complications from RNA tertiary structure. The difference in the appearance of U1A data when plotted by mismatch (Figure 4.3) and the smaller improvement in binding when the consensus sequence in included (Figure 4.4A) demonstrates problems with interrogating affinity using the microarray approach. Because the computational predictions I compared with these data in section a (above) are performed using structures in a high affinity native conformation, the absence of structured RNA in the assay may contribute to the poor results I obtained for the ROC analysis (Figure 4.2). The limits on the RNA length and RNA structure present difficulties in interpreting the RNAcompete results.

The additional steps in the RNAcompete approach required to adapt a DNA microarray to RBP specificity measurements introduce some noise and alter the kinetics of the assay. The differences between the two techniques were discussed in section 1.c (above). In review, RNA compete has additional steps that are affected by binding kinetics. Additionally, the step where RNA is pulled down by the labeled RBP occurs in solution and thus occurs competitively where weakly binding sequences may be displaced. Together with the short heptamer sequences, the experimental differences contribute to greater noise in the measured binding intensities. The greater uncertainty in the intensities could also contribute to the poor result in Figure 4.2.

c. Lessons for improving specificity predictions

While it looks like comparison of predicted and experimental binding affinities should provide useful clues for improving the scoring functions for specificity and affinity prediction, the analysis proved problematic for the case of U1A. U1A pulled

down the unstructured RNA with low affinity leading to an unusual intensity profile. The other two RBPs for which structure and RNAcompete data are available, VTS1 and PTB, were complicated by an interface with a single unpaired base or by a complete protein construct containing four RRMs recognizing a region much greater than the heptamer sequences. However, analysis of RNAcompete results for the SLM2 case suggests that the affinities of some proteins can be well resolved and that higher order binding motifs may be important for RNA binding.

With additional structures and RNAcompete measurements, the approach I have outlined (Figure 4.1) can be used to improve the computational approach for specificity design purposes. With higher confidence in my approach, modeled structures may be used for my approach specific binding predictions. The inclusion of modeled structures would increase the number of proteins that can be compared. The comparison can also be expanded as additional motifs are solved experimentally using the RNAcompete technique. Identifying improperly predicted interactions will provide important information for optimizing the scoring function for specificity prediction. Improvements to binding motif predictions using approaches such as that described in this section also allow us to establish confidence in a specificity based approach to design.

C. Retargeting of RNA Recognition Motif to Bind Micro-RNA Precursors

Clear demonstration of the utility of the specificity predictions is best demonstrated with a successful design application. Previously (Chapter 3.H.2), I showed that my computational approach reproduced the expected position specificity demonstrated in experimental results with the PUF domain from Pumilio1. This result

demonstrated that it is possible to fundamentally alter the binding specificity of proteins with only minor alterations to the protein sequence (Y. Chen & Varani, 2011). Additionally, only small changes in local structure were required (Dong et al., 2011). The approach used to verify the computational approach can be extended to a design problem.

One potential application for a designed RRM is a specific binding to miRNA. A recent review found that several miRNA have a reported role in oncogenesis (Croce, 2009). Yu Chen and Jana Mandic in the Varani laboratory noticed that one miRNA expressed as a precursor to tumor formation exposes a single stranded sequence similar to that recognized by the RRM domain of the Fox1 protein. In this chapter, I describe my approach to designing the target specificity switch using the computational tools I have described. The binding of miRNA that are causal in the dysregulation of genes in cancer cells is an example of a potential role for designing protein domains for binding to RNA.

In this section, I explore the application of RBP specificity predictions to the design of protein-based drugs that bind specific RNA sequences. The known RNA binding domains could form the basis for domains that bind nearly any sequence. I adapt the specificity prediction tools to suggest changes to a RRM that would retarget it to selectively bind a precursor miRNA sequence.

1. Toward a Modular Platform for Sequence Specific RNA Recognition

Proteins that target RNA could provide the basis for a universal platform for binding to specific RNA sequences analogous to the zinc finger platform for DNA. A protein-based platform would be easier to develop than small molecule drugs targeting

RNA structures. While small molecules have been discovered that target specific RNAs (Disney & Guan, 2012), small molecules rely more on interactions with RNA structure than sequence and are not adaptable to new RNA sequences. Proteins may be used for sequence specific binding of RNA at unstructured regions of RNA. As discussed in Chapter 2.A, nature reuses a small number of RNA binding domains (RBDs) for binding most possible sequences and uses RBDs in multiple copies to increase specificity by recognizing longer RNA sequences. Only a few examples of designed peptides or proteins targeting RNA have been demonstrated in the literature. However, the case of designed zinc fingers for sequence specific cleavage of DNA demonstrates what might be possible with a designed library of RNA binding domains. Altering a compact binding domain such as an RRM opens up the possibility of creating protein drugs.

While no therapeutic uses for designed RNA binding domains have been described in literature, peptide-based approaches have been explored for binding RNA targets involved in viral infection. In the Varani laboratory, Davidson et al. (2011) showed that a short peptide fragment could be stabilized and optimized to bind a structured RNA essential to HIV replication. The cell permeability of the cyclic peptidomimetic molecule was demonstrated (Lalonde et al., 2011). However, even much larger peptide fragments may be stabilized for delivery to tissues (Harrison et al., 2010). New techniques for protecting proteins from proteolysis in the bloodstream and for delivering them to specific tissues are under active development (Antosova, Mackova, Kral, & Macek, 2009). The specificity with which peptides and proteins bind their target and the reduced concerns with cytotoxicity may justify the challenges with bioavailability.

The zinc finger is an example of a versatile platform for specifically binding nucleic acid sequences that is currently under development for use in gene therapy in humans. Zinc finger constructs have been designed from libraries to target many of all possible 18 base pair DNA sequences (Maeder et al., 2008). The zinc fingers from the library of engineered zinc finger proteins may be combined in tandem to recognize gene specific DNA sequences (Maeder, Thibodeau-Beganny, Sander, Voytas, & Joung, 2009). The affinity of the tandem zinc fingers can be predicted from the affinities of the component zinc fingers (Sander, Zaback, Joung, Voytas, & Dobbs, 2009). With predictable affinity to specific gene sequences, the zinc finger platform can be used to introduce a double strand break in nearly any gene (Miller et al., 2007). The approach has been used for gene therapy applications in zebra fish (Foley et al., 2009). The ability to rapidly create a construct that will target nearly any DNA sequence demonstrates the flexibility of the protein-based approach to specifically recognize nucleic acids based on sequence.

A promising application of designed zinc fingers is being tested for the treatment of HIV in humans. A designed zinc finger for creating HIV resistant T cells by specifically inducing a double strand break in the gene for the C-C chemokine receptor type 5 (CCR5) (Perez et al., 2008). The HIV virus uses the CCR5 as an entry point into the T cells. Individuals with a deletion in the CCR5 gene have been shown to be resistant to HIV infection. The zinc finger therapy disrupting CCR5 in human CD4+ T cells is currently in clinical trials (`http://clinicaltrials.gov/ct2/show/NCT00842634`). While delivering the designed zinc finger protein is not possible, a method for using adenoviral vector for delivering the gene for the protein to the target cells was

previously developed (Schroers et al., 2004). The zinc finger platform demonstrates a mature application of proteins targeting specific nucleic acid sequences and an alternative to delivering the protein to the target cells.

RNA binding domains may provide the basis for a creating a library of RNA binding proteins similar to the zinc finger platform for DNA binding. The domains listed in Table 2.1 all bind single stranded RNA in a sequence specific manner and may form the basis of a recognition library similar to the zinc finger system (section 1 above). Work by Tanaka group takes advantage of the modular nature of PUF domains and their one to one base correspondence between repeats and number of bases recognized (Cheong & Tanaka Hall, 2006). Thus the development of smaller proteins that specifically recognize RNA would be useful. However, a purely rational approach is not easily extended to proteins that are not modular with respect to the number of repeats and the positions of the site recognized.

2. Support Through Design

Designed specificity switches can be achieved through sequence changes to RBPs. In section Chapter 3.H, I discussed the experimental evidence that altering two positions in a PUF domain from the Pumilio1 protein allowed the position to be modified to be selective for each of the canonical RNA bases. By performing computational exploration of complete sets of mutations at the key protein positions, I reproduced the change in specificity preference *in silico*. I now seek to extend this approach, to discover which residues, when altered, allow for a specificity switch in a different protein domain binding single stranded RNA. We chose to alter an RRM domain to target a sequence in an important micro-RNA precursor.

Experimental and computational approaches to altering DNA binding proteins to recognize alternate DNA sequences have been reported (Buchholz, 2009). A library of zinc finger proteins was developed through selection that can be used in tandem as a near universal platform for binding DNA sequences (Maeder et al., 2009). More similarly design approaches such as foldX have been used to reproduce energy predictions from disease causing haplotypes (Alibés, Nadra, et al., 2010). More similarly, David Baker's research group has published a number of examples of design approaches to altering homing endonuclease binding specificity and affinity (Ashworth et al., 2010; Thyme et al., 2009; Ulge, Baker, & Monnat, 2011). Computational approaches to changing the binding targets of RBPs have not yet been reported in literature.

The work with DNA binding informs my approach, but I attempt to use changes in specificity to guide the residue mutation. Computational design has been used to evaluate the energy changes of mutated DNA binding proteins and to alter the DNA sequence bound by endonucleases. Alibés et al. (2010) demonstrate that their FoldX program can reproduce the PWMs for transcription factors previously predicted by Morozov et al. (2005) using Rosetta. Additionally, with a set of disease causing haplotypes of the Pax6 gene they show that disease-causing mutations have higher predicted $\Delta\Delta G$ values. The FoldX program is a linear combination of physical and statistical terms similar to Rosetta (Guerois, Nielsen, & Serrano, 2002; Das & Baker, 2008). The Pax6 case is an example of computationally assisted approach to understanding biologically relevant changes in binding energy caused by point

mutations. However, Alibés et al. (2010) do not check whether these mutations alter sequence specificity.

The Baker group has reported three experimentally verified examples of computational retargeting of homing endonucleases (Ashworth et al., 2006, 2010; Ulge et al., 2011). The reprogramming are examples of the application of the Rosetta Design (RD) approach (Das & Baker, 2008). RD uses a Monte Carlo approach to explore both structure and sequence space in an automated fashion. In order to select for a protein that best recognizes a selected DNA sequence, the desired recognized sequence is modeled in the structure of the original endonuclease. The program performs alternate sampling of sequence and structure space as it tries to optimize affinity. Checking scores with bases substituted to the other possible base pairs allows calculation of changes in specificity. Because of the allowed sequence and structure space, this approach results in a large number of candidate proteins. The designs are typically manually curated and experimentally checked.

The advantage of the design approach is in the relative ease of experiments that may confirm the design. In the protein-DNA problem, computational design targets could be chosen such that a successful redesign was detected using DNA cleavage assays (Ashworth et al., 2010). We did not come up with such an efficient and scalable approach for use with RNA binding proteins (RPBs). However, the binding of a modified protein with a supplied RNA sequence can still be determined by EMSA experiments with minimal difficulty.

Design must be performed in order to optimize both specificity and affinity. I describe the challenge of modeling specificity and affinity. Then, I describe how the

tools I described in Chapter 3.D can be used to guide a specificity switch in a stepwise manner.

a. Exploring specificity and affinity landscapes

Altering molecular recognition by proteins requires optimizing the modified protein to bind its RNA target tightly and to maximize sequence discrimination. The affinity of binding is measured in terms of binding constants. The sequence discrimination is captured by the specificity calculations I described in the previous chapter. A successful retarget would implement the desired change in specificity while preserving or enhancing the affinity.

The retargeting approach does not seek to quantify affinity or specificity on an absolute scale, but instead seeks to elucidate a path do the designed specificity. As the base is part of a larger binding interface, we may apply approaches that have been previously applied to protein-DNA binding. We assume that the affinity of the protein for the adjacent bases is such that the protein would continue to bind in the same frame if there were no energy contribution from the target base. We want to select the minimum number of residues to change such that the protein prefers another base at the target position without substantially altering the local structure of fold of the protein.

Optimizing affinity and specificity has been implemented in computational approaches for DNA targeting. Groups attempting a similar base swap in protein-DNA systems such as homing endonucleases (HE) have elaborated on the meaning of base preference. Each residue side-chain neighboring a base position may be thought of as contributing to specificity or to affinity (Ashworth & Baker, 2009). In choosing amino

acid swaps, the Rosetta architects try to optimize an amino acid for its affinity and its specificity. The Rosetta approach has concentrated on exploring large structure and sequence space in an automated approach to find a number of designs that optimize affinity and specificity.

Based on the experimental design work with PUF domains and RRMs, I seek to retarget positions through minimum changes to sequence and structure. Figure 4.6 diagrams an approach to exploring the effect of small protein sequence changes on specificity. The approach draws on the specificity approaches I described in the last chapter. I also build upon the approach to calculate affinity and specificity used by Rosetta (Ashworth & Baker, 2009).

The affinity is related to the optimality of each amino acid position for binding. The positional optimality for binding is a Boltzmann distribution of the scores of all possible amino acids at a position with respect to the bound nucleic acid.

$$a_j = \frac{e^{-\Delta G_{j,\mathrm{RNA}}}}{\sum_{i \in \mathrm{aa}} e^{-\Delta G_{i,\mathrm{RNA}}}} \tag{4.3}$$

This represents the affinity, *a*, of amino acid candidate *j* for the base whose affinity is being optimized. In practice, when dealing with small sequence changes and concentrating on amino acids where there is a direct side-chain to base contact, most of the change in residue energy following substitution is due to the interaction with the side-chain. Affinity can be visualized as a variation of a logo (Schneider & Stephens, 1990) with symbols for the 20 amino acids with symbol heights representing the favorability of that amino acid.

The specificity for binding base *k* can be inferred from the Boltzmann energy of the energies of the bases in the presence of a given amino acid substitution.

$$s_j^k = \frac{e^{-\Delta G_{j,\text{residue}}}}{\sum_{i \in \text{na}} e^{-\Delta G_{i,\text{residue}}}} \tag{4.4}$$

Ashworth and Baker (2009) define the specificity conferred by for nucleic acid position *k* of type *j*, s_j^k, as the Boltzmann distribution with respect to all possible bases at position *k*. However, residues are not necessarily additive and could affect or shield neighboring residues. Thus, I concentrate on the score of each base with respect to its entire environment. The relative specificity for the four canonical bases of a residue substitution can be visualized in terms of a logo position (Schneider & Stephens, 1990).

The definitions of affinity and specificity allow quantification of the effect of protein sequence changes. These definitions worked with automated searches in Rosetta Design for homing endonuclease (HE) designs (Ashworth et al., 2006, 2010; Ulge et al., 2011). To maximize specificity, HE designers maximize difference between the specificity with a test amino acid and the average Boltzmann energy with each of the 20 amino acids ($s_i - \overline{s}$). The scoring function and structure search parameters for the RBP design process is not yet ready for an automated search. However, with a scoring function that correctly predicts specificity the specificity and affinity landscapes (Figure 4.6) can guide a rational approach to retargeting a protein to a desired sequence.

Appling this approach to completely redesign the binding surface is unlikely to work with RBPs. Altering the target RNA for a RBP by even one nucleotide could cause

unintended energetic changes along the entire interface (Yamasaki, Nakamura, Terada, & Shimizu, 2007). This will be of greater concern at protein-RNA interfaces than for similar design problems for proteins binding DNA or protein since single stranded RNA is flexible. RBP often bind on the loop regions of RNA stem-loop structures (Svoboda & Cara, 2006). The extended conformation usually observed along a RNA interface does not likely exist in the absence of binding. To get a true estimate of the binding energy we would need to estimate the unbound structure and energy (Castrignanò et al., 2002). I expect considerations of long-range structure change in the protein will become important when several protein sequence changes are made or when the RNA interface is being homology modeled.

I try a more modest approach and try to visualize the effect of a few substitutions to the protein sequence on specificity at a binding site. Here we explore whether this strategy may be ported to the realm of protein-RNA interactions by performing a stepwise search of the specificity landscapes. This approach allows biochemical intuition to guide design choices guided by additional information from the computational approach.

b. Stepwise exploration of specificity landscapes

The calculations with the PUF domain from Pumilio1 described in Chapter 3.H (above) suggested that a couple of well chosen changes to protein sequence could alter the target specificity of an RNA binding protein. In order to evaluate this approach, we attempt to visualize the effect of single and double mutants of proteins at residues contacting a target base edge. In my approach, I query the effect of all single mutants of base contacting protein residues and evaluate the effect on specificity.

While the potential sequence and structure space of the RNA binding site of a protein is very large, we may simplify the problem by remaining close to the native sequence. The lesson from the Pumilio1 example (Chapter 3.H) is that key protein positions determine the base preference. The set of possible single protein mutations at positions with direct side-chain to base interactions is relatively small. We may evaluate each of the contacting protein residue positions to see which play the greatest role in determining specificity. For each candidate specificity-conferring position, I substitute each possible alternate side-chain and perform the additional steps needed to predict specificity at that location. By visualizing the effects of single mutations on specificity we may rationally choose a step that alters the specificity in the desired direction.

As before, a PWM for a position is determined by substituting all canonical bases into the target position and repacking the surrounding residues. In this case the repacked residue includes all the residues in direct contact with the base position where specificity is being altered. The base is scored against all surrounding residues and bases. I also monitor the scores of the mutated residues with respect to its environment and pair scores between residues and the target base. By storing these scores from each of the Monte Carlo minimization runs, I can predict specificity and affinity data from the set of mutations and scoring runs and display the single mutation landscape in a variety of 'views' (Figure 4.6).

From the Pumilio1 PUF repeat 6 example (Chapter 3.H), I noted that the specificity switch was predominantly caused by a single protein residue mutation. When I looked at the complete set of double mutations, I noted that one of the substitutions explained much of the specificity switch. The second mutation can be

thought of as being cooperative or stabilizing. While the specificity of a RBP for a specific base position are not as simple as the case from Pumilio1, I may select test cases for positions with a few strong side-chain to base interactions.

The scores logged from the single mutant calculations can be displayed in two views of the sequence space surrounding a base of interest. In the affinity view, we obtain for each candidate side-chain and each candidate base a weight column reporting the probability of each residue being preferred (Figure 4.6 'affinity view'). When displayed as a logo, the residue one-letter-codes are shown in order of their contribution to affinity. A design application would suggest a residue most preferred with the base we want to be recognized at this position. In an ideal case, the protein-RNA interface would be like a three dimensional puzzle where that residue is only preferred with the base of interest. Unfortunately, the relationship is often more subtle. The specificity view of the substitution landscape helps guide our substitution choices.

The specificity view shows how the base preference changes with each protein residue mutation. The mutation landscape can be seen as a logo for each mutated position with possible residues on the *x*-axis of each logo (Figure 4.6). This allows a quick visual survey of the possible primary mutations. If a residue substitution clearly alters the specificity such that the designed target base is preferred, that mutation path warrants further exploration. However, the selection of residue should also be guided by the rank of the residue in the affinity view. The advantage of this specificity view is that even if the primary mutation does not completely switch the preference, we may see changes in the magnitude of this preference.

For each candidate primary mutation, we can select complimentary mutations to help stabilize the specificity switch. The specificity switch is performed by selecting and performing a protein sequence mutation. The affinity and specificity views can be regenerated to elucidate the next mutation. I expected that as in the Pumilio1 example, there exist one or two protein sequence changes that alter the protein to select any base.

3. Targeting MicroRNAs

A possible novel application of retargeting a compact RNA binding domain is creating proteins that selectively bind precursor microRNA (miRNA) sequences. miRNAs play many important roles in post-transcriptional regulation (Bushati & Cohen, 2007). Over-or under-expression of certain miRNAs can be the first step in a cascade of post-transcriptional events guided by complementary binding of the miRNA with gene transcripts. Thus, a molecule that tightly and specifically binds to the precursor miRNA may inhibit the post-transcriptional cascade by preventing production of the functional mature miRNA.

MicroRNAs (miRNAs) provide a mechanism for regulating gene expression. RISC(RNA-induced silencing complex)-associated miRNAs recognize an mRNA using sequence complementarity. The RISC may inhibit the translation of or initiate the degradation of mRNAs (Fabian et al., 2010). Humans are known to express around 1000 miRNA (Filipowicz et al., 2008), and Bioinformatic analyses show that up to 30% of protein-coding genes in humans are regulated by miRNAs (Filipowicz et al., 2008). Several miRNAs have been linked with gene dysregulation associated with cancer (Esquela-Kerscher & Slack, 2006). In the case where overexpression of a miRNA is implicated in tumorigenesis, a designed molecule that specifically binds precursor

miRNA with high affinity and reduces their expression could form the basis for a therapeutic or investigational tools.

The importance of miRNAs as a target is illustrated by the variety of small-molecule and macro-molecular approaches being developed to target miRNA and its associated proteins (Reichel et al., 2011). Complimentary siRNA-like molecules would be the easiest way to specifically inhibit miRNA, but RNA is subject to degradation pathways and is hard to deliver (Ruth, 2011). Synthetic oligonucleotides such as peptide nucleic acids (PNAs) can bind complementary nucleic acid strands can be delivered to cells (Pooga et al., 1998), and indeed PNAs have recently been designed to target miRNA targets (Gaglione et al., 2011). PNAs have the same drawbacks as other oligonucleotide-based therapeutics. Small molecules could offer a one-off solution to target highly structured RNA, but such an approach would be exceedingly difficult even using fragment-based approaches to drug design (Disney & Guan, 2012). Protein-based approaches are an opportunity to specifically bind multiple base regions of miRNAs and to take advantage of developments in protein-based drugs.

RNA binding domains (RBDs) are an ideal platform on which to design molecules to specifically bind the precursor miRNA by sequence. Many RBDs target loop regions on RNA hairpin structures (Svoboda & Cara, 2006; Westhof & Fritsch, 2011). The domains offer a framework for sequence specific recognition of multiple bases since domains such as the RRM are already known to recognize many of the possible tetramer sequences (Auweter et al., 2006). The development of a complete binding library such as that for zinc fingers binding DNA (section 1 above) would make RBPs a more compelling approach to the sequence specific inhibition of active RNA

molecules in cells. At present for a miRNA of interest, we can search known preferred binding sequences for small RBP domains such as RRM domains for a good starting structure. A good starting structure would recognize as sequence similar to an exposed single stranded sequence on the precursor miRNA we seek to inhibit.

A molecule that targets miRNA must be able to specifically recognize a sequence in a small mostly unstructured RNA. The loop on the precursor miRNA is possible target for a molecule that can be delivered to cells since several steps of miRNA maturation occurs in the cytoplasm (Soifer, Rossi, & Saetrom, 2007). It is plausible that the sequence specificity may be altered either through experimental selection techniques or through computationally guided rational design as described in the preceding section (section 2.b above).

Even with significant difficulties in delivering proteins of the size of RBPs to cells, the protein-based approach offers advantages over small molecule approaches. Since protein components of miRNA processing and the RISC complex are shared among many miRNA, the miRNA is itself the only specific target (Wahid et al., 2010). Small molecules have so far been unable to target a specific miRNA or precursor miRNA sequence with high specificity. The chemical and biological diversity of a protein would allow us to recognize a sequence with higher specificity (Mason, 2010). The absence of toxicity from peptide drugs and a lower chance of invoking an immune response make a protein based drug more favorable than nucleic acid based therapeutics. Artificial nucleic acid analogues may be toxic and are very difficult to deliver. While unmodified peptides have not shown much success as drugs, modified versions can be made which resist degradation. Additionally, the protein may be tied to an existing delivery system.

4. Redesign of Fox1 specificity

The Varani group sought to modify a RNA recognition motif (RRM) to recognize a miRNA target with minimal redesign. We found that the preferred binding sequence of the first RRM domain from the Fox1 was similar to a single stranded region within certain miRNA precursors. The Fox1 RRM domain could be adapted to specifically bind a miRNA precursor with altered binding preference at two nucleobase positions. If the lessons from the Pumilio1 example discussed in the previous chapter (Chapter 3.H) hold, then for each of the two positions there may exist single or double mutations that alter the Fox1 RRM protein to specifically bind the target sequence without significantly altering the structure of the RRM domain. I employed an approach, similar to that used in the computational validation of the Pumilio1 retargeting example, to identify RRM sequence changes that would alter the sequence specificity.

Problem statement: The role several miRNAs play in cancers makes them an attractive target. For example, miR-21 inhibits apoptosis and increases tumorigenicity (Croce, 2009). Inhibition of apoptosis in glioblastoma may result from the expression of miR-21 (Ciafrè et al., 2005). Additionally, a positive feedback loop involving miR-21 may form an epigenetic switch linking inflammation to cancer as well (Iliopoulos, Jaeger, Hirsch, Bulyk, & Struhl, 2010). Inhibiting these miRNAs may represent a viable therapeutic approach to interrupting the signal cascade in the cancers involving miRNAs.

Varani group members Dr. Yu Chen and Javier Castellanos performed a comparison of the consensus binding sequences of well-characterized RNA recognition motif (RRM) domains with miR sequence from the miRBase (Griffiths-Jones, Saini, van Dongen, & Enright, 2007). The first RRM in the Fox1 protein (gene: A2BP2, UniProt ID:

Q9NWB1) is highly specific for the consensus sequence 5'-UGCAUG-3'. The native target of Fox1 differs from the sequence of the target miRNA precursor at two positions. Figure 4.7 shows the target sequence similar to the native target of Fox1 emphasized with bold type. The positions where the preferred Fox1 target differs from the target sequence are indicated in red. Figure 4.7 also shows the RNA in Fox1 structure 2err using the same color scheme to indicate the binding pockets where the RRM binding preference would need to be altered. I note that each of the nucleobases colored in red (B199C and B202G in structure 2err) are characterized by direct contacts with protein side-chains. Thus, retargeting Fox1 to a sequence altered at these two positions is a reasonable task for the computational approach.

To target the protein to specifically bind the intended miRNA precursor, the base preference needs to be changed at two positions. Henceforth, I will refer to the base positions by their position in the RNA hexamer recognized by Fox1 and the one-letter-code preferred at that position (e.g. B199C as 3C and B202G as 6G). Altering the preference for a cytosine at position 3 to a preference for an adenine (3 C→A) would be the first step in retargeting Fox1. Retargeting would be completed by additionally altering the preference for a guanine at position 6 to a preference for a cytosine (3 G→A). The lack of protein contacts with the RNA backbone and ribose at these positions make them conducive to design with the computational tools I have described.

Computational redesign of Fox1: I applied the computational substitution and repacking tools and the Rosetta scoring function to identify residue positions where mutation would alter binding preference. I employ the approach described in section 2

(above), to search for residue positions where single and double amino acid mutations may retarget the Fox1 domain to recognize the desired base at that position.

For the base at residue ID B199 in structure 2err, I identified five residues whose side-chains could make contact with the base: PHE A126, ASN A151, ARG A153, GLY A154 and LYS A156 (Figure 4.8). With the native C B199, the base makes no contacts to the protein backbone. Thus, I expect that specificity may be switched by the protein sequence changes at one or more of the residue positions enumerated.

As outlined in section 2.b (above), I explored the specificity implications of all possible single residue mutations. Each residue mutation was performed 20 times with each canonical RNA base at position 3 (B199). The protein side-chains were repacked around the base position, and the base was allowed to explore a few degrees of freedom around the glycosidic bond. The scoring function retained scores for each base and residue with respect to the complete molecular environment as well as residue pair scores for each residue with respect to the base position. Storing the scores allowed the exploration of the specificity and affinity landscapes with respect to sequence space.

I used an affinity view and the specificity view to identify a primary mutation that would alter the base preference at position B199 from C to A. The affinity view (Figure 4.9) ranks the residues at each position by affinity for the case where each of the four canonical bases is recognized at position 3 (B199). I note that the order of affinity is most dependent on size constraints and that the order of residue affinity is most dependent on whether a purine or pyrimidine is at B199. Individual side-chains do alter the relative preference for the bases.

Ideally a single mutation to the protein structure would significantly change the specificity such that the desired base is preferred at the RNA position. From this, I sought residues with high affinity for A but low affinity for the other bases, especially for the native C. Figure 4.9 suggests that a TYR at A153 may alter the specificity to preferentially bind U since TYR ranks high with the U at position 3 but not as highly with the other bases. This view provides a sampling of the most favorable residues. Since I am comparing the relative residue scores with respect to a fixed base and not comparing the score of that residue with respect to each base, this view does not provide a quantitative view of specificity. Instead this view allows us to determine which residues may be substituted without significantly altering affinity.

In the specificity view (Figure 4.10), the specificity change resulting from each single residue mutation can be easily visualized. I identified some residue mutations that increase the preference for A at position 3 (B199), but only one mutation results in a complete specificity switch to favor the adenine: A156 LYS → TRP. TRP did not rank high for position A156 in the affinity view. Thus we infer that this substitution may significantly alter binding affinity at this site. However, additional mutations could help stabilize the preference for A and perhaps increase affinity.

I performed a second round of mutation calculations with position A156 mutated to TRP. A specificity view for the second mutation is shown in Figure 4.11B. These calculations and the structure in Figure 4.11A suggest that mutating the residue at position A151 to a bulkier or charged residue would make an A at position 3 less favorable. Figure 4.11B suggests that modifications to 2err positions A153 or A126 would not significantly alter the preference for A, but that modification of these

positions could change the degree of preference for A and alter discrimination against the other bases. For example, altering A153 ARG to a PHE or HIS appears to make the desired purine base more favorable with respect to the native C pyrimidine. This provided some predictions that could be tested experimentally.

experimental validation: Dr. Yu Chen performed the A156 K → W mutation experimentally. Expression of the protein with A156W was low and he could find no evidence of binding. Since the TRP residue did not rank well in the affinity view the low binding was not particularly surprising. Since specificity could not be measured, it remains unknown whether this substitution actually selects for A. The same procedure was performed for position B202 where we sought to alter the base preference from G to C. The specificity and affinity plots did not show a clear residue mutation that would provide the desired target switch at this location.

While the single switch prediction did not work as predicted, the expectation is generally that the predictions provide guidance and not a definitive design. Successful redesigns for homing endonucleases (HEs) binding DNA have generally proposed dozens of redesigns and have generally been flexible about which design site and nucleobase switch was achieved (Ashworth et al., 2006). The greater constraint on position and identity of the specificity switch in this Fox1 example made this a more difficult task. Additionally, the HE retargeting examples were accompanied by an easier to perform and more sensitive assay based on the emergence of splice product (Ashworth et al., 2006, 2010). The specific RBP redesign problem we attempted may thus be a more challenging task.

Additionally the RRM motif is a particularly challenging motif to work with from a structural perspective. The prediction of RRM binding is a more challenging task than the PUF domain example because of the larger variation in interface structure. While many RRMs conserve a stacking interaction with the central two bases of the typical tetramer RNA recognition site there is a large variation in the amino acid positions that make contact with the nucleobases (Cléry, Blatter, & Allain, 2008). Some RRMs that conform to the Pfam motif specifically bind RNA using loops adjacent to the typical β-sheet binding surface (Dominguez, Fisette, Chabot, & Allain, 2010). Altering RRM binding specificity at a single RNA position should not significantly alter the overall binding path of the RNA or the residues participating in the specific binding. Since the RRM presents a less rigid binding pattern than some other domains, a higher degree of RNA flexibility may need to be modeled. Still, the compact binding surface and the known binding repertoire of RRMs make them an attractive starting point for design applications.

The most productive path forward with the computational approach is to explore more of the mutations for the B199 position that increase the probability of binding A. Additionally, as discussed with respect to the Pumilio1 example (Chapter 3.H), accurately modeling local RNA backbone flexibility may help improve specificity predictions made using the computational approach. Experimental approaches to altering specificity such as *in vitro* selection remain a viable alternative to small scale computational design for problems such as the Fox1 retargeting problem.

***in vitro* selection – what was missed computationally?** Dr. Yu Chen applied the *in vitro* selection methods he developed to discover mutations to the Fox1 domain that

allow binding to the desired sequence. The *in vitro* selection technique allows coupling of the mutation with its genotype (Y. Chen, Mandic, et al., 2008). The adenine at position 3 (3 C→A), was bound by a Fox1 domain with two substitutions at that position: A151 N→S and A152 E→T. The second position was retargeted to guanidine (6 C→G) with two additional protein mutations: A118 R→E and A147 E→R. The modified protein binds the target sequence with a K_d of 100 nM. Thus, as in the Pumilio example the preferred binding sequence for a RBP domain may be altered at a position with just two mutations.

I performed computational validation of the mutations discovered by the selection methods to verify that the computational method captures the changes to specificity. Figure 4.12 shows the effect of selected double mutations to residue positions contacting RNA positions 3 and 6. The calculations show that the double mutations near each binding site increase the probability of binding the target sequence. For position 3, A151 N→S and A152 E→T increase the probability of binding an A. Additionally the difference in overall prediction for these substitutions as compared with those in Figure 4.10 may be due to the more limited search of side-chain conformations used in the present calculation. For position 6, A118 R→E and A147 E→R increase the probability of binding a C. The computational method correctly predicts the direction of the change in binding probability. However, in neither case does the computational approach predict the specificity switch observed *in vitro*. This suggests that the conceptual approach of using the computational calculations as part of a rational design approach is valid, but that the scoring function or the limited search of RNA structure space was insufficient for the RRM prediction in this case.

Summary of computational approach to Fox1 redesign: The retargeting of the Fox1 RRM domain to a microRNA precursor target sequence demonstrates the feasibility of a possible therapeutic application of RBP design. I applied the computational tools I developed for specificity prediction and validated with the Pumilio1 PUF repeat 6 example to the task of altering the specificity of Fox1. The approach I developed allows rapid calculation and visualization of the effect of protein single mutations on the nucleobase specificity at a RNA binding position. Since we expect that the specificity at many RBP binding positions may be altered by well-chosen single and double mutations to residues that make direct side-chain contacts with the nucleobase, my approach offers a reasonable and unique approach to expanding the binding repertoire of common RBP domains such as the RRM.

In the example, the computational design where the specificity switch was predicted did not show significant binding in the biological assay. The computation suggested that the mutation A156 K→W would switch the specificity at RNA position 3 (B199) from a preference for A to C. However, the computational method also predicted that affinity would be reduced with this mutation. Given the accuracy of the current scoring function as benchmarked in Chapter 3.G, it is not surprising that the specificity landscape at a single RRM position is not perfectly predicted.

I additionally looked at the specificity calculations for the native RRM and double mutants discovered through *in vitro* selection. The double mutants that worked with the selection experiment increased the predicted probability of binding the target sequence, but the specificity switch was not predicted. The failure to predict the residue mutations found by *in vitro* selection is thus due to failure to predict the complete

specificity switch and to my failure to include residue A152 in my initial computational explorations. The results of these comparisons, suggest that with an improved scoring function, the computational approach may still prove useful in guiding RBP retargeting such as that proposed for retargeting Fox1.

D. Summary

In this chapter, I discussed ways to better analyze the performance of scoring functions applied to predicting RBP specificity and discussed an approach to demonstrate the biological utility of the specificity prediction tools. State of the art microarray approaches to measure RNA specificity provide needed data for improving the performance of structure-based scoring functions in specificity prediction. The amount and quality of data, however, remains insufficient because of greater challenges with measuring RBPs than those for measuring DNA binding proteins with protein-binding microarrays. The only example for which both a single domain binding motif and known structures in complex were available, U1A, proved to be a particularly challenging example due to structural transitions upon binding that are well documented in the literature. The specificity tools provide a compelling approach to retargeting RNA binding proteins as more such data becomes available.

The Varani laboratory is pursuing a RBP retargeting application using *in vitro* selection and computational methods. Unlike most small molecule approaches, a method for generating proteins specifically binding short single stranded RNA sequences would be generally applicable. Furthermore, success would demonstrate an understanding of RBPs and be practically useful to biologists.

I adapted concepts elaborated in the previous chapters and concepts from the Rosetta approach to designing homing endonuclease to retarget the Fox1 RRM domain. I showed that in a conservative approach to retargeting an RBP, tools for predicting specificity from structure might be the basis for a step-wise computationally guided rational approach to design. My approach predicted a single mutation to Fox1 that could alter a binding site usually recognizing C to recognize an A needed to target the loop region of a precursor miRNA. A second round suggested additional mutations that would reinforce that preference. Experiments did not show significant binding of the altered Fox1 to the target sequence and, thus, could not demonstrate a change in specificity. I repeated some calculations using double mutations demonstrated to allow Fox1 to bind the target discovered using the by the *in vitro* selection approach. The calculations captured changes in binding probabilities that would make binding the target sequence more probable, but failed to predict a complete specificity switch. The results suggest that the approach may be valid but that improvements in the scoring function and a more comprehensive exploration of mutations that alter specific binding are necessary.

I demonstrate progress on the problem of structure-based computational design of RNA binding proteins. I demonstrated that while no code exists for predicting nucleic acid binding specificity from protein sequence or structure, completely or partially statistical scoring functions can be used to predict binding specificity. Furthermore, specificity prediction is the key component of designing RBPs target alternative sequences. I argued that this limited design application could form the basis for

designing a class of proteins that can be used bind arbitrary RNA sequences in a sequence specific manner.

E . Conclusions

RNA binding proteins (RBPs) play a central role in defining the phenotypic state of the cell (Junhyong Kim & Eberwine, 2010). RBPs participate in all stages of mRNA processing and trafficking and often recognize the RNA in a sequence specific manner (Mittal et al., 2009). The levels of protein in the cell is largely determined at the point of translation by RBPs whose own genes are tightly transcriptionally and translationally controlled (Schwanhausser et al., 2011). All organisms reuse a small set of RNA binding domains with ancient origins for sequence specific binding of RNA (Kerner et al., 2011; Messias & Sattler, 2004). Thus, the ability to predict the RNA sequences bound by RBPs is central to our understanding of gene expression in cells.

I developed computational tools for predicting the sequence specific binding of RNA from protein structure. Structure provides a rich source of information about the physical interactions responsible for the specificity and affinity of the protein interaction with RNA. Structure information has been used in predicting specific interactions between proteins and DNA and between proteins and peptides (King & Bradley, 2010; Morozov et al., 2005). I explored statistical sampling approaches to structure prediction and scoring employing either purely empirical terms or employing a mixture of empirical and physics-based energy terms (Bernard & Samudrala, 2009; Leaver-Fay et al., 2011; Robertson & Varani, 2007). I evaluated the scoring functions and specificity prediction approach using a program developed in the Varani laboratory and with customizations to the Rosetta molecular modeling program.

The key tests involved comparison of the structure based prediction of binding motifs and in the recapitulation of design applications. I showed that the empirical scoring functions (all-atom and generalize) and the Rosetta scoring function substantially predict the independently reported binding motifs from the altered and repacked structures. However, only the Rosetta approach performed well enough to warrant application to design tasks. I showed that the computational approach recapitulated most of the dependence between Pumilio1 PUF repeat 6 double mutants and sequence specificity. I also proposed and applied a method for altering the sequence specificity of the Fox1 RRM domain to recognize a miRNA precursor that could be a therapeutic target in cancer.

The scoring function and design approach needs further development in order to achieve higher accuracy needed for the redesign of RNA binding proteins. The three principal improvements to the scoring function that are likely to improve binding motif prediction are (1) better exploration of local RNA structure, (2) the inclusion of intramolecular scoring terms that enforce correct RNA structure, and (3) improvement to stacking interactions, especially cation-π interactions. These improvements impact both of the central difficulties of statistical scoring functions, namely, the correctness of the scoring function and the algorithms used to search sequence space.

Each of the improvements I suggest address deficiencies I encountered in applying my approach to test cases. Principally, the exploration of local RNA backbone structure appears to be important for correctly differentiating between nucleobases based on chemical properties rather than on size. The ability to correctly minimize the RNA backbone torsion angles may be sufficient for improving specific binding

predictions with single stranded regions of RNA. Recent works have shown that the five RNA backbone dihedral angles can be reduced to two pseudo-angles and treated similarly to the protein backbone (Duarte & Pyle, 1998; Murray, Arendall, Richardson, & Richardson, 2003). Secondly, the RNA intramolecular scores are critically important to predicting specific binding with RNA. These terms are being developed for Rosetta but were not integrated into the present work. Thirdly, from the structure recovery tests, neither the empirical scoring functions nor the Rosetta scoring function recovered the majority of stacking interactions with bases. The stacking interactions (cation-π and π-π) are critically important to affinity for single stranded regions of RNA. Additionally, the stacking interactions have been demonstrated to have clear energetic differences between the bases (Morozova et al., 2006). Improvement to these aspects of the scoring function should allow better performance in design applications.

The rapid growth of structure data and of data from experimental assays of sequence specific interactions of proteins with RNA, makes this the ideal time to further develop computational approaches to specificity prediction and design. The relative difficulty of studying RBP interactions of RNA and the relatively finite structure space employed by nature suggest that computational approaches will play a central role in understanding these interactions. Additionally, designed proteins will confirm our understanding of these interactions and may open up new avenues for therapeutically regulating RNA expression levels in cells.

Figures

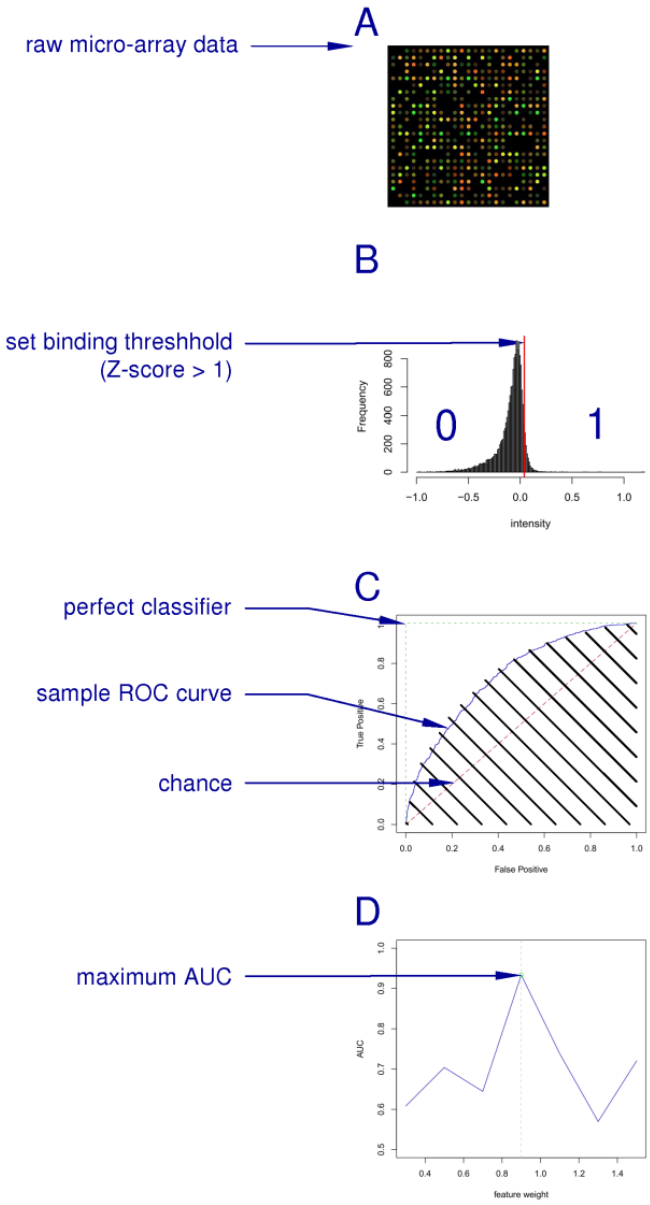


Figure 4.1: Schematic of direct evaluation of predicted and microarray confirmed binding sequences. The emergence of exhaustive binding profiles for RNA binding proteins allow a clear assement of our classifier. The microarray data yield relative binding affinities to all possible heptamer sequences with a small fraction of sequences clearly binding. The ROC curve (**C**) plots true positive against false positive with the ideal curve scoring all binding sequences before encountering a false positive (short dashes) and chance represented by the diagonal (long dashes). (**D**) The area under curve (AUC) summarizes the curve and can be used for optimization of parameter weights.

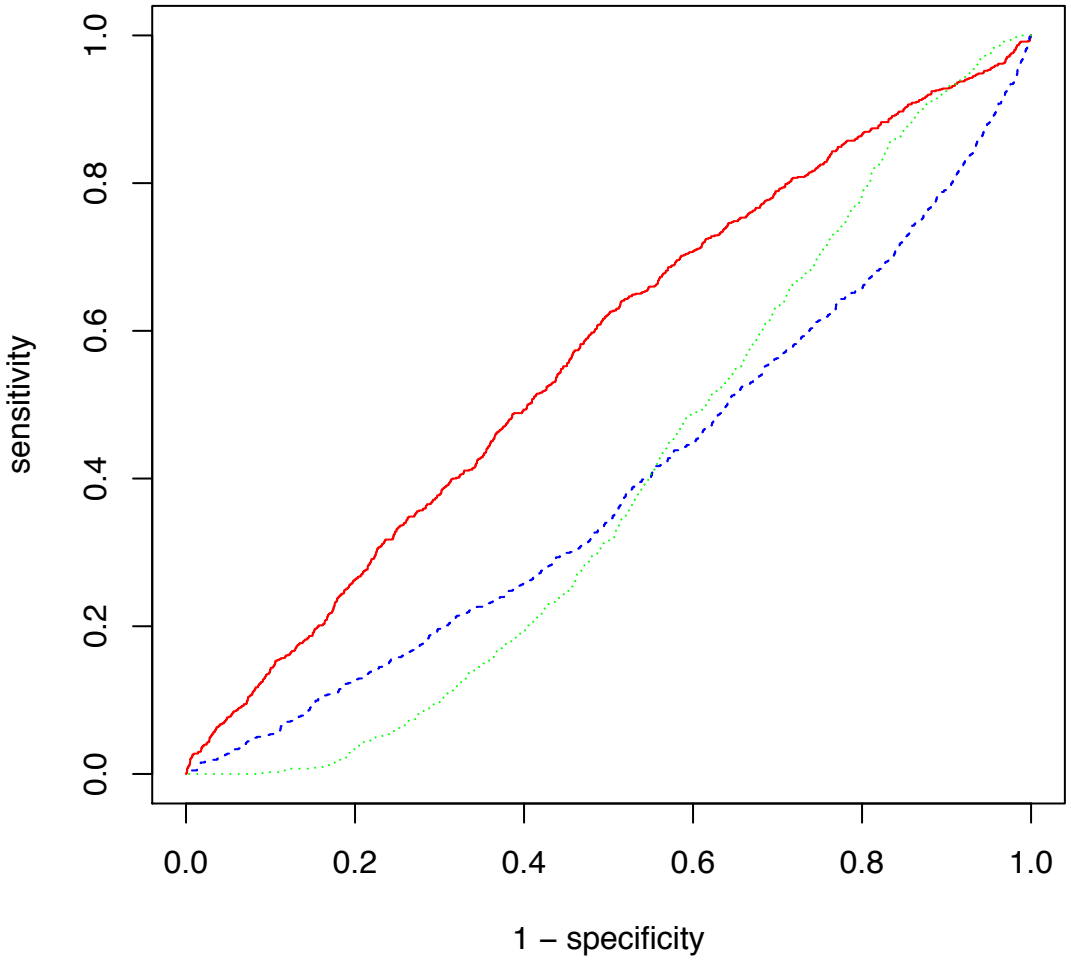


Figure 4.2: Direct comparison of predicted binding sequences for U1A with microarray intensities. The receiver-operator characteristic (ROC) plots the ability of the various computational methods to discriminate sequences that experimentally bind to U1A from sequences that do not. I performed binding calculations using structure `1aud`. I compared calculated interface scores with raw values extracted from the complete heptamer microarray binding data obtained by Ray et al. (2009). The curve assumes that the protein that recognizes a sequence of length 7 binds the complete sequence with perfect specificity at each site, although this is most likely not true. I created a receiver operator characteristic (ROC) for the Rosetta (solid red), generalized (dashed blue), and the all-atom (dotted green) scoring functions. The areas under curve (AUCs) were 0.568, 0.394, 0.398, respectively. Since the pure statistical scoring functions received AUC values less than 0.5, the statistical methods were even worse at finding binding sequences than a random guess.

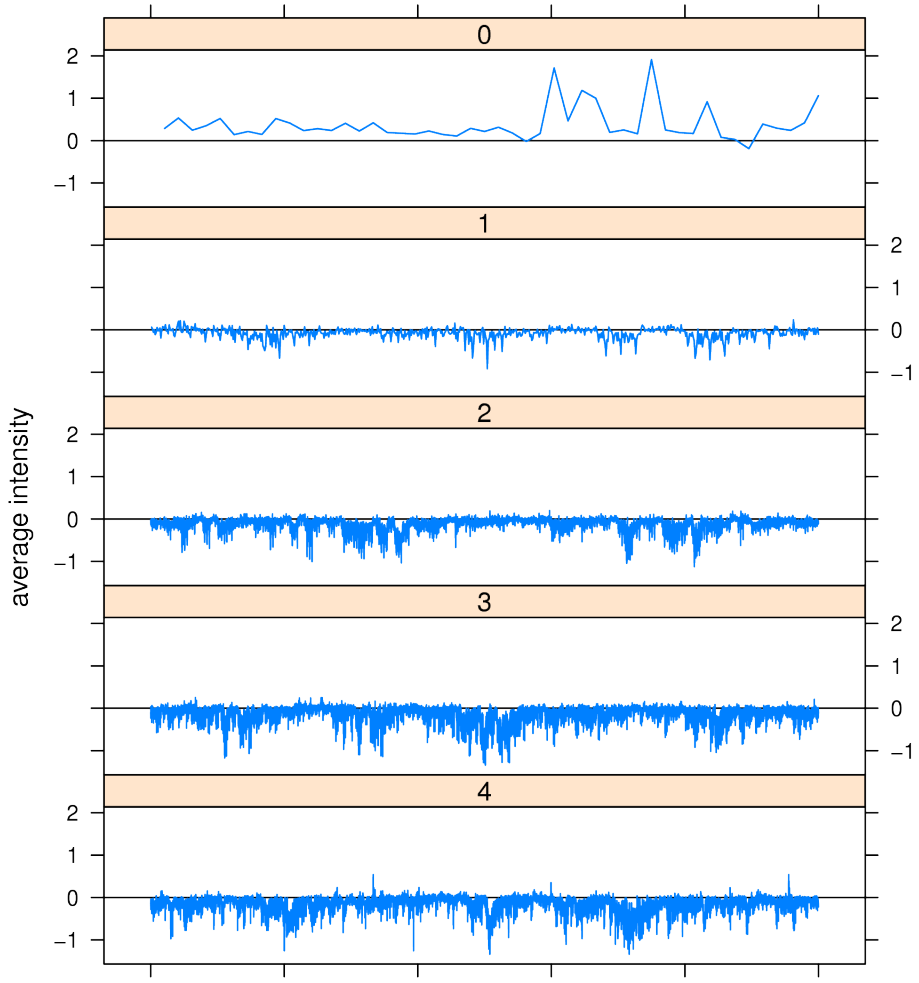


Figure 4.3: Correlation between microarray intensity and mismatches with U1A consensus sequence. I used a technique developed by the Ansari lab (Carlson et al., 2010) to visualize individual intensities and sequence dependent binding in the microarray binding data for U1A (Ray et al., 2009). The technique divides the experimental binding affinities for a recognition sequence seed of length N into N subplots (alternatively displayed as rings) by finding the minimum number of mismatches between seed and target sequence. The panel label corresponds to the minimum number of mismatches. The entries in each panel are lexically sorted by match and flanking sequences to reveal affinity patterns. We used the consensus binding sequence UGCAC as a seed for the plot. Ideally, the plot would show the affinity decreasing with increasing minimum number of mismatches and would show stronger decreases when a particular substitution is more unfavorable. In the case of U1A only exact matches appear to have significantly greater intensities than the mean (Figure 4.4A below) and thus looks different than what would be expected for binding that is dependent more on sequence than structure (Figure 4.5 below).

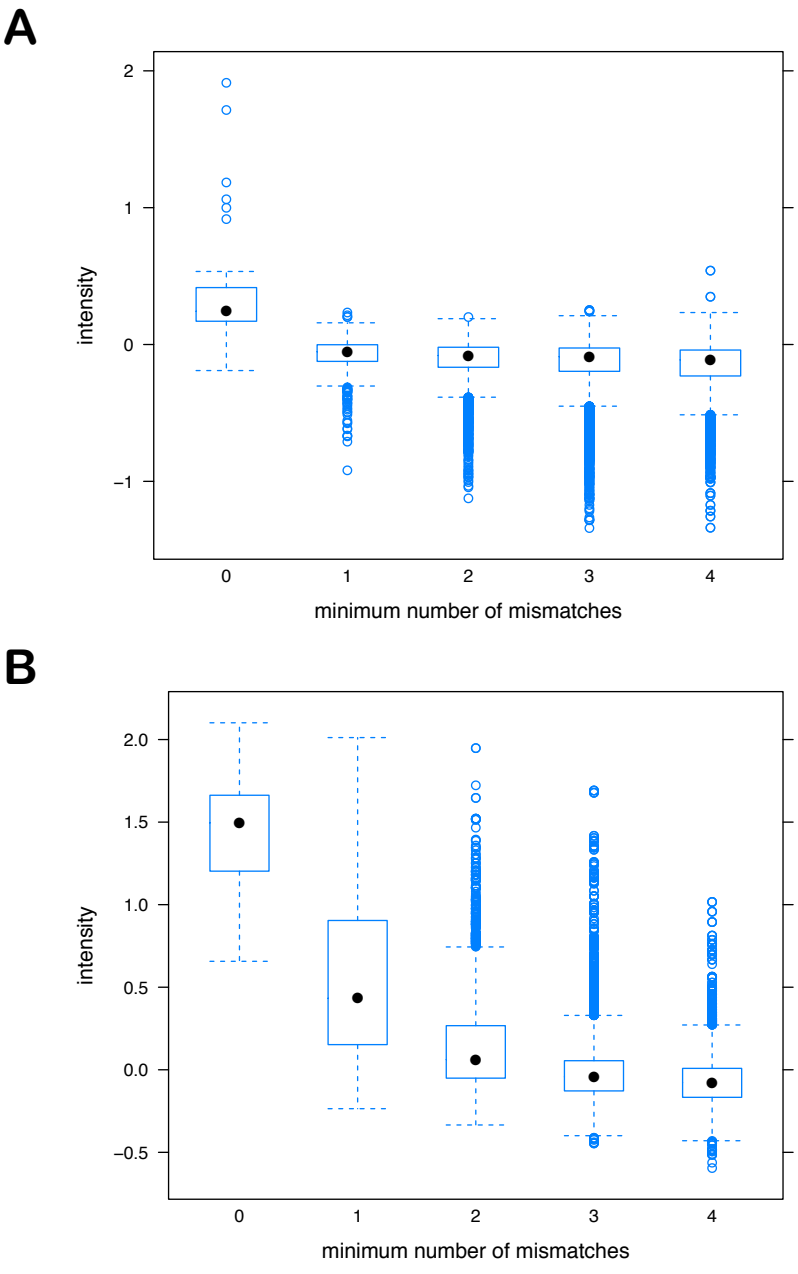


Figure 4.4: Summary of microarray intensities for U1A and SLM2 with mismatch number.
The boxplots show the distribution of microarray intensities in the RNAcompete experiment (Ray et al., 2009), partitioned by the minimum number of mismatches (x-axis) for the aligned consensus sequence to the microarray heptamer sequences. (**A**) Measured intensities due to RNA bound by U1A by minimum number of mismatches to UGCAC. (**B**) Measured intensities due to RNA bound by SLM2 by minimum number of mismatches to AUAAA.

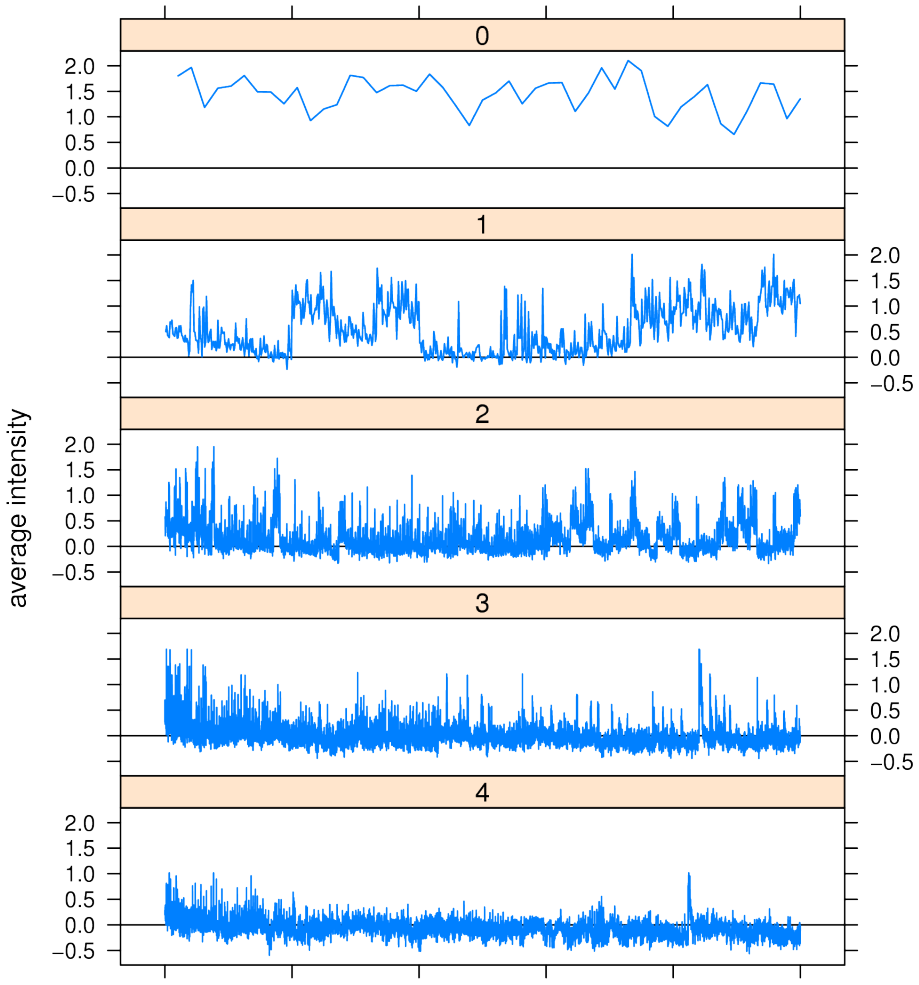


Figure 4.5: Sorted mismatch intensity plot for SLM2 using the consensus sequence AUAAA.
As in Figure 4.3 each panel displays the average intensities for heptamers with the corresponding minimum number of mismatches (**0**, **1**, **2**, **3**, and **4**) with the consensus sequence that with any alignment between the heptamer and the consensus sequence. Within each panel the heptamers are lexically sorted by mismatching pentamer then by the flanking sequences. In panel **0**, heptamers containing an exact pentamer match have intensities a few standard deviations better than the mean intensity but do exhibit some dependence on the flanking sequences. In panels **1**, **2** and 3, we observe that as expected from the derived PWM (`http://rbpdb.ccbr.utoronto.ca/experiments.php?prot_id=1423`) there are clearly positions and sets of positions where mismatches are tolerated. Additionally, the intensities suggest that a higher order relationship among favorable positions may play a greater role in binding affinity.

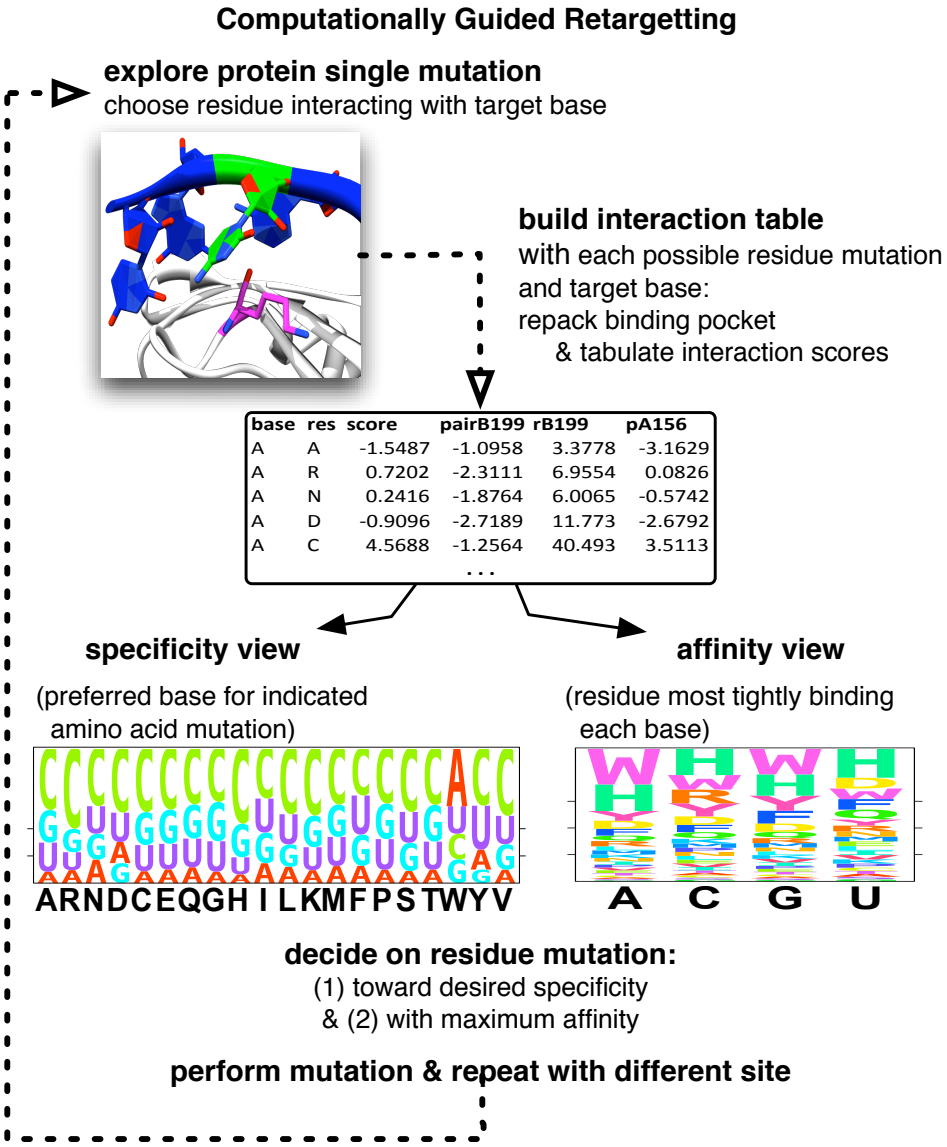
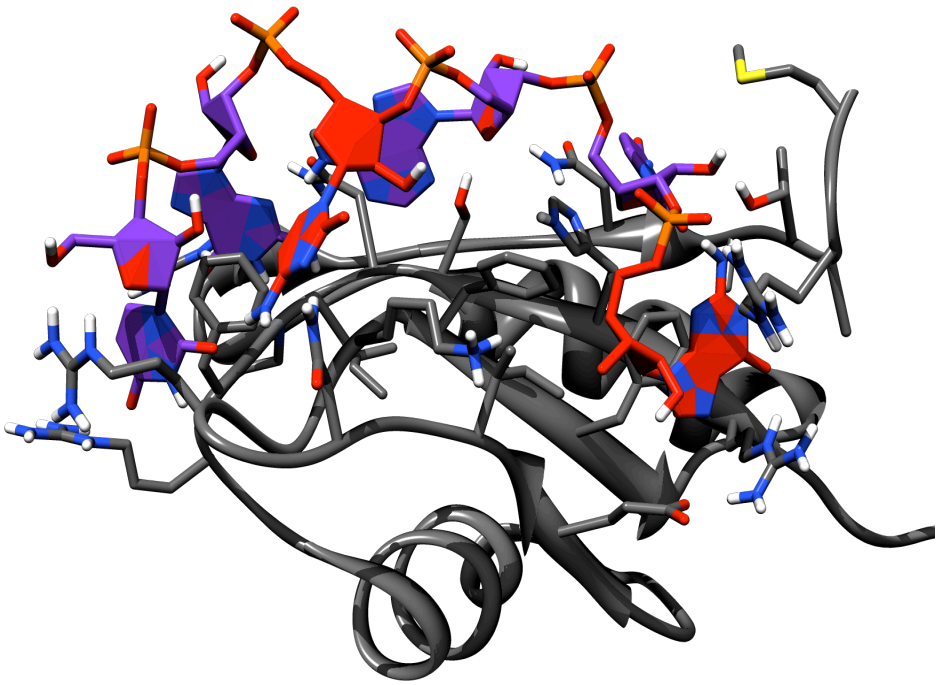


Figure 4.6: Schematic step-wise approach to retarget a RBP to recognize a desired nucleobase. The PUF domain example in the previous chapter (Chapter 3.H) illustrated that mutating one or two residue positions to a different amino acid may alter the base recognized by a RBP. The ease of the computational approach allows us to rapidly assess the effect of all residue point mutations on the specificity and affinity of the protein for a given nucleobase. For a residue position that might alter specificity, I explore the interaction of each possible amino acid with each of the four canonical bases at the target site. The tabulated score of the residue with respect to the base and its environment allows the evaluation of the effect of the residue substitution on the base specificity and on the protein affinity for the RNA as discussed by Ashworth and Baker (2009). The 'specificity view' shows the most favorable base given each possible amino acid at the mutation site. The 'affinity view' shows which residue interacts most favorably with the RNA. After selecting a residue that moves toward the desired base preference and that binds with high affinity, the best residue mutation is selected and search may be repeated for a different interacting residue position.



Fox1: 5' - U G C A U G - 3'

target: 5' - U G A A U C - 3'

Figure 4.7: Comparison of target sequence with Fox1 binding preference and structure.
The structure of Fox1 from PDB structure 2err is shown in complex with its preferred RNA sequence. The bases are colored by correspondence with the target sequence. Matches are shown in purple and mismatches are shown in red. The bases in the structure are colored to correspond with the aligned native and target sequences.

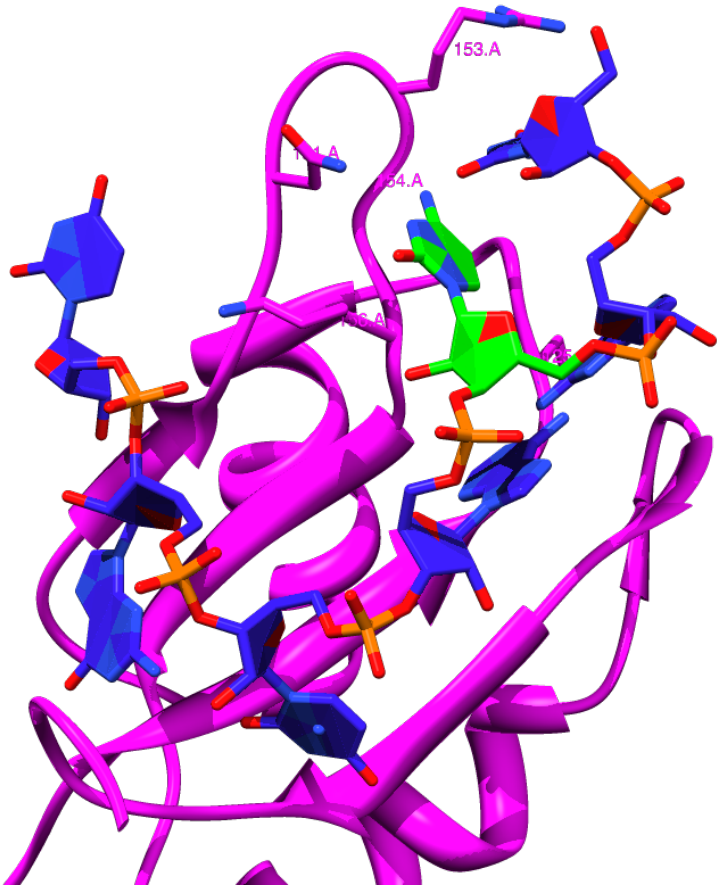


Figure 4.8: Structure of Fox1 design site B199 with contacting residues labeled.
Mutation of position B199 in Fox1 structure 2err would allow the RRM to bind the precursor miRNA loop sequence. The RRM in purple recognizes the five nucleobase single stranded region using mostly contacts with the two central antiparallel strands of the β-sheet. The nucleobase in green is a position we would like to retarget from a preference for C to a preference for A as illustrated in Figure 4.7. Side-chain at residue positions capable of participating in direct contacts with the altered nucleobase position are shown and labeled. I investigated the effect of all possible side-chain mutations at the illustrated residues: A126F, A151N, A153R, A154G and A156K.

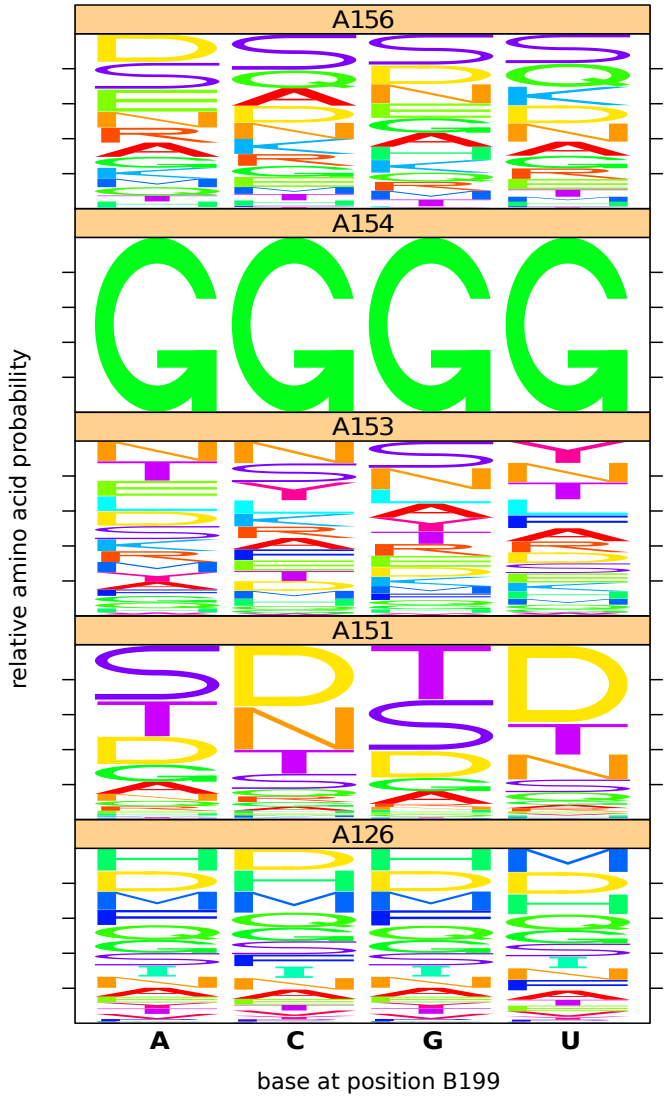


Figure 4.9: Residue preferences at positions contacting for Fox1 position B199 illustrated in logo format. Each panel looks at the relative scores of residue substitutions at a position interacting with nucleic acid position B199 in Fox1 structure 2err. Each panel represents residue substitutions at the indicated position A126, A151, A153, A154 or A156 where the other side-chains are repacked but unchanged from their native identity. For each of the canonical bases at position B199 (*x*-axis), the relative score of each possible residue was determined. Residue height and order was determined by the Boltzmann distribution of minimum scores. The preferred residue reflects that which binds the base with highest affinity. This view helps determine which residue in a position will bind the nucleobase on the *x*-axis most tightly. However, this view does not clearly illustrate specificity, the ability to discriminate between bases.

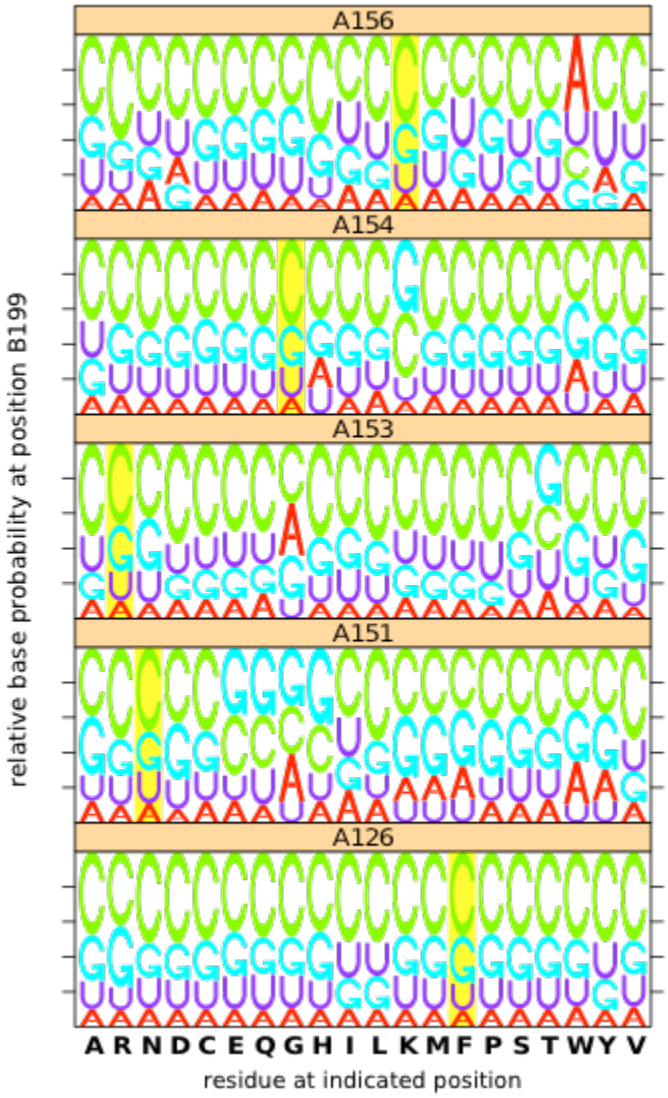


Figure 4.10: Change in base specificity with single mutations at five Fox1 residue positions.
Each panel looks at the effect of residue substitutions on nucleic acid preference at position B199 in Fox1 structure 2err. Each panel represents residue substitutions at the indicated position A126, A151, A153, A154 or A156 where the other side-chains are repacked but unchanged from their native identity. For each of the amino acid substitutions (*x*-axis) at the panel position, the relative preference of each possible base was determined. Base code height and order was determined by the Boltzmann distribution of minimum scores. The base probability reflects the effect of a residue substitution on base specificity. This view helps determine which base would be preferred given the residue substitution at the panel position to the *x*-axis residue. This view quickly shows which single residue mutation may alter base preference at the nucleotide position, though a given substitution may bind all bases with low affinity.By looking at the base preference with each possible single substitution of a contacting residue, I note that a substitution to tryptophan at A156 switches specificity from the native C to an A. Additionally glycine at A151 and A153 appear to increase the probability of binding adenosine.

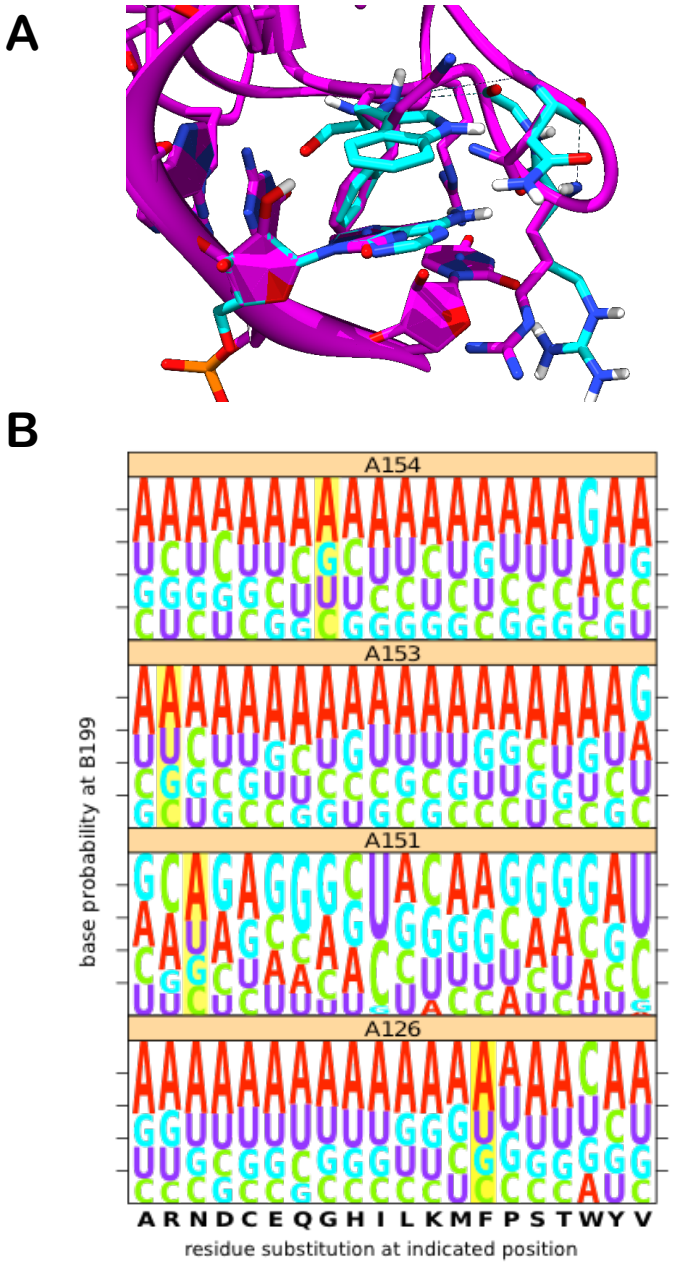


Figure 4.11: Change in base preference at B199 with a second mutation to a TRP mutated Fox1.
The specificity plot in Figure 4.10 suggested that the substitution at A156 from LYS to TRP (A156 K→W) would switch preference at position B199 from C to A. **A** Structure predicted to have lowest energy with A156W (cyan) showing the predicted stacking interaction between A156W and B199A is overlaid on the native structure (purple). A second substitution at one of the four remaining residue positions interacting with B199 could help improve the preference for binding an A. **B** I performed a second round of specificity tests with A156W and a second residue substitution at the position indicated in the panel title to the residue indicated on the *x*-axis. As in Figure 4.10, only one additional substitution is performed and the height of the one letter base code represents its relative probability of binding.

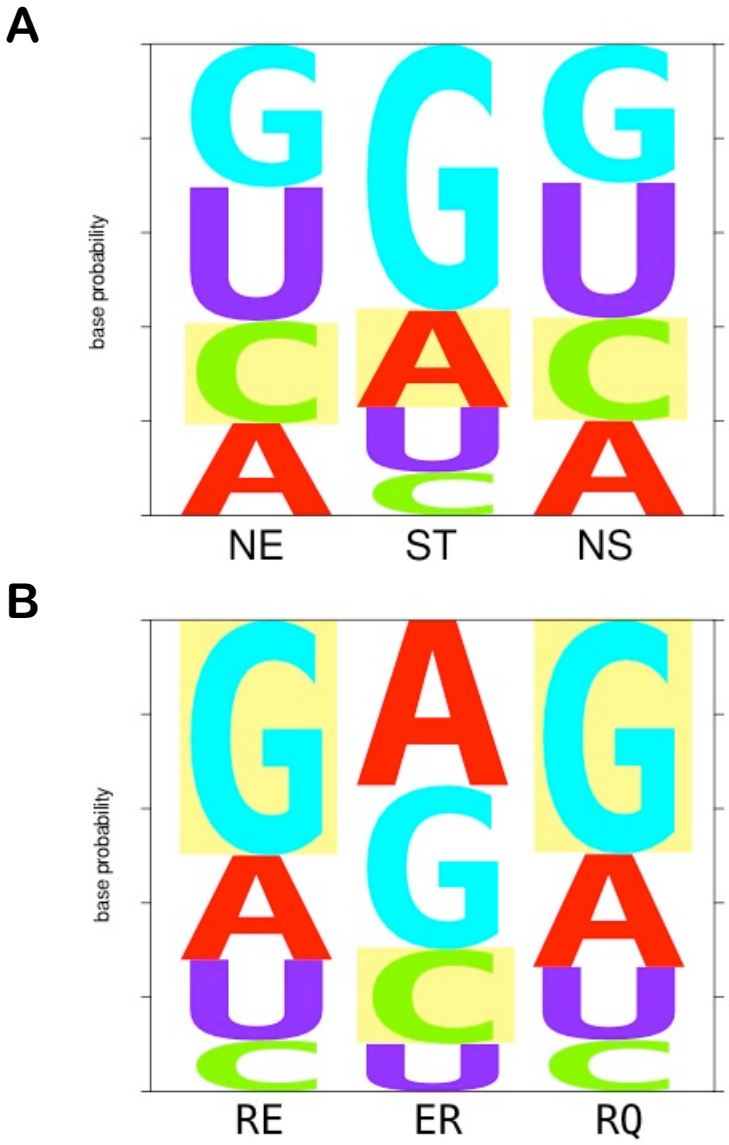


Figure 4.12: Specificity calculations with Fox1 RRM double mutants.
Four mutations (two per binding packet) allow the Fox1 RRM to specifically bind the target sequence instead of its native target. In terms of the changes to the native binding sequence (5'-UGCAUG-3'), specificity switches 3 C→A and 6 G→C are required. Protein positions are reported in terms of residue IDs in the PDB deposited NMR structure 2err. **A** Double mutations to A151 and A152 contacting RNA position 3 in terms of pairs of one-letter-codes on the *x*-axis. The ST mutant binds A in the target better than the Fox1 native NE binds C. **B** Double mutations to A118 and A147 contacting RNA position 6 in terms of pairs of one-letter-codes on the *x*-axis. The ER mutant binds the target C better than the Fox1 native RE binds G. The calculations show that the mutations improve the probability of binding the alternate target sequence.

Bibliography

Alibés, A., Nadra, A. D., De Masi, F., Bulyk, M. L., Serrano, L., & Stricher, F. (2010). Using protein design algorithms to understand the molecular basis of disease caused by protein–DNA interactions: the Pax6 example. *Nucleic Acids Research*, *38*(21), 7422 –7431. doi:10.1093/nar/gkq683

Alibés, A., Serrano, L., & Nadra, A. D. (2010). Structure-Based DNA-Binding Prediction and Design. In J. P. Mackay & D. J. Segal (Eds.), *Engineered Zinc Finger Proteins*, Methods in Molecular Biology (Vol. 649, pp. 77–88). Totowa, NJ: Humana Press. Retrieved from http://dx.doi.org/10.1007/978-1-60761-753-2_4

Allain, F. H.-T., Gubser, C. C., Howe, P. W. A., Nagai, K., Neuhaus, D., & Varani, G. (1996). Specificity of ribonucleoprotein interaction determined by RNA folding during complex formation. *Nature*, *380*(6575), 646–650. doi:10.1038/380646a0

Allen, F. H. (2002). The Cambridge Structural Database: a quarter of a million crystal structures and rising. *Acta Crystallographica Section B Structural Science*, *58*(3), 380–388. doi:10.1107/S0108768102003890

Andrabi, M., Mizuguchi, K., Sarai, A., & Ahmad, S. (2009). Prediction of mono- and di-nucleotide-specific DNA-binding sites in proteins using neural networks. *BMC Structural Biology*, *9*, 30. doi:10.1186/1472-6807-9-30

Andreeva, A., Howorth, D., Chandonia, J.-M., Brenner, S. E., Hubbard, T. J. P., Chothia, C., & Murzin, A. G. (2008). Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Research*, *36*(suppl 1), D419–D425. doi:10.1093/nar/gkm993

Andrusier, N., Mashiach, E., Nussinov, R., & Wolfson, H. J. (2008). Principles of flexible protein-protein docking. *Proteins*, *73*(2), 271–289. doi:10.1002/prot.22170

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science*, *181*(4096), 223–230. doi:10.1126/science.181.4096.223

Antosova, Z., Mackova, M., Kral, V., & Macek, T. (2009). Therapeutic application of peptides and proteins: parenteral forever? *Trends in Biotechnology*, *27*(11), 628–635. doi:10.1016/j.tibtech.2009.07.009

Anunciado, D., Dhar, A., Gruebele, M., & Baranger, A. M. (2011). Multistep Kinetics of the U1A–SL2 RNA Complex Dissociation. *Journal of Molecular Biology*, *408*(5), 896–908. doi:10.1016/j.jmb.2011.02.054

Ashworth, J., & Baker, D. (2009). Assessment of the optimization of affinity and specificity at protein–DNA interfaces. *Nucleic Acids Research*, *37*(10), e73. doi:10.1093/nar/gkp242

Ashworth, J., Havranek, J. J., Duarte, C. M., Sussman, D., Monnat, R. J. J., Stoddard, B. L., & Baker, D. (2006). Computational redesign of endonuclease DNA binding and cleavage specificity. *Nature*, *441*(7093), 656–659. doi:10.1038/nature04818

Ashworth, J., Taylor, G. K., Havranek, J. J., Quadri, S. A., Stoddard, B. L., & Baker, D. (2010). Computational reprogramming of homing endonuclease specificity at multiple adjacent base pairs. *Nucleic Acids Research*, *38*(16), 5601 –5608. doi:10.1093/nar/gkq283

Auweter, S. D., Oberstrass, F. C., & Allain, F. H.-T. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Research*, *34*(17), 4943–59. doi:10.1093/nar/gkl620

Avery, J. (2003). *Information theory and evolution*. River Edge N.J.: World Scientific. Retrieved from http://books.google.com/books?id=1agMKiYn2CkC

Badis, G., Berger, M. F., Philippakis, A. A., Talukder, S., Gehrke, A. R., Jaeger, S. A., Chan, E. T., et al. (2009). Diversity and complexity in DNA recognition by transcription factors. *Science*, *324*(5935), 1720–1723. doi:10.1126/science.1162327

Bahadur, R. P., Zacharias, M., & Janin, J. (2008). Dissecting protein-RNA recognition sites. *Nucleic Acids Research*, *36*(8), 2705–2716. doi:10.1093/nar/gkn102

Bembom, O. (2007). *seqLogo: Sequence logos for DNA sequence alignments*.

Berezikov, E. (2011). Evolution of microRNA diversity and regulation in animals. *Nature Reviews Genetics*, *12*(12), 846–860. doi:10.1038/nrg3079

Berger, M. F., & Bulyk, M. L. (2006). Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. In M. Bina (Ed.), *Gene Mapping, Discovery, and Expression*, Methods Molecular Biology (Vol. 338, pp. 245–260). Retrieved from http://dx.doi.org/10.1385/1-59745-097-9:245

Berger, M. F., & Bulyk, M. L. (2009). Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nature Protocols*, *4*(3), 393–411. doi:10.1038/nprot.2008.195

Berget, S. M., Moore, C., & Sharp, P. A. (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences*, *74*(8), 3171 –3175.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., et al. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235 –242. doi:10.1093/nar/28.1.235

Bernard, B., & Samudrala, R. (2009). A generalized knowledge-based discriminatory function for biomolecular interactions. *Proteins*, *76*(1), 115–128. doi:10.1002/prot.22323

Beuth, B., Garcia-Mayoral, M. F., Taylor, I. A., & Ramos, A. (2007). Scaffold-independent analysis of RNA-protein interactions: the Nova-1 KH3-RNA complex. *Journal of the American Chemical Society*, *129*(33), 10205–10210. doi:10.1021/ja072365q

Boas, F. E., & Harbury, P. B. (2007). Potential energy functions for protein design. *Current Opinion in Structural Biology*, *17*(2), 199–204. doi:10.1016/j.sbi.2007.03.006

Bonasio, R., Tu, S., & Reinberg, D. (2010). Molecular Signals of Epigenetic States. *Science*, *330*(6004), 612–616. doi:10.1126/science.1191078

Boutz, P. L., Stoilov, P., Li, Q., Lin, C.-H., Chawla, G., Ostrow, K., Shiue, L., et al. (2007). A post-transcriptional regulatory switch in polypyrimidine tract-binding proteins reprograms alternative splicing in developing neurons. *Genes & Development*, *21*(13), 1636 –1652. doi:10.1101/gad.1558107

Boyer, L. A., Lee, T. I., Cole, M. F., Johnstone, S. E., Levine, S. S., Zucker, J. P., Guenther, M. G., et al. (2005). Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell*, *122*(6), 947–956. doi:10.1016/j.cell.2005.08.020

Bradley, P., Malmström, L., Qian, B., Schonbrun, J., Chivian, D., Kim, D. E., Meiler, J., et al. (2005). Free modeling with Rosetta in CASP6. *Proteins*, *61 Suppl 7*, 128–134. doi:10.1002/prot.20729

Buchholz, F. (2009). Engineering DNA processing enzymes for the postgenomic era. *Current Opinion in Biotechnology*, *20*(4), 383–389. doi:10.1016/j.copbio.2009.07.005

Bushati, N., & Cohen, S. M. (2007). microRNA functions. *Annual Review of Cell and Developmental Biology*, *23*, 175–205. doi:10.1146/annurev.cellbio.23.090506.123406

Caetano-Anollés, D., Kim, K. M., Mittenthal, J. E., & Caetano-Anollés, G. (2010). Proteome Evolution and the Metabolic Origins of Translation and Cellular Life. *Journal of Molecular Evolution*, *72*(1), 14–33. doi:10.1007/s00239-010-9400-9

Carlson, C. D., Warren, C. L., Hauschild, K. E., Ozers, M. S., Qadir, N., Bhimsaria, D., Lee, Y., et al. (2010). Specificity landscapes of DNA binding molecules elucidate biological function. *Proceedings of the National Academy of Sciences*, *107*(10), 4544–4549. doi:10.1073/pnas.0914023107

Cartegni, L., Wang, J., Zhu, Z., Zhang, M. Q., & Krainer, A. R. (2003). ESEfinder: a web resource to identify exonic splicing enhancers. *Nucleic Acids Research*, *31*(13), 3568 –3571. doi:10.1093/nar/gkg616

Castle, J. C., Zhang, C., Shah, J. K., Kulkarni, A. V., Kalsotra, A., Cooper, T. A., & Johnson, J. M. (2008). Expression of 24,426 human alternative splicing events and predicted cis regulation in 48 tissues and cell lines. *Nature Genetics*, *40*, 1419–1425. doi:10.1038/ng.264

Castrignanò, T., Chillemi, G., Varani, G., & Desideri, A. (2002). Molecular dynamics simulation of the RNA complex of a double-stranded RNA-binding domain reveals dynamic features of the intermolecular interface and its hydration. *Biophysical Journal*, *83*(6), 3542–3552. doi:10.1016/S0006-3495(02)75354-X

Chen, B. E., Kondo, M., Garnier, A., Watson, F. L., Püettmann-Holgado, R., Lamar, D. R., & Schmucker, D. (2006). The Molecular Diversity of Dscam Is Functionally Required for Neuronal Wiring Specificity in Drosophila. *Cell*, *125*(3), 607–620. doi:10.1016/j.cell.2006.03.034

Chen, J. R., Chang, B. H., Allen, J. E., Stiffler, M. A., & MacBeath, G. (2008). Predicting PDZ domain-peptide interactions from primary sequences. *Nature Biotechnology*, *26*(9), 1041–1045. doi:10.1038/nbt.1489

Chen, X., Hughes, T. R., & Morris, Q. (2007). RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. *Bioinformatics*, *23*(13), i72–79. doi:10.1093/bioinformatics/btm224

Chen, Y., Kortemme, T., Robertson, T., Baker, D., & Varani, G. (2004). A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. *Nucleic Acids Research*, *32*(17), 5147–5162. doi:10.1093/nar/gkh785

Chen, Y., Mandic, J., & Varani, G. (2008). Cell-free selection of RNA-binding proteins using in vitro compartmentalization. *Nucleic Acids Research*, *36*(19), e128. doi:10.1093/nar/gkn559

Chen, Y., & Varani, G. (2005). Protein families and RNA recognition. *FEBS Journal*, *272*(9), 2088–2097. doi:10.1111/j.1742-4658.2005.04650.x

Chen, Y., & Varani, G. (2011). Finding the Missing Code of RNA Recognition by PUF Proteins. *Chemistry & Biology*, *18*. doi:10.1016/j.chembiol.2011.07.001

Cheong, C.-G., & Tanaka Hall, T. M. (2006). Engineering RNA sequence specificity of Pumilio repeats. *Proceedings of the National Academy of Sciences*, *103*(37), 13635–13639. doi:10.1073/pnas.0606294103

Chevalier, B. S., Kortemme, T., Chadsey, M. S., Baker, D., Monnat Jr., R. J., & Stoddard, B. L. (2002). Design, Activity, and Structure of a Highly Specific Artificial Endonuclease. *Molecular Cell*, *10*(4), 895–905. doi:10.1016/S1097-2765(02)00690-1

Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, *357*(6379), 543–544. doi:10.1038/357543a0

Chothia, C., & Finkelstein, A. V. (1990). The Classification and Origins of Protein Folding Patterns. *Annual Review of Biochemistry*, *59*(1), 1007–1035. doi:10.1146/annurev.bi.59.070190.005043

Chothia, C., & Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826.

Ciafrè, S. A., Galardi, S., Mangiola, A., Ferracin, M., Liu, C.-G., Sabatino, G., Negrini, M., et al. (2005). Extensive modulation of a set of microRNAs in primary glioblastoma. *Biochemical and Biophysical Research Communications*, *334*(4), 1351–1358. doi:10.1016/j.bbrc.2005.07.030

Cléry, A., Blatter, M., & Allain, F. H.-T. (2008). RNA recognition motifs: boring? Not quite. *Current Opinion in Structural Biology*, *18*(3), 290–298. doi:10.1016/j.sbi.2008.04.002

Cook, K. B., Kazan, H., Zuberi, K., Morris, Q., & Hughes, T. R. (2010). RBPDB: a database of RNA-binding specificities. *Nucleic Acids Research*, *39*(Database), D301–D308. doi:10.1093/nar/gkq1069

Cooper, W. J., & Waters, M. L. (2005). Molecular recognition with designed peptides and proteins. *Current Opinion in Chemical Biology*, *9*(6), 627–631. doi:10.1016/j.cbpa.2005.10.015

Costa, F. F. (2008). Non-coding RNAs, epigenetics and complexity. *Gene*, *410*(1), 9–17. doi:10.1016/j.gene.2007.12.008

Costa, F. F. (2009). Non-coding RNAs and new opportunities for the private sector. *Drug Discovery Today*, *14*(9-10), 446–452. doi:10.1016/j.drudis.2009.01.008

Crick, F. H. C. (1970). Central Dogma of Molecular Biology. *Nature*, *227*(5258), 561–563. doi:10.1038/227561a0

Croce, C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. *Nature Reviews Genetics*, *10*(10), 704–714. doi:10.1038/nrg2634

Crooks, G. E., Hon, G., Chandonia, J.-M., & Brenner, S. E. (2004). WebLogo: A Sequence Logo Generator. *Genome Research*, *14*(6), 1188–1190. doi:10.1101/gr.849004

Danner, S., & Belasco, J. G. (2001). T7 phage display: a novel genetic selection system for cloning RNA-binding proteins from cDNA libraries. *Proceedings of the National Academy of Sciences*, *98*(23), 12954–12959. doi:10.1073/pnas.211439598

Das, R., André, I., Shen, Y., Wu, Y., Lemak, A., Bansal, S., Arrowsmith, C. H., et al. (2009). Simultaneous prediction of protein folding and docking at high resolution. *Proceedings of the National Academy of Sciences*, *106*(45), 18978 –18983. doi:10.1073/pnas.0904407106

Das, R., & Baker, D. (2007). Automated de novo prediction of native-like RNA tertiary structures. *Proceedings of the National Academy of Sciences*, *104*(37), 14664–14669. doi:10.1073/pnas.0703836104

Das, R., & Baker, D. (2008). Macromolecular Modeling with Rosetta. *Annual Review of Biochemistry*, *77*(1), 363–382. doi:10.1146/annurev.biochem.77.062906.171838

Das, R., Karanicolas, J., & Baker, D. (2010). Atomic accuracy in predicting and designing noncanonical RNA structure. *Nature Methods*, *7*(4), 291–294. doi:10.1038/nmeth.1433

Davidson, A., Begley, D. W., Lau, C., & Varani, G. (2011). A small-molecule probe induces a conformation in HIV TAR RNA capable of binding drug-like fragments. *Journal of Molecular Biology*, *410*(5), 984–996. doi:10.1016/j.jmb.2011.03.039

Disney, M. D., & Guan, L. (2012). Recent Advances in Developing Small Molecules Targeting RNA. *ACS Chemical Biology*, *7*(1), 73–86. doi:10.1021/cb200447r

Djordjevic, M. (2007). SELEX experiments: New prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering*, *24*(2), 179–189. doi:10.1016/j.bioeng.2007.03.001

Dominguez, C., Fisette, J.-F., Chabot, B., & Allain, F. H.-T. (2010). Structural basis of G-tract recognition and encaging by hnRNP F quasi-RRMs. *Nature Structural & Molecular Biology*, *17*(7), 853–861. doi:10.1038/nsmb.1814

Dong, S., Wang, Y., Cassidy-Amstutz, C., Lu, G., Bigler, R., Jezyk, M. R., Li, C., et al. (2011). Specific and Modular Binding Code for Cytosine Recognition in Pumilio/FBF (PUF) RNA-binding Domains. *Journal of Biological Chemistry*, *286*(30), 26732 – 26742. doi:10.1074/jbc.M111.244889

Duarte, C. M., & Pyle, A. M. (1998). Stepping through an RNA structure: A novel approach to conformational analysis. *Journal of Molecular Biology*, *284*(5), 1465–1478. doi:10.1006/jmbi.1998.2233

Dunbrack, R. L. (2002). Rotamer libraries in the 21st century. *Current Opinion in Structural Biology*, *12*(4), 431–440. doi:10.1016/S0959-440X(02)00344-5

Dunbrack, R. L., & Karplus, M. (1993). Backbone-dependent rotamer library for proteins. Application to side-chain prediction. *Journal of Molecular Biology*, *230*(2), 543–574. doi:10.1006/jmbi.1993.1170

Ellis, J. J., Broom, M., & Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins*, *66*(4), 903–911. doi:10.1002/prot.21211

Elmen, J., Lindow, M., Schutz, S., Lawrence, M., Petri, A., Obad, S., Lindholm, M., et al. (2008). LNA-mediated microRNA silencing in non-human primates. *Nature*, *452*(7189), 896–899. doi:10.1038/nature06783

Esquela-Kerscher, A., & Slack, F. J. (2006). Oncomirs [mdash] microRNAs with a role in cancer. *Nature Reviews Cancer*, *6*(4), 259–269. doi:10.1038/nrc1840

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nature Reviews Genetics*, *8*(4), 286–298. doi:10.1038/nrg2005

Fabian, M. R., Sonenberg, N., & Filipowicz, W. (2010). Regulation of mRNA Translation and Stability by microRNAs. *Annual Review of Biochemistry*, *79*(1), 351–379. doi:10.1146/annurev-biochem-060308-103103

Farwer, J., Packer, M. J., & Hunter, C. A. (2006). Prediction of atomic structure from sequence for double helical DNA oligomers. *Biopolymers*, *81*(1), 51–61. doi:10.1002/bip.20377

Feliu, E., Aloy, P., & Oliva, B. (2011). On the analysis of protein–protein interactions via knowledge-based potentials for the prediction of protein–protein docking. *Protein Science*, *20*(3), 529–541. doi:10.1002/pro.585

Ferrada, E., & Melo, F. (2009). Effective knowledge-based potentials. *Protein Science*, *18*(7), 1469–1485. doi:10.1002/pro.166

Ferrada, E., Vergara, I. A., & Melo, F. (2007). A knowledge-based potential with an accurate description of local interactions improves discrimination between native and near-native protein conformations. *Cell Biochemistry and Biophysics*, *49*(2), 111–124.

Filipovska, A., Razif, M. F. M., Nygård, K. K. A., & Rackham, O. (2011). A universal code for RNA recognition by PUF proteins. *Nature Chemical Biology*, *7*(7), 425–427. doi:10.1038/nchembio.577

Filipowicz, W., Bhattacharyya, S. N., & Sonenberg, N. (2008). Mechanisms of post-transcriptional regulation by microRNAs: are the answers in sight? *Nature Reviews Genetics*, *9*(2), 102–114. doi:10.1038/nrg2290

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H.-R., Ceric, G., et al. (2008). The Pfam protein families database. *Nucleic Acids Research*, *36*(suppl_1), D281–288. doi:10.1093/nar/gkm960

Fisher, J., & Henzinger, T. A. (2007). Executable cell biology. *Nature Biotechnology*, *25*(11), 1239–1249. doi:10.1038/nbt1356

Flores, T. P., Orengo, C. A., Moss, D. S., & Thornton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science*, *2*(11), 1811–1826. doi:10.1002/pro.5560021104

Foley, J. E., Yeh, J.-R. J., Maeder, M. L., Reyon, D., Sander, J. D., Peterson, R. T., & Joung, J. K. (2009). Rapid Mutation of Endogenous Zebrafish Genes Using Zinc Finger Nucleases Made by Oligomerized Pool ENgineering (OPEN). *PLoS ONE*, *4*(2), e4348. doi:10.1371/journal.pone.0004348

Fulton, D. L., Denarier, E., Friedman, H. C., Wasserman, W. W., & Peterson, A. C. (2011). Towards resolving the transcription factor network controlling myelin gene expression. *Nucleic Acids Research*, *39*(18), 7974 –7991. doi:10.1093/nar/gkr326

Fuxreiter, M., Simon, I., & Bondos, S. (2011). Dynamic protein–DNA recognition: beyond what can be seen. *Trends in Biochemical Sciences*, *36*(8), 415–423. doi:10.1016/j.tibs.2011.04.006

Gaglione, M., Milano, G., Chambery, A., Moggio, L., Romanelli, A., & Messere, A. (2011). PNA-based artificial nucleases as antisense and anti-miRNA oligonucleotide agents. *Molecular BioSystems*, *7*(8), 2490–2499. doi:10.1039/C1MB05131H

Gao, M., & Skolnick, J. (2010). Structural space of protein–protein interfaces is degenerate, close to complete, and highly connected. *Proceedings of the National Academy of Sciences*, *107*(52), 22517–22522. doi:10.1073/pnas.1012820107

García-Mayoral, M. F., Díaz-Moreno, I., Hollingworth, D., & Ramos, A. (2008). The sequence selectivity of KSRP explains its flexibility in the recognition of the RNA targets. *Nucleic Acids Research*, *36*(16), 5290–5296. doi:10.1093/nar/gkn509

Garner, M. M., & Revzin, A. (1981). A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the Escherichia coli lactose operon regulatory system. *Nucleic Acids Research*, *9*(13), 3047–3060. doi:10.1093/nar/9.13.3047

Gentleman, R. C., Carey, V. J., Bates, D. M., Bolstad, B., Dettling, M., Dudoit, S., Ellis, B., et al. (2004). Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, *5*(10), R80. doi:10.1186/gb-2004-5-10-r80

Gharaibeh, R. Z., Newton, J. M., Weller, J. W., & Gibas, C. J. (2010). Application of Equilibrium Models of Solution Hybridization to Microarray Design and Analysis. *PLoS ONE*, *5*(6), e11048. doi:10.1371/journal.pone.0011048

Gilbert, W. (1978). Why genes in pieces? *Nature*, *271*(5645), 501. doi:10.1038/271501a0

Glisovic, T., Bachorik, J. L., Yong, J., & Dreyfuss, G. (2008). RNA-binding proteins and post-transcriptional gene regulation. *FEBS Letters*, *582*(14), 1977–1986. doi:10.1016/j.febslet.2008.03.004

Godin, K. S., & Varani, G. (2007). How Arginine-Rich Domains Coordinate mRNA Maturation Events. *RNA Biology*, *4*, 69–75. doi:10.4161/rna.4.2.4869

Goldberg, A. D., Allis, C. D., & Bernstein, E. (2007). Epigenetics: A Landscape Takes Shape. *Cell*, *128*(4), 635–638. doi:10.1016/j.cell.2007.02.006

Graveley, B. R. (2000). Sorting out the complexity of SR protein functions. *RNA*, *6*(9), 1197–211.

Griffiths-Jones, S., Saini, H. K., van Dongen, S., & Enright, A. J. (2007). miRBase: tools for microRNA genomics. *Nucleic Acids Research*, *36*(Database), D154–D158. doi:10.1093/nar/gkm952

Guerois, R., Nielsen, J. E., & Serrano, L. (2002). Predicting Changes in the Stability of Proteins and Protein Complexes: A Study of More Than 1000 Mutations. *Journal of Molecular Biology*, *320*(2), 369–387. doi:10.1016/S0022-2836(02)00442-4

Gupta, A., & Gribskov, M. (2011). The Role of RNA Sequence and Structure in RNA–Protein Interactions. *Journal of Molecular Biology*, *409*(4), 574–587. doi:10.1016/j.jmb.2011.04.007

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome Biology*, *8*(2), R24. doi:10.1186/gb-2007-8-2-r24

Hafner, M., Landthaler, M., Burger, L., Khorshid, M., Hausser, J., Berninger, P., Rothballer, A., et al. (2010). Transcriptome-wide Identification of RNA-Binding Protein and MicroRNA Target Sites by PAR-CLIP. *Cell*, *141*(1), 129–141. doi:10.1016/j.cell.2010.03.009

Hamelryck, T., Borg, M., Paluszewski, M., Paulsen, J., Frellsen, J., Andreetta, C., Boomsma, W., et al. (2010). Potentials of Mean Force for Protein Structure Prediction Vindicated, Formalized and Generalized. *PLoS ONE*, *5*(11), e13714. doi:10.1371/journal.pone.0013714

Harrison, R. S., Shepherd, N. E., Hoang, H. N., Ruiz-Gómez, G., Hill, T. A., Driver, R. W., Desai, V. S., et al. (2010). Downsizing human, bacterial, and viral proteins to short water-stable alpha helices that maintain biological potency. *Proceedings of the National Academy of Sciences*, *107*(26), 11686–11691. doi:10.1073/pnas.1002498107

Havranek, J. J., Duarte, C. M., & Baker, D. (2004). A simple physical model for the prediction and design of protein-DNA interactions. *Journal of Molecular Biology*, *344*(1), 59–70. doi:10.1016/j.jmb.2004.09.029

Hendlich, M., Lackner, P., Weitckus, S., Floeckner, H., Froschauer, R., Gottsbacher, K., Casari, G., et al. (1990). Identification of native protein folds amongst a large number of incorrect models. The calculation of low energy conformations from potentials of mean force. *Journal of Molecular Biology*, *216*(1), 167–180. doi:10.1016/S0022-2836(05)80068-3

Hieronymus, H., & Silver, P. A. (2003). Genome-wide analysis of RNA-protein interactions illustrates specificity of the mRNA export machinery. *Nature Genetics*, *33*(2), 155–161. doi:10.1038/ng1080

Hinman, M., & Lou, H. (2008). Diverse molecular functions of Hu proteins. *Cellular and Molecular Life Sciences (CMLS)*. doi:10.1007/s00018-008-8252-6

Hoffman, M. M., Khrapov, M. A., Cox, J. C., Yao, J., Tong, L., & Ellington, A. D. (2004). AANT: the Amino Acid-Nucleotide Interaction Database. *Nucleic Acids Research*, *32*(Database issue), D174–181. doi:10.1093/nar/gkh128

Hogan, D. J., Riordan, D. P., Gerber, A. P., Herschlag, D., & Brown, P. O. (2008). Diverse RNA-Binding Proteins Interact with Functionally Related Sets of RNAs, Suggesting an Extensive Regulatory System. *PLoS Biology*, *6*(10), e255. doi:10.1371/journal.pbio.0060255

Huang, E. S., Samudrala, R., & Park, B. H. (2000). Scoring functions for ab initio protein structure prediction. *Methods in Molecular Biology*, *143*, 223–245. doi:10.1385/1-59259-368-2:223

Huang, S.-Y., & Zou, X. (2008). An iterative knowledge-based scoring function for protein-protein recognition. *Proteins*, *72*(2), 557–579. doi:10.1002/prot.21949

Iliopoulos, D., Jaeger, S. A., Hirsch, H. A., Bulyk, M. L., & Struhl, K. (2010). STAT3 Activation of miR-21 and miR-181b-1 via PTEN and CYLD Are Part of the Epigenetic Switch Linking Inflammation to Cancer. *Molecular Cell*, *39*(4), 493–506. doi:10.1016/j.molcel.2010.07.023

International Human Genome Sequencing Consortium. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931–945. doi:10.1038/nature03001

Jacob, F., & Monod, J. (1961). Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, *3*(3), 318–356. doi:10.1016/S0022-2836(61)80072-7

Janga, S. C., & Mittal, N. (2011). Construction, Structure and Dynamics of Post-Transcriptional Regulatory Network Directed by RNA-Binding Proteins. In L. J. Collins (Ed.), *RNA Infrastructure and Networks*, Advances in Experimental Medicine and Biology (Vol. 722, pp. 103–117). New York, NY: Springer New York. Retrieved from http://dx.doi.org/10.1007/978-1-4614-0332-6_7

Jaynes, E. T. (1957). Information Theory and Statistical Mechanics. *Physical Review*, *106*(4), 620–630. doi:10.1103/PhysRev.106.620

Jaynes, E. T. (2003). *Probability theory : the logic of science*. New York: Cambridge University Press. Retrieved from http://books.google.com/books?id=tTN4HuUNXjgC

Jones, S., Daley, D. T., Luscombe, N. M., Berman, H. M., & Thornton, J. M. (2001). Protein-RNA interactions: a structural analysis. *Nucleic Acids Research*, *29*(4), 943–954. doi:10.1093/nar/29.4.943

Karanicolas, J., Corn, J. E., Chen, I., Joachimiak, L. A., Dym, O., Peck, S. H., Albeck, S., et al. (2011). A De Novo Protein Binding Pair By Computational Design and Directed Evolution. *Molecular Cell*, *42*(2), 250–260. doi:10.1016/j.molcel.2011.03.010

Keene, J. D. (2001). Ribonucleoprotein infrastructure regulating the flow of genetic information between the genome and the proteome. *Proceedings of the National Academy of Sciences*, *98*(13), 7018–7024. doi:10.1073/pnas.111145598

Keene, J. D. (2007). RNA regulons: coordination of post-transcriptional events. *Nature Reviews Genetics*, *8*(7), 533–543. doi:10.1038/nrg2111

Keene, J. D., & Tenenbaum, S. A. (2002). Eukaryotic mRNPs May Represent Posttranscriptional Operons. *Molecular Cell*, *9*(6), 1161–1167. doi:10.1016/S1097-2765(02)00559-2

Kerner, P., Degnan, S. M., Marchand, L., Degnan, B. M., & Vervoort, M. (2011). Evolution of RNA-Binding Proteins in Animals: Insights from Genome-Wide Analysis in the Sponge Amphimedon queenslandica. *Molecular Biology and Evolution*, *28*(8), 2289–2303. doi:10.1093/molbev/msr046

Kim, Junhyong, & Eberwine, J. (2010). RNA: state memory and mediator of cellular phenotype. *Trends in Cell Biology*, *20*(6), 311–318. doi:10.1016/j.tcb.2010.03.003

Kim, Jonghwan, Chu, J., Shen, X., Wang, J., & Orkin, S. H. (2008). An Extended Transcriptional Network for Pluripotency of Embryonic Stem Cells. *Cell*, *132*(6), 1049–1061. doi:10.1016/j.cell.2008.02.039

King, C. A., & Bradley, P. (2010). Structure-based prediction of protein-peptide specificity in rosetta. *Proteins: Structure, Function, and Bioinformatics*, 3437–3449. doi:10.1002/prot.22851

Kishore, S., Jaskiewicz, L., Burger, L., Hausser, J., Khorshid, M., & Zavolan, M. (2011). A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins. *Nature Methods*, *8*. doi:10.1038/nmeth.1608

Kishore, S., Luber, S., & Zavolan, M. (2010). Deciphering the role of RNA-binding proteins in the post-transcriptional control of gene expression. *Briefings in Functional Genomics*, *9*(5-6), 391–404. doi:10.1093/bfgp/elq028

Kortemme, T., Morozov, A. V., & Baker, D. (2003). An Orientation-dependent Hydrogen Bonding Potential Improves Prediction of Specificity and Structure for Proteins and Protein-Protein Complexes. *Journal of Molecular Biology*, *326*(4), 1239–1259. doi:10.1016/S0022-2836(03)00021-4

Krzanowski, W. J. (2009). *ROC Curves for Continuous Data*. Boca Raton: CRC Press. Retrieved from http://books.google.com/books?id=UZHwdiwOs4QC

Laing, C., & Schlick, T. (2011). Computational approaches to RNA structure prediction, analysis, and design. *Current Opinion in Structural Biology*, *21*(3), 306–318. doi:10.1016/j.sbi.2011.03.015

Lalonde, M. S., Lobritz, M. A., Ratcliff, A., Chamanian, M., Athanassiou, Z., Tyagi, M., Wong, J., et al. (2011). Inhibition of Both HIV-1 Reverse Transcription and Gene Expression by a Cyclic Peptide that Binds the Tat-Transactivating Response Element (TAR) RNA. *PLoS Pathogens*, *7*(5), e1002038. doi:10.1371/journal.ppat.1002038

Lane, D., Prentki, P., & Chandler, M. (1992). Use of gel retardation to analyze protein-nucleic acid interactions. *Microbiological Reviews*, *56*(4), 509–528.

Leaver-Fay, A., Tyka, M., Lewis, S. M., Lange, O. F., Thompson, J., Jacak, R., Kaufman, K., et al. (2011). ROSETTA3: an object-oriented software suite for the simulation and design of macromolecules. *Methods in Enzymology*, *487*, 545–574. doi:10.1016/B978-0-12-381270-4.00019-6

Lejeune, D., Delsaux, N., Charloteaux, B., Thomas, A., & Brasseur, R. (2005). Protein–nucleic acid recognition: Statistical analysis of atomic interactions and influence of DNA structure. *Proteins: Structure, Function, and Bioinformatics*, *61*(2), 258–271. doi:10.1002/prot.20607

Lennard-Jones, J. E., & Devonshire, A. F. (1937). Critical Phenomena in Gases. I. *Proceedings of the Royal Society of London. Series A - Mathematical and Physical Sciences*, *163*(912), 53 –70. doi:10.1098/rspa.1937.0210

Lensink, M. F., & Wodak, S. J. (2010a). Docking and scoring protein interactions: CAPRI 2009. *Proteins*, *78*(15), 3073–3084. doi:10.1002/prot.22818

Lensink, M. F., & Wodak, S. J. (2010b). Blind predictions of protein interfaces by docking calculations in CAPRI. *Proteins: Structure, Function, and Bioinformatics*, *78*(15), 3085–3095. doi:10.1002/prot.22850

Leontis, N. B., & Westhof, E. (2001). Geometric nomenclature and classification of RNA base pairs. *RNA*, *7*(4), 499–512.

Li, Q., Lee, J.-A., & Black, D. L. (2007). Neuronal regulation of alternative pre-mRNA splicing. *Nature Reviews Neuroscience*, *8*(11), 819–831. doi:10.1038/nrn2237

Li, Xiaofan, Moal, I. H., & Bates, P. A. (2010). Detection and refinement of encounter complexes for protein–protein docking: Taking account of macromolecular crowding. *Proteins: Structure, Function, and Bioinformatics*, *78*(15), 3189–3196. doi:10.1002/prot.22770

Li, Xiao, Quon, G., Lipshitz, H. D., & Morris, Q. (2010). Predicting in vivo binding sites of RNA-binding proteins using mRNA secondary structure. *RNA*, *16*(6), 1096–1107. doi:10.1261/rna.2017210

Licatalosi, D. D., Mele, A., Fak, J. J., Ule, J., Kayikci, M., Chi, S. W., Clark, T. A., et al. (2008). HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, *456*(7221), 464–469. doi:10.1038/nature07488

Liu, J., Montelione, G. T., & Rost, B. (2007). Novel leverage of structural genomics. *Nature Biotechnology*, *25*(8), 849–851. doi:10.1038/nbt0807-849

Liu-Yesucevitz, L., Bassell, G. J., Gitler, A. D., Hart, A. C., Klann, E., Richter, J. D., Warren, S. T., et al. (2011). Local RNA Translation at the Synapse and in Disease. *The Journal of Neuroscience*, *31*(45), 16086 –16093. doi:10.1523/JNEUROSCI.4105-11.2011

Loewenstein, W. (1999). *The touchstone of life : molecular information, cell communication, and the foundations of life*. New York: Oxford University Press. Retrieved from http://books.google.com/books?id=g82cx_99NbQC

Lu, H., Lu, L., & Skolnick, J. (2003). Development of Unified Statistical Potentials Describing Protein-Protein Interactions. *Biophysical Journal*, *84*(3), 1895–1901. doi:10.1016/S0006-3495(03)74997-2

Lu, H., & Skolnick, J. (2001). A distance-dependent atomic knowledge-based potential for improved protein structure selection. *Proteins*, *44*(3), 223–232. doi:10.1002/prot.1087

Lu, M., Dousis, A. D., & Ma, J. (2008a). OPUS-PSP: An Orientation-dependent Statistical All-atom Potential Derived from Side-chain Packing. *Journal of Molecular Biology*, *376*(1), 288–301. doi:10.1016/j.jmb.2007.11.033

Lu, M., Dousis, A. D., & Ma, J. (2008b). OPUS-Rota: a fast and accurate method for side-chain modeling. *Protein Science*, *17*(9), 1576–1585. doi:10.1110/ps.035022.108

Lukong, K. E., Chang, K., Khandjian, E. W., & Richard, S. (2008). RNA-binding proteins in human genetic disease. *Trends in Genetics*, *24*(8), 416–425. doi:10.1016/j.tig.2008.05.004

MacKerell, A. D., & Nilsson, L. (2008). Molecular dynamics simulations of nucleic acid-protein complexes. *Current Opinion in Structural Biology*, *18*(2), 194–199. doi:10.1016/j.sbi.2007.12.012

Mack, G. S. (2007). MicroRNA gets down to business. *Nature Biotechnology*, *25*(6), 631–638. doi:10.1038/nbt0607-631

Mackay, J. P., Font, J., & Segal, D. J. (2011). The prospects for designer single-stranded RNA-binding proteins. *Nature Structural & Molecular Biology*, *18*(3), 256–261. doi:10.1038/nsmb.2005

Maeda, T., Imanishi, Y., & Palczewski, K. (2003). Rhodopsin phosphorylation: 30 years later. *Progress in Retinal and Eye Research*, *22*(4), 417–434. doi:10.1016/S1350-9462(03)00017-X

Maeder, M. L., Thibodeau-Beganny, S., Osiak, A., Wright, D. A., Anthony, R. M., Eichtinger, M., Jiang, T., et al. (2008). Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular Cell*, *31*(2), 294–301. doi:10.1016/j.molcel.2008.06.016

Maeder, M. L., Thibodeau-Beganny, S., Sander, J. D., Voytas, D. F., & Joung, J. K. (2009). Oligomerized pool engineering (OPEN): an "open-source" protocol for making customized zinc-finger arrays. *Nature Protocols*, *4*(10), 1471–1501. doi:10.1038/nprot.2009.98

Mansfield, K. D., & Keene, J. D. (2009). The ribonome: a dominant force in co-ordinating gene expression. *Biology of the Cell*, *101*(3), 169. doi:10.1042/BC20080055

Mason, J. M. (2010). Design and development of peptides and peptide mimetics as antagonists for therapeutic intervention. *Future Medicinal Chemistry*, *2*(12), 1813–1822. doi:10.4155/fmc.10.259

McDonald, I. K., & Thornton, J. M. (1994). Satisfying hydrogen bonding potential in proteins. *Journal of Molecular Biology*, *238*(5), 777–793. doi:10.1006/jmbi.1994.1334

Mehler, M. F., & Mattick, J. S. (2007). Noncoding RNAs and RNA Editing in Brain Development, Functional Diversification, and Neurological Disease. *Physiological Reviews*, *87*(3), 799–823. doi:10.1152/physrev.00036.2006

Melo, S. A., & Esteller, M. (2011). Dysregulation of microRNAs in cancer: Playing with fire. *FEBS Letters*, *585*(13), 2087–2099. doi:10.1016/j.febslet.2010.08.009

Messias, A. C., & Sattler, M. (2004). Structural basis of single-stranded RNA recognition. *Accounts of Chemical Research*, *37*(5), 279–287. doi:10.1021/ar030034m

Mikheikin, A. L., Surzhikov, S. A., Zasedateleva, O. A., Vasiliskov, V. A., Pan'kov, S., Grechishnikova, I. V., Kisselev, L. L., et al. (2008). An RNA microchip containing immobilized oligoribonucleotides with protective groups at 2'-O-positions. *BioTechniques*, *44*, 77–83. doi:10.2144/000112677

Miller, J. C., Holmes, M. C., Wang, J., Guschin, D. Y., Lee, Y.-L., Rupniewski, I., Beausejour, C. M., et al. (2007). An improved zinc-finger nuclease architecture for highly specific genome editing. *Nature Biotechnology*, *25*(7), 778–785. doi:10.1038/nbt1319

Mittal, N., Roy, N., Babu, M. M., & Janga, S. C. (2009). Dissecting the expression dynamics of RNA-binding proteins in posttranscriptional regulatory networks. *Proceedings of the National Academy of Sciences*, *106*(48), 20300 –20305. doi:10.1073/pnas.0906940106

Mittal, N., Scherrer, T., Gerber, A. P., & Janga, S. C. (2011). Interplay between Posttranscriptional and Posttranslational Interactions of RNA-Binding Proteins. *Journal of Molecular Biology*, *409*(3), 466–479. doi:10.1016/j.jmb.2011.03.064

Moore, M. J., Schwartzfarb, E. M., Silver, P. A., & Yu, M. C. (2006). Differential Recruitment of the Splicing Machinery during Transcription Predicts Genome-Wide Patterns of mRNA Splicing. *Molecular Cell*, *24*(6), 903–915. doi:10.1016/j.molcel.2006.12.006

Morozov, A. V., Havranek, J. J., Baker, D., & Siggia, E. D. (2005). Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Research*, *33*(18), 5781–5798. doi:10.1093/nar/gki875

Morozova, N., Allers, J., Myers, J. C., & Shamoo, Y. (2006). Protein-RNA interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. *Bioinformatics*, *22*(22), 2746–2752. doi:10.1093/bioinformatics/btl470

Morris, Q., Bulyk, M. L., & Hughes, T. R. (2011). Jury remains out on simple models of transcription factor specificity. *Nature Biotechnology*, *29*(6), 483–484. doi:10.1038/nbt.1892

Moult, J., Fidelis, K., Kryshtafovych, A., & Tramontano, A. (2011). Critical assessment of methods of protein structure prediction (CASP)—round IX. *Proteins: Structure, Function, and Bioinformatics*, *79*(S10), 1–5. doi:10.1002/prot.23200

Murray, L. J. W., Arendall, W. B. 3rd, Richardson, D. C., & Richardson, J. S. (2003). RNA backbone is rotameric. *Proceedings of the National Academy of Sciences*, *100*(24), 13904–13909. doi:10.1073/pnas.1835769100

Myers, J. C., & Shamoo, Y. (2004). Human UP1 as a model for understanding purine recognition in the family of proteins containing the RNA recognition motif (RRM). *Journal of Molecular Biology*, *342*(3), 743–756. doi:10.1016/j.jmb.2004.07.029

Nadra, A. D., Serrano, L., & Alibés, A. (2011). DNA-binding specificity prediction with FoldX. *Methods in Enzymology*, *498*, 3–18. doi:10.1016/B978-0-12-385120-8.00001-2

Newburger, D. E., & Bulyk, M. L. (2009). UniPROBE: an online database of protein binding microarray data on protein-DNA interactions. *Nucleic Acids Research*, *37*(Database issue), D77–82. doi:10.1093/nar/gkn660

Newman, M. E. J., & Barkema, G. T. (1999). *Monte Carlo methods in statistical physics*. New York: Oxford University Press. Retrieved from http://books.google.com/books?id=J5aLdDN4uFwC

Nilsen, T. W., & Graveley, B. R. (2010). Expansion of the eukaryotic proteome by alternative splicing. *Nature*, *463*(7280), 457–463. doi:10.1038/nature08909

Nowacki, M., Shetty, K., & Landweber, L. F. (2011). RNA-Mediated Epigenetic Programming of Genome Rearrangements. *Annual Review of Genomics and Human Genetics*, *12*(1), 367–389. doi:10.1146/annurev-genom-082410-101420

Osbourn, A. E., & Field, B. (2009). Operons. *Cellular and Molecular Life Sciences*, *66*(23), 3755–3775. doi:10.1007/s00018-009-0114-3

Pabo, C. O., & Nekludova, L. (2000). Geometric analysis and comparison of protein-DNA interfaces: why is there no simple code for recognition? *Journal of Molecular Biology*, *301*(3), 597–624. doi:10.1006/jmbi.2000.3918

Pabo, C. O., & Sauer, R. T. (1984). Protein-DNA Recognition. *Annual Review of Biochemistry*, *53*, 293–321. doi:10.1146/annurev.bi.53.070184.001453

Pabo, C. O., & Sauer, R. T. (1992). Transcription Factors: Structural Families and Principles of DNA Recognition. *Annual Review of Biochemistry*, *61*, 1053–1095. doi:10.1146/annurev.bi.61.070192.005201

Paillard, G., & Lavery, R. (2004). Analyzing protein-DNA recognition mechanisms. *Structure*, *12*(1), 113–122. doi:10.1016/j.str.2003.11.022

Pantazes, R. J., Grisewood, M. J., & Maranas, C. D. (2011). Recent advances in computational protein design. *Current Opinion in Structural Biology*, *21*(4), 467–472. doi:10.1016/j.sbi.2011.04.005

Perez, E. E., Wang, J., Miller, J. C., Jouvenot, Y., Kim, K. A., Liu, O., Wang, N., et al. (2008). Establishment of HIV-1 resistance in CD4+ T cells by genome editing using zinc-finger nucleases. *Nature Biotechnology*, *26*(7), 808–816. doi:10.1038/nbt1410

Perutz, M. F. (1983). Species adaptation in a protein molecule. *Molecular Biology and Evolution*, *1*(1), 1 –28.

Philippakis, A. A., Qureshi, A. M., Berger, M. F., & Bulyk, M. L. (2008). Design of compact, universal DNA microarrays for protein binding microarray experiments. *Journal of Computational Biology*, *15*(7), 655–665. doi:10.1089/cmb.2007.0114

Piana, S., Lindorff-Larsen, K., & Shaw, D. E. (2011). How Robust Are Protein Folding Simulations with Respect to Force Field Parameterization? *Biophysical Journal*, *100*, L47–L49. doi:10.1016/j.bpj.2011.03.051

Piva, F., Giulietti, M., Nocchi, L., & Principato, G. (2009). SpliceAid: a database of experimental RNA target motifs bound by splicing proteins in humans. *Bioinformatics*, *25*(9), 1211–1213. doi:10.1093/bioinformatics/btp124

Pons, C., Solernou, A., Perez-Cano, L., Grosdidier, S., & Fernandez-Recio, J. (2010). Optimization of pyDock for the new CAPRI challenges: Docking of homology-based models, domain–domain assembly and protein-RNA binding. *Proteins: Structure, Function, and Bioinformatics*, *78*(15), 3182–3188. doi:10.1002/prot.22773

Pooga, M., Soomets, U., Hallbrink, M., Valkna, A., Saar, K., Rezaei, K., Kahl, U., et al. (1998). Cell penetrating PNA constructs regulate galanin receptor levels and modify pain transmission in vivo. *Nature Biotechnology*, *16*(9), 857–861. doi:10.1038/nbt0998-857

Portales-Casamar, E., Thongjuea, S., Kwon, A. T., Arenillas, D., Zhao, X., Valen, E., Yusuf, D., et al. (2010). JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *38*(Database issue), D105–110. doi:10.1093/nar/gkp950

Prasanth, K. V., & Spector, D. L. (2007). Eukaryotic regulatory RNAs: an answer to the "genome complexity" conundrum. *Genes & Development*, *21*(1), 11 –42. doi:10.1101/gad.1484207

Qin, F., Chen, Y., Wu, M., Li, Y., Zhang, J., & Chen, H.-F. (2010). Induced fit or conformational selection for RNA/U1A folding. *RNA*, *16*(5), 1053 –1061. doi:10.1261/rna.2008110

R Development Core Team. (2011). *R: A Language and Environment for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org/

Raman, S., Vernon, R., Thompson, J., Tyka, M., Sadreyev, R., Pei, J., Kim, D., et al. (2009). Structure prediction for CASP8 with all-atom refinement using Rosetta. *Proteins: Structure, Function, and Bioinformatics*, *77*(S9), 89–99. doi:10.1002/prot.22540

Ray, D., Kazan, H., Chan, E. T., Castillo, L. P., Chaudhry, S., Talukder, S., Blencowe, B. J., et al. (2009). Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins. *Nature Biotechnology*. doi:10.1038/nbt.1550

Read, R. J., & Chavali, G. (2007). Assessment of CASP7 predictions in the high accuracy template-based modeling category. *Proteins: Structure, Function, and Bioinformatics*, *69*(S8), 27–37. doi:10.1002/prot.21662

Redfern, O. C., Dessailly, B., & Orengo, C. A. (2008). Exploring the structure and function paradigm. *Current Opinion in Structural Biology*, *18*(3), 394–402. doi:10.1016/j.sbi.2008.05.007

Reichel, M., Li, J., & Millar, A. A. (2011). Silencing the silencer: strategies to inhibit microRNA activity. *Biotechnology Letters*, *33*(7), 1285–1292. doi:10.1007/s10529-011-0590-z

Ritchie, D. W. (2008). Recent Progress and Future Directions in Protein-Protein Docking. *Current Protein & Peptide Science*, *9*(1), 1–15. doi:10.2174/138920308783565741

Robertson, T. (2007). *Development and validation of statistical potential functions for the prediction of protein nucleic-acid interactions from structure* (Ph.D.). University of Washington, Seattle, Washington. Retrieved from http://hdl.handle.net/1773/9268

Robertson, T., & Varani, G. (2007). An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. *Proteins*, *66*(2), 359–374. doi:10.1002/prot.21162

Rohl, C. A., Strauss, C. E. M., Misura, K. M. S., & Baker, D. (2004). Protein Structure Prediction Using Rosetta. *Numerical Computer Methods, Part D*, Methods in Enzymology (Vol. 383, pp. 66–93). Academic Press. Retrieved from http://dx.doi.org/10.1016/S0076-6879(04)83004-0

Rudel, D., & Sommer, R. J. (2003). The evolution of developmental mechanisms. *Developmental Biology*, *264*(1), 15–37. doi:10.1016/S0012-1606(03)00353-1

Ruth, D. (2011). Polymer therapeutics as nanomedicines: new perspectives. *Current Opinion in Biotechnology*, *22*(4), 492–501. doi:10.1016/j.copbio.2011.05.507

Rykunov, D., & Fiser, A. (2010). New statistical potential for quality assessment of protein models and a survey of energy functions. *BMC Bioinformatics*, *11*. doi:10.1186/1471-2105-11-128

Samudrala, R., & Moult, J. (1998). An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. *Journal of Molecular Biology*, *275*(5), 895–916. doi:10.1006/jmbi.1997.1479

Sander, J. D., Zaback, P., Joung, J. K., Voytas, D. F., & Dobbs, D. (2009). An affinity-based scoring scheme for predicting DNA-binding activities of modularly assembled zinc-finger proteins. *Nucleic Acids Research*, *37*(2), 506 –515. doi:10.1093/nar/gkn962

Sarkar, D. (2008). *Lattice multivariate data visualization with R*. Use R! New York: Springer Science. Retrieved from http://dx.doi.org/10.1007/978-0-387-75969-2

Sato, A. K., Viswanathan, M., Kent, R. B., & Wood, C. R. (2006). Therapeutic peptides: technological advances driving peptides into development. *Current Opinion in Biotechnology*, *17*(6), 638–642. doi:10.1016/j.copbio.2006.10.002

Schaeffer, R. D., Fersht, A., & Daggett, V. (2008). Combining experiment and simulation in protein folding: closing the gap for small model systems. *Current Opinion in Structural Biology*, *18*(1), 4–9. doi:10.1016/j.sbi.2007.11.007

Scharf, A. N. D., & Imhof, A. (2011). Every methyl counts – Epigenetic calculus. *FEBS Letters*, *585*(13), 2001–2007. doi:10.1016/j.febslet.2010.11.029

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097–6100. doi:10.1093/nar/18.20.6097

Schneider, T. D., Stormo, G. D., Haemer, J. S., & Gold, L. (1982). A design for computer nucleic-acid-sequence storage, retrieval, and manipulation. *Nucleic Acids Research*, *10*(9), 3013–3024. doi:10.1093/nar/10.9.3013

Schroers, R., Hildebrandt, Y., Hasenkamp, J., Glass, B., Lieber, A., Wulf, G., & Piesche, M. (2004). Gene transfer into human T lymphocytes and natural killer cells by

Ad5/F35 chimeric adenoviral vectors. *Experimental Hematology*, *32*(6), 536–546. doi:10.1016/j.exphem.2004.03.010

Schuster-Bockler, B., Schultz, J., & Rahmann, S. (2004). HMM Logos for visualization of protein families. *BMC Bioinformatics*, *5*(1), 7. doi:10.1186/1471-2105-5-7

Schwanhausser, B., Busse, D., Li, N., Dittmar, G., Schuchhardt, J., Wolf, J., Chen, W., et al. (2011). Global quantification of mammalian gene expression control. *Nature*, *473*(7347), 337–342. doi:10.1038/nature10098

Shannon, C. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*.

Sharp, P. A. (2009). The centrality of RNA. *Cell*, *136*(4), 577–580. doi:10.1016/j.cell.2009.02.007

Shaw, D. E., Maragakis, P., Lindorff-Larsen, K., Piana, S., Dror, R. O., Eastwood, M. P., Bank, J. A., et al. (2010). Atomic-Level Characterization of the Structural Dynamics of Proteins. *Science*, *330*(6002), 341 –346. doi:10.1126/science.1187409

Shenoy, S. R., & Jayaram, B. (2010). Proteins: sequence to structure and function--current status. *Current Protein & Peptide Science*, *11*(7), 498–514.

Sippl, M. J. (1990). Calculation of conformational ensembles from potentials of mean force. An approach to the knowledge-based prediction of local structures in globular proteins. *Journal of Molecular Biology*, *213*(4), 859–883. doi:10.1016/S0022-2836(05)80269-4

Sippl, M. J. (1995). Knowledge-based potentials for proteins. *Current Opinion in Structural Biology*, *5*(2), 229–235. doi:10.1016/0959-440X(95)80081-6

Sippl, M. J., Ortner, M., Jaritz, M., Lackner, P., & Flockner, H. (1996). Helmholtz free energies of atom pair interactions in proteins. *Folding and Design*, *1*(4), 289–298. doi:10.1016/S1359-0278(96)00042-9

Smith, C. A., & Kortemme, T. (2010). Structure-based prediction of the peptide sequence space recognized by natural and synthetic PDZ domains. *Journal of Molecular Biology*, *402*(2), 460–474. doi:10.1016/j.jmb.2010.07.032

Smith, C. A., & Kortemme, T. (2011). Predicting the Tolerated Sequences for Proteins and Protein Interfaces Using RosettaBackrub Flexible Backbone Design. *PLoS ONE*, *6*(7), e20451. doi:10.1371/journal.pone.0020451

Smith, T. F., Lee, J. C., Gutell, R. R., & Hartman, H. (2008). The origin and evolution of the ribosome. *Biology Direct*, *3*, 16. doi:10.1186/1745-6150-3-16

Soifer, H. S., Rossi, J. J., & Saetrom, P. (2007). MicroRNAs in Disease and Potential Therapeutic Applications. *Molecular Therapy*, *15*(12), 2070–2079. doi:10.1038/sj.mt.6300311

Solis, A. D., & Rackovsky, S. R. (2006). Improvement of statistical potentials and threading score functions using information maximization. *Proteins: Structure, Function, and Bioinformatics*, *62*(4), 892–908. doi:10.1002/prot.20501

Solis, A. D., & Rackovsky, S. R. (2008). Information and discrimination in pairwise contact potentials. *Proteins: Structure, Function, and Bioinformatics*, *71*(3), 1071–1087. doi:10.1002/prot.21733

Solis, A. D., & Rackovsky, S. R. (2009). Information-theoretic analysis of the reference state in contact potentials used for protein structure prediction. *Proteins: Structure, Function, and Bioinformatics*, *78*(6), 1382–1397. doi:10.1002/prot.22652

Sood, V. D., & Baker, D. (2006). Recapitulation and Design of Protein Binding Peptide Structures and Sequences. *Journal of Molecular Biology*, *357*(3), 917–927. doi:10.1016/j.jmb.2006.01.045

Sossin, W. S., & DesGroseillers, L. (2006). Intracellular Trafficking of RNA in Neurons. *Traffic*, *7*(12), 1581–1589. doi:10.1111/j.1600-0854.2006.00500.x

Stamm, S., Ben-Ari, S., Rafalska, I., Tang, Y., Zhang, Z., Toiber, D., Thanaraj, T. A., et al. (2005). Function of alternative splicing. *Gene*, *344*, 1–20. doi:10.1016/j.gene.2004.10.022

Steffen, P., Voß, B., Rehmsmeier, M., Reeder, J., & Giegerich, R. (2006). RNAshapes: an integrated RNA analysis package based on abstract shapes. *Bioinformatics*, *22*(4), 500 –503. doi:10.1093/bioinformatics/btk010

Stormo, G. D. (1998). Information Content and Free Energy in DNA-Protein Interactions. *Journal of Theoretical Biology*, *195*(1), 135–137. doi:10.1006/jtbi.1998.0785

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics*, *16*(1), 16–23. doi:10.1093/bioinformatics/16.1.16

Stormo, G. D., & Zhao, Y. (2010). Determining the specificity of protein–DNA interactions. *Nature Reviews Genetics*, *11*(11), 751–760. doi:10.1038/nrg2845

Su, Y., Zhou, A., Xia, X., Li, W., & Sun, Z. (2009). Quantitative prediction of protein–protein binding affinity with a potential of mean force considering volume correction. *Protein Science*, *18*(12), 2550–2558. doi:10.1002/pro.257

Sul, J.-Y., Wu, C. K., Zeng, F., Jochems, J., Lee, M. T., Kim, T. K., Peritz, T., et al. (2009). Transcriptome transfer produces a predictable cellular phenotype. *Proceedings of the National Academy of Sciences*, *106*(18), 7624 –7629. doi:10.1073/pnas.0902161106

Svoboda, P., & Cara, A. D. (2006). Hairpin RNA: a secondary structure of primary importance. *Cellular and Molecular Life Sciences*, *63*(7-8), 901–908. doi:10.1007/s00018-005-5558-5

Sykes, M. T., & Levitt, M. (2005). Describing RNA Structure by Libraries of Clustered Nucleotide Doublets. *Journal of Molecular Biology*, *351*(1), 26–38. doi:10.1016/j.jmb.2005.06.024

Taft, R. J., Pheasant, M., & Mattick, J. S. (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *BioEssays*, *29*(3), 288–299. doi:10.1002/bies.20544

Tanaka, E., Bailey, T., Grant, C. E., Noble, W. S., & Keich, U. (2011). Improved similarity scores for comparing motifs. *Bioinformatics*, *27*(12), 1603–1609. doi:10.1093/bioinformatics/btr257

Tang, Y., & Nilsson, L. (1999). Molecular dynamics simulations of the complex between human U1A protein and hairpin II of U1 small nuclear RNA and of free RNA in solution. *Biophysical Journal*, *77*(3), 1284–1305. doi:10.1016/S0006-3495(99)76979-1

Thyme, S. B., Jarjour, J., Takeuchi, R., Havranek, J. J., Ashworth, J., Scharenberg, A. M., Stoddard, B. L., et al. (2009). Exploitation of binding energy for catalysis and design. *Nature*, *461*(7268), 1300–1304. doi:10.1038/nature08508

Treger, M., & Westhof, E. (2001). Statistical analysis of atomic contacts at RNA-protein interfaces. *Journal of Molecular Recognition*, *14*(4), 199–214. doi:10.1002/jmr.534

Tsai, D. E., Harper, D. S., & Keene, J. D. (1991). U1-snRNP-A protein selects a ten nucleotide consensus sequence from a degenerate RNA pool presented in various structural contexts. *Nucleic Acids Research*, *19*(18), 4931–4936. doi:10.1093/nar/19.18.4931

Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, *249*(4968), 505–510. doi:10.1126/science.2200121

Ule, J., Jensen, K. B., Ruggiu, M., Mele, A., Ule, A., & Darnell, R. B. (2003). CLIP Identifies Nova-Regulated RNA Networks in the Brain. *Science*, *302*(5648), 1212 –1215. doi:10.1126/science.1090095

Ule, J., Stefani, G., Mele, A., Ruggiu, M., Wang, X., Taneri, B., Gaasterland, T., et al. (2006). An RNA map predicting Nova-dependent splicing regulation. *Nature*, *444*(7119), 580–586. doi:10.1038/nature05304

Ulge, U. Y., Baker, D. A., & Monnat, R. J. (2011). Comprehensive computational design of mCreI homing endonuclease cleavage specificity for genome engineering. *Nucleic Acids Research*, *39*(10), 4330 –4339. doi:10.1093/nar/gkr022

Vidal, M., Cusick, M. E., & Barabási, A.-L. (2011). Interactome Networks and Human Disease. *Cell*, *144*(6), 986–998. doi:10.1016/j.cell.2011.02.016

Wade, P. A. (2001). Methyl CpG-binding proteins and transcriptional repression. *BioEssays*, *23*(12), 1131–1137. doi:10.1002/bies.10008

Wahid, F., Shehzad, A., Khan, T., & Kim, Y. Y. (2010). MicroRNAs: Synthesis, mechanism, function, and recent clinical trials. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research*, *1803*(11), 1231–1243. doi:10.1016/j.bbamcr.2010.06.013

Waldman, Y. Y., Tuller, T., Shlomi, T., Sharan, R., & Ruppin, E. (2010). Translation efficiency in humans: tissue specificity, global optimization and differences between developmental stages. *Nucleic Acids Research*, *38*(9), 2964 –2974. doi:10.1093/nar/gkq009

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., et al. (2008). Alternative isoform regulation in human tissue transcriptomes. *Nature*, *456*(7221), 470–476. doi:10.1038/nature07509

Wang, G., & Dunbrack, R. L. (2005). PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Research*, *33*(Web Server issue), W94–98. doi:10.1093/nar/gki402

Wang, H., & Laughton, C. A. (2009). Evaluation of molecular modelling methods to predict the sequence-selectivity of DNA minor groove binding ligands. *Physical Chemistry Chemical Physics*, *11*(45), 10722–10728. doi:10.1039/B911702D

Wang, Xin, Wang, K., Radovich, M., Wang, Y., Wang, G., Feng, W., Sanford, J. R., et al. (2009). Genome-wide prediction of cis-acting RNA elements regulating tissue-specific pre-mRNA alternative splicing. *BMC Genomics*, *10 Suppl 1*, S4. doi:10.1186/1471-2164-10-S1-S4

Wang, Xiaoqiang, McLachlan, J., Zamore, P. D., & Tanaka Hall, T. M. (2002). Modular recognition of RNA by a human pumilio-homology domain. *Cell*, *110*(4), 501–512. doi:10.1016/S0092-8674(02)00873-5

Wasserman, W. W., & Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. *Nature Reviews Genetics*, *5*(4), 276–287. doi:10.1038/nrg1315

Watson, J. D., & Crick, F. H. C. (1953). Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *Nature*, *171*(4356), 737–738. doi:10.1038/171737a0

Wereszczynski, J., & McCammon, J. A. (2011). Statistical Mechanics and Molecular Dynamics in Evaluating Thermodynamic Properties of Biomolecular Recognition. *Quarterly Reviews of Biophysics*, *FirstView*, 1–25. doi:10.1017/S0033583511000096

Westhof, E., & Fritsch, V. (2011). The Endless Subtleties of RNA-Protein Complexes. *Structure*, *19*(7), 902–903. doi:10.1016/j.str.2011.06.006

Whisstock, J. C., & Lesk, A. M. (2003). Prediction of Protein Function from Protein Sequence and Structure. *Quarterly Reviews of Biophysics*, *36*(03), 307–340. doi:10.1017/S0033583503003901

Wickham, H. (2011). The Split-Apply-Combine Strategy for Data Analysis. *Journal of Statistical Software*, *40*(1), 1–29.

Wilbur, W. J. (1985). On the PAM matrix model of protein evolution. *Molecular Biology and Evolution*, *2*(5), 434 –447.

Wilden, U., & Kuehn, H. (1982). Light-dependent phosphorylation of rhodopsin: number of phosphorylation sites. *Biochemistry*, *21*(12), 3014–3022. doi:10.1021/bi00541a032

Williamson, A. J. K., & Whetton, A. D. (2011). The requirement for proteomics to unravel stem cell regulatory mechanisms. *Journal of Cellular Physiology*, *226*(10), 2478–2483. doi:10.1002/jcp.22610

Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Briefings in Bioinformatics*, *9*(4), 326 –332. doi:10.1093/bib/bbn016

Wingender, E., Dietze, P., Karas, H., & Knüppel, R. (1996). TRANSFAC: A Database on Transcription Factors and Their DNA Binding Sites. *Nucleic Acids Research*, *24*(1), 238 –241. doi:10.1093/nar/24.1.238

Wintjens, R., Liévin, J., Rooman, M., & Buisine, E. (2000). Contribution of cation-pi interactions to the stability of protein-DNA complexes. *Journal of Molecular Biology*, *302*(2), 395–410. doi:10.1006/jmbi.2000.4040

Wu, T., Wang, J., Liu, C., Zhang, Y., Shi, B., Zhu, X., Zhang, Z., et al. (2006). NPInter: the noncoding RNAs and protein related biomacromolecules interaction database. *Nucleic Acids Research*, *34*(Database issue), D150–152. doi:10.1093/nar/gkj025

Xie, Z., Hu, S., Qian, J., Blackshaw, S., & Zhu, H. (2011). Systematic characterization of protein-DNA interactions. *Cellular and Molecular Life Sciences*, *68*(10), 1657–1668. doi:10.1007/s00018-010-0617-y

Xu, B., Yang, Y., Liang, H., & Zhou, Y. (2009). An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. *Proteins: Structure, Function, and Bioinformatics*, *76*(3), 718. doi:10.1002/prot.22384

Yamasaki, S., Nakamura, S., Terada, T., & Shimizu, K. (2007). Mechanism of the Difference in the Binding Affinity of E. coli tRNAGln to Glutaminyl-tRNA Synthetase Caused by Noninterface Nucleotides in Variable Loop. *Biophysical Journal*, *92*(1), 192–200. doi:10.1529/biophysj.106.093351

Yanover, C., & Bradley, P. (2011). Extensive protein and DNA backbone sampling improves structure-based specificity prediction for C2H2 zinc fingers. *Nucleic Acids Research*, *39*(11), 4564 –4576. doi:10.1093/nar/gkr048

Zeisel, A., Kostler, W. J., Molotski, N., Tsai, J. M., Krauthgamer, R., Jacob-Hirsch, J., Rechavi, G., et al. (2011). Coupled pre-mRNA and mRNA dynamics unveil operational strategies underlying transcriptional responses to stimuli. *Molecular Systems Biology*, *7*. doi:10.1038/msb.2011.62

Zhang, C., Liu, S., Zhu, Q., & Zhou, Y. (2005). A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. *Journal of Medicinal Chemistry*, *48*(7), 2325–2335. doi:10.1021/jm049314d

Zhang, Y. (2009). Protein Structure Prediction: Is It Useful? *Current Opinion in Structural Biology*, *19*(2), 145–155. doi:10.1016/j.sbi.2009.02.005

Zhang, Y., Hubner, I. A., Arakaki, A. K., Shakhnovich, E., & Skolnick, J. (2006). On the origin and highly likely completeness of single-domain protein structures. *Proceedings of the National Academy of Sciences*, *103*(8), 2605 –2610. doi:10.1073/pnas.0509379103

Zhang, Y., & Skolnick, J. (2005). The protein structure prediction problem could be solved using the current PDB library. *Proceedings of the National Academy of Sciences*, *102*(4), 1029 –1034. doi:10.1073/pnas.0407152101

Zhao, H., Yang, Y., & Zhou, Y. (2011). Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. *Nucleic Acids Research*, *39*(8), 3017 –3025. doi:10.1093/nar/gkq1266

Zhao, Y., & Stormo, G. D. (2011). Quantitative analysis demonstrates most transcription factors require only simple models of specificity. *Nature Biotechnology*, *29*(6), 480–483. doi:10.1038/nbt.1893

Zheng, S., Robertson, T., & Varani, G. (2007). A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. *The FEBS Journal*, *274*(24), 6378–6391. doi:10.1111/j.1742-4658.2007.06155.x

Zhou, H., & Zhou, Y. (2002). Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. *Protein Science*, *11*(11), 2714–2726. doi:10.1110/ps.0217002

de Vries, S. J., Melquiond, A. S. J., Kastritis, P. L., Karaca, E., Bordogna, A., van Dijk, M., Rodrigues, J. P. G. L. M., et al. (2010). Strengths and weaknesses of data-driven docking in critical assessment of prediction of interactions. *Proteins*, *78*(15), 3242–3249. doi:10.1002/prot.22814

van Driel, R., Fransz, P. F., & Verschure, P. J. (2003). The eukaryotic genome: a system regulated at different hierarchical levels. *Journal of Cell Science*, *116*(20), 4067 –4075. doi:10.1242/jcs.00779

van Kouwenhove, M., Kedde, M., & Agami, R. (2011). MicroRNA regulation by RNA-binding proteins and its implications for cancer. *Nature Reviews Cancer*, *11*(9), 644–656. doi:10.1038/nrc3107

van Nimwegen, E. (2007). Finding regulatory elements and regulatory motifs: a general probabilistic framework. *BMC Bioinformatics*, *8*(Suppl 6), S4. doi:10.1186/1471-2105-8-S6-S4

Šponer, J., Leszczynski, J., & Hobza, P. (2001). Electronic properties, hydrogen bonding, stacking, and cation binding of DNA and RNA bases. *Biopolymers*, *61*(1), 3–31. doi:10.1002/1097-0282(2001)61:1<3::AID-BIP10048>3.0.CO;2-4

Appendix 1. Code

A . Sequence Logo Code

I employed the concept of a sequence logo to illustrate concepts related to specifically recognized motifs. Schneider and Stephens (1990) first described the sequence logo as a means for displaying position weight matrix (PWM) data. Stormo (1998) described the mathematics of the PWM. The sequence logo provides a visual means of conveying the information in a position independent weight matrix and is easy to read.

A number of computational tools have been created for creating sequence logo. The weblogo interface is an online tool for creating logos from sequence alignments (Crooks, Hon, Chandonia, & Brenner, 2004). Protein logos also appear in literature (Schuster-Bockler, Schultz, & Rahmann, 2004). However, none of those provided versatile tools. The seqLogo package for the Bioconductor project provided the basis for a more generalizable sequence logo generator (Bembom, 2007; Gentleman et al., 2004). Bioconductor provides a general framework for genome analysis within the R Project for Statistical Computing (R Development Core Team, 2011). The seqLogo package was written in a manner that was highly specialized for the display of a single sequence logo.

I extended and rewrote the seqLogo package such that it could be used for displaying RNA and protein motifs and so that it can be used for displaying additional data from an R dataframe. My improvements thus include (1) an extended alphabet

containing all 26 English letters such that any one-letter-code may be used and (2) a panel function for use with the R lattice graphics package (Sarkar, 2008). The new code imports letter shapes from an ordered list of letters in a postscript file (A.1.b below). Thus, the alphabet is complete and any installed font may be easily used for logo creation. The code is now presented as a panel function (see A.1.a below) for use with the existing barchart function in the `lattice` package (Sarkar, 2008). The used of lattice graphics allows more information about entries in position weight matrices to be stored in a more statistics friendly dataframe format. Among the advantages of using R dataframes are the ability to process and summarize data using the native `aggregate` function or the sophisticated `plyr` package (Wickham, 2011). While I borrowed the grid graphics functions and was guided by the conceptual framework of the original seqLogo function (Bembom, 2007), my extensions represent a major rewriting of the graphics code and implement a major conceptual shift to take advantage of a more modern graphics system.

My sequence logo code allowed the adaptation of sequence logo motifs for use with new molecule types and for comparison of logo columns that represent positions and data that do not conform to the constraints of a traditional motif. The updated sequence logo code is presented in A.1 below. A usage example is given in A.2 below. This code was used to produce the logos used throughout this dissertation.

1. Lattice compatible sequence logo code

a. Sequence logo panel function

```
require(lattice)
require(plyr)
require(foreach)
require(grid)
```

```
source("new_pwm_alphabets.R")

# color scheme
logotheme <- function(N=4)
  {
    rc <- function(N)
      rainbow(N, s=1, v=1, start=0, end=max(1,N-1)/N, gamma=1, alpha=1)

    theme <- canonical.theme(TRUE)
    theme$superpose.polygon$col <- rc(N)
    theme$background <- NULL
    theme
  }

# create shape information for a sequence logo column
columnBuilder <- function(alphabet, colors)
  {
    letterShapes <- getLetters(alphabet)

    # set up color table.  Any undefined colors become black.
    dc <- length(alphabet) - length(colors)
    if (dc > 0)
      colors <- c(colors, rep(gray(0),dc))
    names(colors) <- alphabet

    customizeLetter <- function(letter, x.pos, y.pos, h, w)
      {
        letterdata <- letterShapes[[letter]]
        cl <- list()
        cl$x <- letterdata$x*w + x.pos - w/2
        cl$y <- letterdata$y*h + y.pos
        cl$id <- letterdata$id
        cl$fill.letter <- letterdata$fill.letter
        cl
      }

    # pass 5 vectors defining the position of letters
    function(letters, x, y, h, w)
      {
        # will only include letters that are defined in the alphabet
        ok <- letters %in% alphabet
        column.data <- foreach(letter=letters[ok], x=x[ok], y=y[ok],
                               h=h[ok], w=w[ok], .combine=regionAdd) %do%
             customizeLetter(as.character(letter), x, y, h, w)

        # convert fill.letter to a color and store in fill
        column.data$fill <- colors[column.data$fill.letter]
        column.data
      }
  }

# make the PWM column as a grid.polygon object
panel.PWMColumn <- function(x, y, groups, col, width, height,...)
  {
    cb <- columnBuilder(levels(groups), col)
```

```
    pd <- cb(as.character(groups), x, y, height, width)
    grid.polygon(x=unit(pd$x,"native"), y=unit(pd$y,"native"), id=pd$id,
                 gp=gpar(fill=pd$fill,col="transparent"),...)
  }

# a PWM panel for lattice for use with barchart function
panel.pwm <- function (x, y,  groups = NULL,
   col = if (is.null(groups)) plot.polygon$col else superpose.polygon$col,
   box.ratio = 1, box.width = box.ratio/(.2 + box.ratio),
   reference = TRUE,
   border = if (is.null(groups)) plot.polygon$border else
       superpose.polygon$border,
   lty = if (is.null(groups)) plot.polygon$lty else superpose.polygon$lty,
   lwd = if (is.null(groups)) plot.polygon$lwd else superpose.polygon$lwd,
   fill.col=NULL, ...)
{
    plot.polygon <- trellis.par.get("plot.polygon")
    superpose.polygon <- trellis.par.get("superpose.polygon")
    reference.line <- trellis.par.get("reference.line")
    keep <- (function(x, y, groups, subscripts, ...) {
        !is.na(x) & !is.na(y) & if (is.null(groups))
            TRUE
        else !is.na(groups[subscripts])
    })(x = x, y = y, groups = groups, ...)
    if (!any(keep))
        return()
    x <- as.numeric(x[keep])
    y <- as.numeric(y[keep])
    if (!is.null(groups))
      {
        groupSub <- function(groups, subscripts, ...)
          groups[subscripts[keep]]
        if (!is.factor(groups))
          groups <- factor(groups)
        nvals <- nlevels(groups)
      }

    if (is.null(groups))
      panel.rect(x = x, y = rep(0, length(x)), col = col,
                 border = border, lty = lty, lwd = lwd,
                 width = rep(width, length(x)),
                 height = y, just = c("centre","bottom"))
    else
      {
        col <- rep(col, length.out = nvals)
        width <- box.width
        for (i in unique(x))
          {
            ok <- x == i
            ord <- sort.list(y[ok], decreasing=FALSE)
            pos <- y[ok][ord] > 0
            nok <- sum(pos, na.rm = TRUE)
            if (nok > 0)
              panel.PWMColumn(x=rep(i, nok),
                              y=cumsum(c(0,y[ok][ord][pos][-nok])),
```

```
                                groups=groups[ok][ord][pos],
                                   col=if(!is.null(fill.col)) fill.col
                                         else col, #col=col,
                                width = rep(width,nok),
                                   height = y[ok][ord][pos])
          }
      }
  }
```

b. Extended sequence logo alphabet

```
library(grImport)
library(foreach)
library(grid)

## shape code extends original seqLogo shape list format
# functional "functor" (clusure) for combining regions
regionAdder <- function()
  {
    x <- NULL
    y <- NULL
    id <- NULL
    fill.letter <- NULL

    add <- function(letterdata)
      {
        x <<- c(x, letterdata$x)
        y <<- c(y, letterdata$y)
        id <<- c(id, letterdata$id + ifelse(is.null(id),0,max(id)))
        fill.letter <<- c(fill.letter, letterdata$fill.letter)
        TRUE
      }

    get <- function()
      list(x=x, y=y, id=id, fill.letter=fill.letter)

    list(addRegion=add, getCombined=get)
  }

# add a list of region definitions
regionAddList <- function(regdata)
  {
    ra <- regionAdder()
    sapply(regdata, ra$addRegion)
    ra$getCombined()
  }

# add all regions passed as args.
regionAdd <- function(...)
  {
    regdata <- list(...)
    regionAddList(regdata)
  }
```

```
## Code for representing letters from
setClass("letter", representation(x="numeric", y="numeric",
                                  r="numeric", ch="character"))

getPSLetterObjects <- function(psfile)
{
  PostScriptTrace("~/work/R/alpha_bs_bold.ps")
  alphabet <- readPicture("~/work/R/alpha_bs_bold.ps.xml")

  # basic alphabet character names
  AtoZ <- foreach(chcode=as.raw(seq(65,65+25,1)), .combine=c) %do%
       {rawToChar(chcode)}

  make.regions <- function(x)
    {
      step.func <- function(val)
        {
          v <- 0
          foreach(i=val, .combine=c) %do% {v <- v+i; v}
        }
      step.func(names(x) == 'move')
    }

  make.unit.path <- function(unscaled.path, size=NULL)
    {
      r <- range(unscaled.path)

      # if size is defined then
      if (!is.null(size))
        {
          mid <- mean(r)
          r <- c(mid-size/2,mid+size/2)
        }

      # use range make a unit-size vector, also assume and ignore last move
      (unscaled.path[seq(1, length(unscaled.path)-1)] - r[[1]])/(r[[2]] - r[[1]])
    }

  # make a polygon letter object from
  make.letter <- function(letterch, path, xs=NULL, ys=NULL)
    {
      x <- make.unit.path(path@x, xs)
      y <- make.unit.path(path@y, ys)
      reg <- make.regions(x)

      # warn in case where regions from y are not equal to regions form x
      if ( !all(reg == make.regions(y)) )
          warning("invalid letter definition: moves in x and y differ")

      # v is coordinates,
      close.paths <- function(v, r)
        {
          rpath <- v[names(v)=='move']
          v <- v[reg == r]
```

```
          if (r>2)
            {
              rp <- rpath[seq(1,r-1)]
              v <- c(rp, v, rev(rp))
            }
          v
        }
      x <- foreach(i=unique(reg), .combine=c) %do% close.paths(x, i)
      y <- foreach(i=unique(reg), .combine=c) %do% close.paths(y, i)

      new("letter", x=x, y=y, r=rep(1,length(x)), ch=letterch)
    }

learn.block.width <- function(alphabet)
    {
      rvec <- foreach(alpha.paths=alphabet@paths, .combine=c) %do%
                  {r=range(alpha.paths@x); abs(r[2]-r[1])}
      max(rvec)
    }

  # build the alphabet as described above
  make.alphabet <- function(alphabet)
    {
      letter.width <- learn.block.width(alphabet)
      alpha <- foreach(path=alphabet@paths, ch = AtoZ) %do%
                  make.letter(ch, path, letter.width)
      names(alpha) <- AtoZ
      alpha
    }

  # return list of letters
  make.alphabet(alphabet)
}


# make old-style alphabet to bridge the gap to a new class-based letter drawing
getPSLetters <- { function()
  {
    PSLetterObjs <- getPSLetterObjects("alpha_bs_bold.ps")

    ## defines a letter region -- a filled segment or an entire letter
    letterRegion <- function(letter,x,y,id=1)
        list(fill.letter=letter,x=x,y=y,id=rep(id,length(x=x)))

    letterObj2Letter <- function(letterObj)
        letterRegion(letterObj@ch, letterObj@x, letterObj@y)

    ## unknown letters are replaced with a box of the correct height.
    letterUnknown <- function(letter)
      {
        #A box is used for unknown regions
        x <- c(0,1,1,0)
        y <- c(0,0,1,1)
        letterRegion(letter,x,y)
      }
```

```
  getLetterRegion <- function(l)
  {
    if(! l %in% names(PSLetterObjs))
      return(letterUnknown(l))
    letterObj2Letter(PSLetterObjs[[l]])
  }

  ## build the alphabet list, use letterUnknown (box) for any undefined letters
  buildLetterList <- function(alphabet=NULL)
    {
      if (is.null(alphabet))
        alphabet <- names(PSLetterObjs)
      letters <- foreach(l=alphabet) %do% getLetterRegion(l)
      names(letters) <- alphabet
      letters
    }

  ## encapsulation hides the letter building code from the user
  buildLetterList
}}()
```

2. Usage for sequence logos using R lattice

The sequence logos are created from a standard R dataframe using the same approach that would be used to create a grouped barplot using the lattice package. The `barchart` function is more sophisticated than the native `barplot` graphics function in that it allows multipanel graphics to be created based on information in additional columns in a dataframe. The use of features from the Lattice plot library allows for taking advantage of more modern features of the plot system.

This approach has some additional advantages over the sequence logo code in the current version of Bioconductor (Bembom, 2007; Gentleman et al., 2004). The use of the lattice plot system allows for easy creation of multi-panel plots comparing experimental or computational logos created using different conditions or criteria. Additionally, the information used to build the logos may now be taken from standard R dataframes. Dataframes provide a more statistical format for experimental information and different conditions can be represented as factors in data frame columns. Tools for

applying transformations to dataframes such as `transform` and `aggregate` may be applied to the data. More advanced tools such as those provided by the `plyr` and `foreach` libraries may also be applied in analysis.

The code at the end of this section demonstrates the use of the panel function `panel.pwm` in the section above for plotting a randomly generated PWM column:

```
library(lattice)

vals <- sample(1:10, 4)
tmp <- data.frame(pos='P1', base=c("A","C","G","U"), prob=vals/sum(vals))

samplelogo <- with(tmp, barchart(prob ~ pos, group=base, panel=panel.pwm,
                          ylim=c(0,1), par.settings=logotheme(4),
                          scales=list(y=list(draw=FALSE), x=list(draw=FALSE)),
                                      ylab=NULL))

print(samplelogo)
```

B. Sample Scoring Function in Rosetta

I performed base specificity tests using a mixture of features that were in Rosetta3 in early 2010 and custom scoring and selection modules. Leaver-Fay et al. (2011) recently described the modern Rosetta framework. The framework contained many parts that were particular to DNA that needed to be modified. Additionally the framework did not contain function objects for custom residue selections or for monitoring the energies of particular interactions. The following function demonstrates a custom scoring function in the C++ language to query and report interactions that determine base specificity:

```
//! test specificity
void spec_test(string id = "")
{
    // read structure
    tt << "reading structure" << endl;
    pose::PoseOP pose_op(new pose::Pose);
    io::pdb::pose_from_pdb( *pose_op, options::start_file() );
```

```
// alter protein sequence
vector<bjerre::ResidueID> mut_resid;
utility::vector1<Size> res_list;
string res_mut_file( options::option[bjerre_options::residue_mut_file] );
bjerre::read_resid_file(res_mut_file, mut_resid);
boost::sort(mut_resid);
res_list.resize(mut_resid.size());
boost::transform(mut_resid, res_list.begin(), bjerre::pos_lookup(*pose_op));
bjerre::thread_protein_sequence(*pose_op,
                     options::option[bjerre_options::new_residue_seq],
                      res_list);

// create score function
scoring::ScoreFunctionOP scorefxn_op(new scoring::ScoreFunction);
scorefxn_op->set_energy_method_options( scoring::methods::
               EnergyMethodOptions().exclude_DNA_DNA( false ) );


// get score weights from database file
scorefxn_op->add_weights_from_database_file(options::
                                option[bjerre_options::score_weights]);
// alter score weights as specifiesd w/ -score_comp & -score_weight options
bjerre::set_score_weights_from_options(*scorefxn_op);
scoring::ScoreFunctionInfoOP sfxninfo( scorefxn_op->info());

// get energies
ti << "total score " << (*scorefxn_op)(*pose_op) << endl;
ti << "scores present" << endl;
sfxninfo->scores_present().print();
ti << "total score summary" << endl;
pose_op->energies().total_energies().print();

// read resid selection from file or by proximity (bases for preference test)
utility::vector1<Size> pos_list;
string res_sele_file = options::option[bjerre_options::design_residue_file];

tt << "reading position list: ";
utility::vector1<bjerre::ResidueID> resid_list;
bjerre::read_resid_file(res_sele_file, resid_list);
bjerre::print_vec(tt, resid_list);
// lookup positions
pos_list.resize(resid_list.size());
boost::transform(resid_list, pos_list.begin(), bjerre::pos_lookup(*pose_op));

// prefix used for intermediate structure files
string prefix;
if (!string(options::option[bjerre_options::design_out_prefix]).empty())
  {
        tt << "will output strutures with the prefix " << prefix << endl;
        prefix = options::option[bjerre_options::design_out_prefix] + id;
  }

// position preferences for nucleic acid bases
{
```

```
if (!options::option[bjerre_options::disable_minimization] and
             !options::option[bjerre_options::disable_pre_minimization])
    bjerre::minimize_protein_sidechains(*pose_op, *scorefxn_op, res_list);

bjerre::sequence_space::MonomerSet rna_set;
bjerre::sequence_space::make_rna_monomer_set(rna_set);

// if a moving residue set is not selected use a Neighbor packer task and
// monitor all residues contacting NA otherwise read the residues to move
// from the file and use a fixed packer task...
bjerre::sequence_space::task_gen_sp task_gen_ptr;
vector<Size> moving_residues;

// set packing residues
task_gen_ptr = bjerre::sequence_space::
               neighbor_task_gen_sp( new bjerre::sequence_space::
                                     NeighborPackerTaskGenerator(scorefxn_op,
                                     options::option[bjerre_options::
                                                     rna_interface_cutoff],
                                     options::option[bjerre_options::
                                                     pack_rna]) );
tt << "selecting residues contacting selected bases as movable residues"
   << endl;

vector<Size> tmp;
bjerre::sequence_space::get_sidechain_contact_neighbors(tmp, *pose_op,
                                                        pos_list, 6.0);
boost::sort(tmp);
boost::set_union(tmp, res_list, back_inserter(moving_residues));

pos_min_ptr = bjerre::sequence_space::
              pos_minimizer_sel_prot_sc_sp(new bjerre::sequence_space::
                        pos_minimizer_sel_prot_sc(*scorefxn_op,
                                                  task_gen_ptr) );

string spec_file( options::option[bjerre_options::spec_out_file] + id );

// scorers of interest
bjerre::sequence_space::pos_scorer_explicit_sp
      new_ps_ptr(new bjerre::sequence_space::
                     position_scorer_explicit(1.0, 1.0, false));

// attach PWM  records
bjerre::sequence_space::pos_scorer_record_design_sp
      new_rec_ptr = make_design_record(*pose_op, moving_residues,
                                       0.0, id);
new_ps_ptr->add_pos_scorer_record( new_rec_ptr );

// combined scorer
bjerre::sequence_space::pos_multi_scorer_sp
      combined_ps_ptr(new bjerre::sequence_space::
                           position_multi_scorer() );
combined_ps_ptr->add_scorer(new_ps_ptr);

bjerre::sequence_space::rna_subs_provider_sp
```

```
            subs_provider_ptr(new bjerre::sequence_space::
                                rna_subs_provider(scorefxn_op, task_gen_ptr,
                                                  combined_ps_ptr,
                                                  pos_min_ptr));

        // structure writers for intermediate and final structure output.
        bjerre::sequence_space::pos_scorer_explicit_sp
        pos_scorer_ptr(new bjerre::sequence_space::
                           position_scorer_explicit(1.0, 1.0, false));

        if (!prefix.empty())
              bjerre::sequence_space::
              make_write_delta_final(prefix, subs_provider_ptr, task_gen_ptr,
                                     pos_scorer_ptr, rna_set);

        // preference test
        bjerre::sequence_space::preference_test pref_test
              = for_each(pos_list.begin(), pos_list.end(),
                         bjerre::sequence_space::
                           preference_test(pose_op, subs_provider_ptr,
                                           rna_set) );

        // write
        if (!spec_file.empty())
              {
                    bjerre::safely_lock_file(spec_file);
                    ofstream score_output_stream(spec_file.c_str(), ios::app);
                    pref_test.print_scores(score_output_stream);
                    bjerre::safely_unlock_file(spec_file);
              }

        new_rec_ptr->write_record(spec_file+".new");
    }
}
```

Appendix 2. Math

A. Lennard-Jones Equation

The 12-6 Lennard-Jones potential is strongly repulsive at very small *r* (Lennard-Jones & Devonshire, 1937).

$$V(r) = 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right]$$

B. Hamming Distance

The Hamming distance is the minimum number of changes that must be made to a sequence in order to restore the correct sequence. To evaluate the overall effectiveness of a scoring function in my computational structure-based approach to predicting the preferred binding target sequence of a protein, I employ the average Hamming distance metric. For each structure, the Hamming distance is equal to the fraction of incorrect predictions in a structure:

$$d = \frac{1}{N}\sum_{i=1}^{N}\left(s_i^{\mathrm{a}} \neq s_i^{\mathrm{p}}\right)$$

s^{a} and s^{p} are the actual and predicted sequences respectively. In evaluating the performance of scoring function, I report the average per-structure Hamming distance. The average Hamming distance is the average over the structures included in the set of test structures.

Vita

Daniel Bjerre was born in Salmon, Idaho in 1980 and raised in Reno, Nevada. In 2002, he graduated from Swarthmore College with a B.A. in Physics. He earned a M.S. in Physics from Portland State University in 2006 for his work in the field of biophysics. In 2012, he was awarded a Ph.D. in Biochemistry from the University of Washington.